Université de Montréal

Contributions à la sonification d'image et à la classification de sons

par

Ohini Kafui TOFFA

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Novembre, 2021

Université de Montréal
Faculté des arts et des sciences

Cette thèse intitulée:

Contributions à la sonification d'image et à la classification de sons

présentée par:

Ohini Kafui TOFFA

a été évaluée par un jury composé des personnes suivantes:

_____

Jean Meunier
président-rapporteur

_____

Max Mignotte
directeur de recherche

_____

Sébastien Roy
membre du jury

Alessandro Lameiras Koerich
examinateur externe

François Cavayas
représentant du doyen de la FES

RÉSUMÉ

L'objectif de cette thèse est d'étudier d'une part le problème de sonification d'image et de le solutionner à travers de nouveaux modèles de correspondance entre domaines visuel et sonore. D'autre part d'étudier le problème de la classification de son et de le résoudre avec des méthodes ayant fait leurs preuves dans le domaine de la reconnaissance d'image.

La sonification d'image est la traduction de données d'image (forme, couleur, texture, objet) en sons. Il est utilisé dans les domaines de l'assistance visuelle et de l'accessibilité des images pour les personnes malvoyantes. En raison de sa complexité, un système de sonification d'image qui traduit correctement les données d'image en son de manière intuitive n'est pas facile à concevoir.

Notre première contribution est de proposer un nouveau système de sonification d'image de bas-niveau qui utilise une approche hiérarchique basée sur les caractéristiques visuelles. Il traduit, à l'aide de notes musicales, la plupart des propriétés d'une image (couleur, gradient, contour, texture, région) vers le domaine audio, de manière très prévisible et donc est facilement ensuite décodable par l'être humain.

Notre deuxième contribution est une application Android de sonification de haut niveau qui est complémentaire à notre première contribution car elle implémente la traduction des objets et du contenu sémantique de l'image. Il propose également une base de données pour la sonification d'image.

Finalement dans le domaine de l'audio, notre dernière contribution généralise le motif binaire local (LBP) à 1D et le combine avec des descripteurs audio pour faire de la classification de sons environnementaux. La méthode proposée surpasse les résultats des méthodes qui utilisent des algorithmes d'apprentissage automatique classiques et est plus rapide que toutes les méthodes de réseau neuronal convolutif. Il représente un meilleur choix lorsqu'il y a une rareté des données ou une puissance de calcul minimale.

Mots clés :

Personnes malvoyantes, synthèse audio, retour auditif, écran tactile, accessibilité

image, classification de sons environnementaux, modèle binaire local, apprentissage automatique, spectrogramme de signal audio

Abstract

The objective of this thesis is to study on the one hand the problem of image sonification and to solve it through new models of mapping between visual and sound domains. On the other hand, to study the problem of sound classification and to solve it with methods which have proven track record in the field of image recognition.

Image sonification is the translation of image data (shape, color, texture, objects) into sounds. It is used in vision assistance and image accessibility domains for visual impaired people. Due to its complexity, an image sonification system that properly conveys the image data to sound in an intuitive way is not easy to design.

Our first contribution is to propose a new low-level image sonification system which uses an hierarchical visual feature-based approach to translate, using musical notes, most of the properties of an image (color, gradient, edge, texture, region) to the audio domain, in a very predictable way in which is then easily decodable by the human being.

Our second contribution is a high-level sonification Android application which is complementary to our first contribution because it implements the translation to the audio domain of the objects and the semantic content of an image. It also proposes a dataset for an image sonification.

Finally, in the audio domain, our third contribution generalizes the Local Binary Pattern (LBP) to 1D and combines it with audio features for an environmental sound classification task. The proposed method outperforms the results of methods that uses handcrafted features with classical machine learning algorithms and is faster than any convolutional neural network methods. It represents a better choice when there is data scarcity or minimal computing power.

Keywords:

Visually impaired, sound synthesis, auditory feedback, touch screen, image accessibility, ESC, Local Binary Pattern, Local Phase Quantization, Machine Learning, Audio Signal Spectrogram

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES APPENDICES

LISTE DES SIGLES

AD	Différence Angulaire(*Angular Difference*)
CI	Intensité du Pixel Central(*Central Pixel Intensity*)
CNN	Réseaux de Neurones Convolutifs(*Convolutional Neural Networks*)
CQT	Transformée de la Constante Q(*Constant-Q Transform*)
CRP	Tracé de Récurrence Croisée(*Cross-Recurrence Plot*)
CWT	Transformée en Ondelettes Continue(*Continuous Wavelet Transform*)
ELBP	Motif Binaire Local Étendu(*Extended Local Binary Pattern*)
ESC	Classification de Sons Environnementaux(*Environmental Sound Classification*)
FFT	Transformé de Fourier Rapide(*Fast Fourier Transform*)
FWT	Transformée en Ondelettes Rapide(*Fast Wavelet Transform*)
GFCC	Coefficients de Cepstrum de Fréquence de Gammatone Gammatone Frequency Cepstral Coefficients
GMM	Modèles de Mixture Gaussienne(*Gaussian Mixture Models*)
HSV	Teinte Saturation Valeur(*Hue Saturation Value*)
HSL	Teinte Saturation Luminance(*Hue Saturation Luminance*)
ICA	Analyse en Composantes Independantes(*Independent Component Analysis*)
kNN	k plus proches voisins(*k-Nearest Neighbor*)
LBP	Motif Binaire Local(*Local Binary Pattern*)
LPCC	Prédiction Linéaire des Coefficients Cepstraux(*Linear Prediction Cepstral Coefficients*)

LPQ	Quantification de Phase Locale(*Local Phase Quantization*)
MDS	Mise à l'Échelle Multi-Dimensionnelle (*Multi-Dimensional Scaling*)
MFCC	Coefficients de Cepstrum de Fréquence de Mel(*Mel Frequency Cepstral Coefficients*)
MPEG-7	Quantification de Phase Locale(*Local Phase Quantization*)
MRI	Image par Résonance Magnétique(*Magnetic Resonance Imaging*)
NI	Intensité du Voisin(*Neighbor Intensity*)
PCA	Analyse en Composantes Principales(*Principal Component Analisys*)
RBM	Machine Restreinte de Boltzmann(*Restricted Boltzmann Machine*)
RGB	Rouge Vert Bleu(*Red Green Blue*)
RD	Différence Radiale(*Radial Difference*)
SFuF	Fréquence Fondamentale à Court Terme(*Short-time Fundamental Frequency*)
STFT	Short-Time Fourier transform(*Transformé de Fourier à Court Terme*)
SVH	Système Visuel Humain
SVM	Machine à Vecteur de Support(*Support Vector Machine*)
PET	Tomographie à Émission de Positrons(*Positron Emission Tomography*)
VoI	Variation de l'Information(*Variation of Information*)
VoIP	Voix sur IP(*Variation over IP*)
ZCR	Fréquence de Passage à Zéro(*Zero Crossing Rate*)

Je dédie cette thèse à:

Mon père.
Pour avoir toujours cru en moi et pour m'avoir poussé à retourner sur les bancs universitaires afin de faire mon doctorat.

L'âme de ma chère mère
Qui nous a quitté si tôt.

Ma tendre épouse.
Pour son soutien indéfectible malgré mon peu de disponibilité, partagé entre mes études de doctorat et mon emploi à temps plein.

Mes testeurs sans qui je n'y serai jamais arrivé.

Mes enfants, mon frère, mes sœurs, ma famille, mes amis et tous ceux qui me sont chers.

REMERCIEMENTS

Je tiens à exprimer toute ma gratitude à mon directeur de thèse le professeur Max Mignotte, pour avoir accepté de diriger mes travaux de recherche, bien que je fusse à temps partiel, pour sa disponibilité permanente, ses conseils, sa perspicacité et ses compétences dans le domaine de l'analyse d'image et de l'audio.

Je tiens également à remercier les membres du jury pour m'avoir fait l'honneur d'accepter d'évaluer cette thèse.

Mes remerciements vont finalement a l'endroit des professeurs du département et de tous les responsables de mon unité académique.

CHAPTER 1

Introduction

1.1 Contexte de recherche

La sonification d'image consiste à traduire le contenu visuel d'une image en sons descriptifs de ce contenu. Elle permet à une personne mal voyante d'avoir une expérience sonore du contenu visuel d'une image ou vient enrichir le contenu visuel existant.

C'est une technique simple par sa définition mais complexe en pratique car elle est multimodale et nécessite des connaissances en vision par ordinateur couplées à des connaissances en synthèse de son. Elle repose essentiellement sur un bon mappage entre l'image et l'audio, entre le 2D et le 1D ce qui cause un grand défi car pour obtenir un résultat intuitif et facilement compréhensible il faut beaucoup d'heuristiques et de l'imagination.

Le choix des caractéristiques visuelles influence beaucoup le mappage audio et les différentes méthodes proposées au fil des années nous permettent de dégager deux catégories de sonification :

- la sonification de bas niveau : désigne la conversion en sons non-vocaux des caractéristiques visuelles abstraites comme les couleurs, les contours, les textures.

- la sonification de haut niveau : désigne la conversion en sons vocaux du contenu sémantique comme les classes d'objets.

Dans cette thèse, nous nous intéressons, dans un premier temps, à la sonification de bas-niveau car le système de mappage audio étant très compliqué, beaucoup de méthodes vont se limiter à peu de caractéristiques visuelles offrant une compréhension limitée du contenu de l'image à l'utilisateur. Pour régler ce problème, nous proposons une démarche hiérarchique allant du bas niveau (pixel) au haut niveau (segmentation) et combinant le bas niveau (contour de couleur et texture), les niveaux moyen et haut (dégradé ou répartition des couleurs pour chaque région de l'image) pour aller chercher beaucoup

de caractéristiques visuelles ce qui nous permet de générer en temps réel un audio riche et varié, basé sur la position spatiale du curseur de l'utilisateur dans l'image.

Dans un deuxième temps, nous nous intéressons à la sonification de haut-niveau qui consiste à décrire d'une manière vocale les objets identifiés dans l'image. Nous pensons que le fait d'utiliser de la voix pour décrire un objet qu'une personne non-voyante n'a peut-être jamais vu de sa vie à part l'avoir entendu à travers sa manifestation sonore (aboiement du chien par exemple), altère un peu l'expérience sonore de l'utilisateur. C'est pourquoi nous proposons une sonification de moyen-niveau qui consiste à mapper des sons naturels que produit un objet ou un animal à sa classe d'objet au lieu de le décrire d'une manière vocale. Cela nous amène à proposer dans notre thèse une base de données audio contenant des sons correspondant à des classes d'objets ainsi qu'une base de données d'images dédiée à la recherche sur la sonification car l'une des premières difficultés à laquelle nous avons fait face pendant nos recherches est l'inexistence d'un tel outil de recherche. Nous proposons que les sonification de bas-niveau et de haut-niveau peuvent être complémentaire à travers une application mobile Android qui permet d'expérimenter chacune des deux modes de sonification à travers des actions tactile offrant ainsi à l'utilisateur une expérience immersive complète et globale de l'image à travers du contenu sonore.

Notre thèse étant un travail qui s'intéresse à la fois au domaine visuel et audio, nous nous intéressons dans un troisième temps à la possibilité d'utiliser des algorithmes ayant fait leur preuve dans le domaine visuel dans le domaine audio. Pour cela, nous généralisons le motif binaire local, un classificateur de texture 2D ayant fait ses preuves dans l'imagerie, en 1D et l'utilisons pour classifier les textures sonores, plus précisément les sons environnementaux. Nous comparons les résultats du nouveau descripteur à des descripteurs audios existants et le combinons à d'autres descripteurs pour obtenir un résultat plus performant. Nous comparons également leurs coûts de calcul à ceux des méthodes utilisant des réseaux de neurones.

1.1.1 Sonification

La sonification est la traduction de données en sons, originalement non vocaux. Les données peuvent être de n'importe quel type, par exemple du type environnement 3D sous-marin dans le cas du sonar [54] qui utilise son système d'écholocation pour convertir les formes géométriques et les propriétés des objets détectés en son. Elles peuvent aussi être du type climatique comme dans [28] où des textures sonores traduisent des métaphores climatiques, du type sportif où les actions d'athlètes d'élites en aviron peuvent déclencher une rétroaction auditive positive ou négative pendant leurs séances d'entrainement [77] ou des quantités physiques [21].

La sonification d'image est l'application de la sonification au domaine de l'accessibilité d'image, un domaine émergent qui devient de plus en plus important et utile à cause d'un nombre croissant de personnes non voyantes ou en perte d'autonomie visuelle (croissance et vieillissement de la population). Elle se développe aussi grâce à une présence accrue des écrans tactiles, des tablettes, des téléphones intelligents et la puissance de calcul des CPUs et GPUs. Elle trouve son application dans les technologies d'assistance aux personnes malvoyantes [7, 52], la sonification des tableaux d'art dans les musées digitaux [11, 41], la composition musicale [94], le partage de photos [99], les médias sociaux [93]...

On distingue souvent deux types de sonification : bas-niveau et haut-niveau. La sonification d'image de bas-niveau vise à traduire les caractéristiques de l'image en un son non-vocal en mappant les données entre le domaine visuel et le domaine audio, entre 2 dimensions indépendantes du temps de l'image et une dimension dépendante du temps de l'audio. Trouver le bon mappage entre l'image et l'audio n'est pas trivial et nécessite beaucoup d'heuristiques pour créer un système de sonification intuitif et facile à interpréter par un auditeur qui doit entendre un son pour visualiser un contenu. Dans le domaine des technologies d'assistance pour les personnes aveugles, la sonification de bas niveau a été initialement utilisée pour développer des systèmes de navigation pour améliorer la mobilité des personnes malvoyantes à l'aide de caméras [7, 13, 17, 52]. En termes d'expérience directe avec une image naturelle ou synthétique, les chercheurs pré-

fèrent souvent utiliser la même approche que l'alphabet Braille avec des écrans tactiles imprimés en 3D pour couvrir une tablette ou un écran tactile haptique qui transforme une image virtuelle en une image physique ou un retour tactile pour les écrans contenant une forme de guides physiques superposés à l'écran et reconnue par l'application sous-jacente [27, 37]. Peu de travaux ont utilisé la sonification pour aider les personnes malvoyantes à reconnaître des objets, à ressentir les couleurs, le dégradé et la texture, à percevoir les contours et les formes d'une image ou d'une peinture synthétique et naturelle, afin de tirer des conclusions sur le contenu de l'image (ou des trames vidéo). Différentes méthodes de conversion ont été dédiées à un petit nombre de caractéristiques visuelles à la fois. La luminosité du pixel est convertie au volume du son généré [96] ou des notes musicales [51], le motif de texture en un signal périodique [50], les couleurs à l'enveloppe d'onde, la forme d'onde à la fréquence de son d'un oscillateur [34], le contour à la fréquence d'un son d'oscillateur [97]. Des méthodes plus sophistiquées ont été développées comme celle dans [76] qui exploite la richesse de la couleur en mappant l'espace colorimétrique HSV à différentes caractéristiques du son : Teinte (H) à la fréquence fondamentale, Saturation (S) à l'enveloppe spectrale du signal, Valeur (V) à l'intensité du son synthétisé. Les auteurs dans [10], ont utilisé le concept de couleur, mélange de couleurs avec la combinaison d'entités acoustiques et le degré de rugosité sur des régions naturelles pré-classées et des contours.

La sonification de haut-niveau quant à elle essaie de produire une description vocale du contenu sémantique de l'image comme les objets. Pour détecter ce contenu sémantique certains utiliseront un moteur de reconnaissance d'objets comme *Sudol et Al* [83] avec LookTel un système qui capture à distance le flux vidéo de la caméra d'un smartphone en utilisant un signal 3G, le traite avec un moteur de reconnaissance et retourne en temps réel le nom de l'objet à l'aide d'un moteur de synthèse vocale. Dans [9], les auteurs ont utilisé une Machine à Vecteur de Support (SVM) et un sac de mots visuels disponible dans OpenCV pour détecter la présence d'objets dans la scène et informer l'utilisateur via une sortie vocale. Bien que les récentes percées dans l'apprentissage automatique utilisant l'apprentissage profond [42] aient repoussé les limites de l'analyse sémantique d'image, il reste très compliqué de comprendre pleinement le contenu d'une

image. C'est pourquoi certaines recherches [56] préfèrent utiliser le crowdsourcing en temps réel et la technique d'annotation d'images pour produire une description vocale aux personnes malvoyantes. Certaines méthodes vont analyser manuellement le contenu de l'image, notamment dans le domaine de l'art, afin de créer une carte d'exploration ou une version imprimée en 3D de l'image originale, puis guider l'utilisateur à l'aide de la commande vocale et du retour haptique [11, 41, 68, 71]. De telles méthodes ont l'avantage de décrire non seulement les objets dans l'image mais aussi la couleur, la luminosité et la texture puisqu'elle est faite manuellement. Mais la plupart des méthodes de sonification de haut niveau qui sont automatisées ne décrivent que les objets identifiés dans l'image sans pouvoir fournir des informations complètes sur les contours, la forme, la variation de couleur et la texture. Une telle sonification n'est pas en mesure d'interpréter les dessins abstraits et la peinture.

1.1.2 Classement de Sons Environmentaux

La sonification permet d'aller de l'image vers l'audio. Une fois dans le domaine de l'audio, d'autres problématiques comme la classification automatique de son existent. La classification de sons environnementaux est l'identification des sons quotidiens générés par les activités humaines ou par la nature, par exemple un chien qui aboie, un feu crépitant, des pleurs de bébé, etc. Contrairement à la musique et à la parole, les sons environnementaux n'ont pas de structure car ils ont en réalité des origines diverses. Leur reconnaissance et leur classification sont l'un des domaines les plus importants du traitement du signal audio, offrant diverses applications : audition de robot, détection de contenu répréhensible, surveillance routière, maison intelligente et surveillance, détection de coups de feu [5, 38, 70, 79, 90]...

Comme pour la reconnaissance vocale, la segmentation audio et d'autres sujets de traitement du signal audio, l'ESC repose sur l'extraction de données spécifiques et de caractéristiques sonores efficaces dans les domaines temporels ou fréquentiels [14, 18, 98]. Certaines de ces caractéristiques sont la fréquence fondamentale à court terme (SFuF) [98], les filtres de Gabor [5, 90], l'énergie de courte durée (STE) [98], le taux de pas-

sage par zéro (ZCR) [98], la transformation à Q constant (CQT) [78], les coefficients cepstraux de fréquence gamma (GFCC) [82, 88], le chromagramme [81], et le coefficient cepstral de fréquence mel (MFCC) [47]. Ce dernier est le plus couramment utilisé pour l'ESC. Bien que le spectrogramme du signal audio 1D soit une image 2D (temps $\times$ fréquence), il n'est pas courant d'utiliser des fonctionnalités d'image pour classer le contenu audio. Avec la croissance et la popularité de l'apprentissage profond dans la classification d'image [43], de nombreux travaux ont commencé à utiliser les réseaux de neurones convolutifs (CNN) ou un mélange de spectrogramme, MFCC et tracé de récurrence croisée (CRP) [12], pour classer les spectrogrammes de sons [64]. Néanmoins, à notre connaissance, peu de travaux de recherche exploitants des descripteurs d'images ont été publiés.

De nombreux travaux ont remplacé les classificateurs classiques avec des réseaux de neurones convolutionnels (CNN) capable de mieux apprendre les caractéristiques temps-fréquence en utilisant le partage du poids et la mise en commun. Dans ce contexte, Huzaifah [32] a comparé CQT, CWT et la transformée de Fourier à court terme (STFT) sur les CNN. Sharma et *al.* [80] ont implémenté un CNN profond de plusieurs canaux de caractéristiques composées de MFCC, GFCC, CQT et chromagramme. D'autre part, le CNN est parfois directement appliqué au signal avec apprentissage de bout en bout [2, 86], ou à son spectrogramme sans extraction préalable de caractéristiques [64]. Les auteurs de SoundNet [6] ont formé leur réseau en transférant les connaissances discriminantes des réseaux de reconnaissance visuelle vers les réseaux sonores. Boddapati et *al.* [12] ont obtenu une bonne classification en utilisant l'apprentissage par transfert avec les réseaux GoogleLeNet et AlexNet. Il convient également de mentionner le mécanisme d'attention temporelle [45], la machine de Boltzmann restreinte (RBM) [72], le réseau très profond [19], la technique inter-classes [87], pour ne citer que ceux-là parmi une longue liste de méthodes d'apprentissage profond qui ont abordé l'ESC avec succès. Bien que les méthodes CNN fonctionnent mieux que les méthodes classiques et conventionnelles, elles sont confrontées à des problèmes de rareté des données ou de manque de diversité dans les jeux de données. Ces problèmes sont généralement résolus par les techniques d'augmentation telles que le *time stretching, le pitch shifting*, la compression

du bruit de fond [64, 73].

1.1.3 Contributions

Le but de cette thèse est l'étude des problèmes de sonification d'image et de la classification de sons environnementaux se déroulant dans deux espaces différents que sont l'image (2D) et le temps (1D). La difficulté de créer un système de sonification intuitif, la complexité de faire une interprétation complète du contenu d'une image et le peu de travaux dans le domaine rendent notre thèse pertinente car nous y proposons des modèles ayant pour but de traduire le plus de caractéristiques possibles de l'image et d'offrir à l'utilisateur une expérience sonore du contenu global de l'image à travers des actions tactiles. Avec la classification des sons environnementaux, nous explorons l'utilisation de descripteurs d'image dans le domaine temporel tout en faisant ressortir l'avantage des méthodes d'apprentissage classiques dans certaines situations face à des méthodes basées sur les réseaux de neurones.

1.1.3.1 Sonification de bas niveau

Comme expliqué précédemment, la résolution du problème de sonification d'image de bas niveau se résume à trouver le bon mappage entre l'image et l'audio. Cela nécessite beaucoup d'ingéniosité pour générer des sons qui traduisent des informations abstraites que représentent la couleur, le contour, le gradient, la texture. Nous pouvons citer des méthodes de base constituant à mapper une ou deux caractéristiques de l'image à l'enveloppe ou la fréquence du son généré à partir d'un oscilloscope ou d'une note de musique [34, 50, 51, 96, 97]. Il existe également des méthodes plus élaborées exploitant l'espace de couleur HSV pour aller chercher beaucoup plus de caractéristiques dans l'image qui vont être mappées sur la fréquence, l'enveloppe, le volume du son ou des entités acoustiques comme le son de tam-tam [10, 76].

Dans notre travail, nous proposons une approche hiérarchique qui permet de traduire le plus de caractéristiques possibles de l'image allant du bas niveau (pixel) au haut niveau (segmentation) et combinant bas niveau (contour de couleur et texture), niveau intermé-

diaire et haut niveau (dégradé ou répartition des couleurs pour chaque région de l'image). L'espace de couleur RGB étant difficilement compréhensible par la perception visuelle humaine, car basée sur des quantités de rouge, vert et bleu, nous proposons d'utiliser l'espace de couleur HSL pour son intuitivité dans la représentation de la couleur en des termes simple et compréhensible pour l'humain comme la teinte, la saturation et la luminance. C'est également un espace approprié pour la description de peinture, donc pour la conversion du monde artistique visuel vers le monde artistique musical. Vu le nombre et la richesse des caractéristiques visuelles (teinte, saturation, luminance, segmentation, contour, gradient, texture) que nous voulons traduire, la nécessité d'un modèle intuitif, le principe que les malvoyants développent une oreille musicale au fil des années [23] nous proposons de choisir un instrument de musique capable de véhiculer une richesse de sons inégalée comme un grand piano à 8 octaves, chaque octave possédant 7 notes. En se basant sur la position du curseur dans l'image et sa zone d'appartenance en termes de segmentation, un mappage image-audio peut s'établir entre la teinte et la note, la luminance et l'octave, la saturation et la pureté, la rugosité et le rythme, l'histogramme du gradient et l'enveloppe. Comparé aux méthodes existantes, le modèle proposé a pour ambition d'être le plus riche en caractéristiques visuelles et le plus intuitif.

1.1.3.2 Sonification de haut niveau

Comme décrit précédemment, la sonification de haut-niveau repose sur l'identification du contenu sémantique de l'image suivie d'une description vocale du contenu sémantique de l'image. Dans ce domaine, nous pouvons citer les méthodes automatiques basées sur l'apprentissage machine du contenu sémantique suivi d'un engin de synthèse vocal [9, 83] et les méthodes manuelles basées sur le support de personnes ne souffrant pas de problèmes de vue pour analyser et décrire le contenu de l'image [11, 41, 56, 68, 71]. Les méthodes manuelles ont l'avantage de décrire à la fois le contenu sémantique et abstrait de l'image, donc d'offrir une sonification d'image complète alors que les méthodes automatisées se limitent au haut niveau.

Dans notre thèse nous ambitionnons une sonification complète automatisée. Nous pensons que le fait d'utiliser de la voix pour décrire un objet qu'une personne non-

voyante n'a peut-être jamais vu de sa vie à part l'avoir entendu (aboiement du chien par exemple), altère un peu l'expérience sonore de l'utilisateur. Pour cela, nous proposons un système de sonification moyen-niveau qui consiste à mapper des sons naturels que produit un objet ou un animal à sa classe d'objet au lieu de le décrire d'une manière vocale. Nous proposons une base de données audio contenant des sons correspondants au 20 classes d'objets de la base de données d'image PASCAL VOC 2012 [24] : *avion, vélo, oiseau, bateau, bouteille, bus, voiture, chat, chaise, vache, table à manger, chien, cheval, moto, personne, pot de plante, mouton, canapé, train, tv*. Nous proposons également une base de données d'image dédiée à la recherche sur la sonification et organisée en couleur, gradient, texture pour la sonification de bas-niveau, et en 20 classes d'objets pour la sonification de haut niveau, une première dans le domaine. Nous regroupons le modèle de sonification de bas-niveau proposé dans la première contribution et le modèle de sonification de moyen-niveau dans une application Android de sonification complète exploitant les actions tactiles (appuyer, maintenir, glisser) pour activer les deux modes de sonification, offrant ainsi à l'utilisateur une expérience immersive complète et détaillée du contenu visuel de l'image à travers du contenu sonore.

1.1.3.3 Classification de sons environnementaux

La problématique dans le domaine de la classification de sons environnementaux repose sur le fait qu'ils n'ont pas de structure comme la voix ou la musique et posent donc un défi dans le choix des descripteurs. Les travaux dans le domaine se regroupent en deux grandes familles. Les méthodes classiques sont basées sur l'extraction de caractéristiques audio du signal (SFuF, MFCC, STE, ZCR, CQT, GFCC, filtres de Gabor, le chromagramme) suivie de classification avec des algorithmes conventionnels comme les machines à vecteur de support (SVM), le forêts aléatoire, les k-plus proches voisins (kNN) [5, 47, 78, 81, 82, 88, 90, 98]. Les méthodes d'apprentissage profond sont plus efficaces dans l'apprentissage des caractéristiques temps-fréquence du signal audio en utilisant des techniques de partage du poids, de mise en commun ou d'attention des réseaux de neurones convolutionnels [2, 6, 12, 19, 32, 45, 64, 72, 80, 86].

Dans nos travaux, nous proposons d'augmenter l'efficacité des méthodes tradition-

nelles en adaptant un descripteur d'image ayant fait ses preuves dans le domaine d'image au domaine audio. Pour cela, nous généralisons les descripteurs de texture très efficaces que sont le LBP et sa version en fréquentielle LPQ au domaine audio pour créer le LBP-1D et le LPQ-1D. Nous appliquons également le LBP original, le LBP étendu (ELBP) et le LBP variant (VAR-LBP) au spectrogramme du son et comparons l'efficacité des descripteurs d'image aux descripteurs audio. Nous utilisons la collaboration de caractéristiques pour construire une caractéristique plus efficace et supérieure dans la reconnaissance des sons environnementaux. En dépit de la bonne précision obtenue par les méthodes CNN, ils sont très demandants en données, en temps et ressources computationnelles. Dans nos travaux, nous allons également essayer de démontrer qu'au niveau de l'ESC, les méthodes classiques peuvent des fois représenter un bon compromis entre précision et rapidité en présence d'une quantité relativement faible de données, de manque de diversité au niveau des données ou de manque de puissance de calcul.

1.2 Organisation du travail

Le plan de notre thèse est structuré comme suit : après l'introduction au chapitre I, nous présentons dans le chapitre II un nouveau modèle de sonification de bas-niveau utilisant une approche hiérarchique et des notes musicales pour convertir une image en son. Notre système est intuitif et exploite plusieurs caractéristiques de l'image comme la couleur, le gradient, la texture. Dans le chapitre IV, nous proposons un système de sonification de moyen niveau utilisant du son non vocal pour décrire le contenu sémantique des images et complémenter notre système de sonification de bas-niveau. Nous proposons également une base de données de sonification d'image composée d'audio et d'images, le tout implémenté dans une application Android (Annexe 1) fonctionnant sur téléphone ou tablette à écran tactile. Dans le chapitre III, nous généralisons les motifs binaires locaux (LBP), un descripteur de texture assez connu dans le domaine de la reconnaissance de textures d'image (en raison de sa simplicité d'implémentation mais surtout de son efficacité) , à une dimension (1D) et nous le combinons avec d'autres descripteurs sonores pour classifier les sons environnementaux. Notre approche utilise ainsi

une caractéristique visuelle couramment utilisée en traitement d'images et la généralise dans ce travail pour aborder et résoudre efficacement un problème de classification de sons. Finalement, dans la conclusion générale, nous tirons un bilan de nos travaux et proposons quelques perspectives de recherche associées à cette étude.

Les principales publications dans les journaux internationaux reliées à nos travaux sont les suivantes :

1.3 Publications et soumissions

- O. K. Toffa and M. Mignotte, "A Hierarchical Visual Feature-Based Approach For Image Sonification," in IEEE Transactions on Multimedia, vol. 23, pp. 706-715, 2021, doi : 10.1109/TMM.2020.2987710.

- O.K. Toffa and M. Mignotte, Dataset and Semantic Based-Approach For Image Sonification, in Multimedia Tools and Application, May 2022, doi : 10.1007/s11042-022-12914-z.

- O. K. Toffa and M. Mignotte, "Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration," in IEEE Transactions on Multimedia, vol. 23, pp. 3978-3985,2021, doi : 10.1109/TMM.2020.3035275.

CHAPITRE 2

A Hierarchical Visual Feature-Based Approach For Image Sonification

Dans ce chapitre, nous présentons notre article publié dans IEEE Transactions on MultiMedia, intitulé : **O. K. Toffa and M. Mignotte, "A Hierarchical Visual Feature-Based Approach For Image Sonification," in IEEE Transactions on Multimedia, vol. 23, pp. 706-715, 2021, doi : 10.1109/TMM.2020.2987710**.
Nous exposons ce dernier dans sa langue originale de publication.

Abstract

This paper presents a new image sonification system that strives to help visually impaired users access visual information *via* an audio (easily decodable) signal that is generated in real time when the users explore the image on a touch screen or with a pointer. The sonified signal, which is generated for each position within the image, tries to capture the most useful and discriminant local information about the image content at different levels of abstraction, ranging from low-level (at the pixel level) to high-level (segmentation) and combining low-level (color edges and texture), mid-level and high-level (gradient or color distribution for each region of the image) features. The proposed system mainly uses musical notes at several octaves, the notion of timbre, and loudness but also uses pitch, rhythm and the distortion effect in an intuitive way to sonify the image content both locally and globally. To this end, we use perceptually meaningful mappings, in which the properties of an image are directly reflected in the audio domain, in a very predictable way. The listener can then draw simple and reliable conclusions about the image by quickly decoding the sonified result.

Keywords

Sonification, visually impaired, sound synthesis, auditory feedback, audio mapping.

2.1 Introduction

Sonification is the translation of data into sound. More generally and precisely, it is the use of non-speech audio to convey information or perceptualize data. In fact, this field has greatly progressed over the past century and currently now constitutes an established area of research. One of the earliest and most successful applications of sonification is the Geiger counter, which was invented in 1908 and which uses the rate of clicking to convey the level of radiation being detected in the immediate vicinity of the device. One of the most recent and technologically advanced applications is SONAR [54], which uses echo location, very similar to that used by bats and marine mammals (whales, dolphin, etc.), to convey information about the 3D underwater environment, not only about

the geometry (*i.e.*, position, shape, orientation of one or several near or far objects) but also, to a certain extent, about the surface properties of the detected objects, the nature of the sediments lying on the seafloor and/or the structure of the seabed. With the evolution of the computing technology and the presence of tactile screens, smartphones, tablets and wearable devices, sonification has become more interactive [20] and is used in emergency services, aircraft cockpits, assistive technologies, climate sciences [28], elite sports [77], multimodal interactive environments [84], engineering analyses and simulations and interpretations based on the sonification of physical quantities [21].

Hence, the idea behind image sonification is to find ways to translate the image data, which describe shape, color and texture (sometimes depth information), into sounds. This is a recent field that has naturally emerged after the development of image and sound processing techniques. Such sonification may be particularly useful for finding ways to represent information that would be accessible to users with visual impairments. This technique is also beneficial in circumstances where visual representations would be impossible to use or to enrich a graphical realization [26], for human-computer interactions or for medical applications [22]. In these latter application cases, auditory feedback can complement visual data without requiring a surgeon to constantly monitor the screen or to help him or her to understand critical and additional useful information.

Previous work on image sonification can be roughly divided into two categories. In high-level (symbolic) sonification, visual information is translated into natural speech language. This field is still in its infancy since it is very difficult today, if not impossible, to fully understand the semantic content of all images. Let us note that the obvious limitation of such sonification is that it is limited to images composed of objects that have obvious semantic representations. For example, it is not clear how to sonify complex shapes, color, textures or variations of these visual cues and abstract drawings and paintings. In contrast, low-level image sonification aims to transpose image features or visual information into an abstract non-verbal audio signal [97]. This work falls into this latter category. Let us also add that this type of sonification can be quite complementary to a high-level sonification type for the previously mentioned reasons. Hence, image so-

nification can be viewed as a data conversion or a data mapping between the visual and audio domains. Nevertheless, it is crucial to understand that the time-independent two-dimensional nature of an image and the temporal nature of a sound makes this conversion nontrivial, especially since a well-designed sonification system must make intuitive sense, and the listener must be able to effectively extract and discern and real-time interpret important audio features.

In biomedical applications, low-level sonification has often been used for providing audio feedback for heart rate variability, Doppler ultrasound and electroencephalography signals [74] and sometimes used for manual positioning of surgical instruments and surgical navigational system [36]. Few works have been dedicated to image sonification except [22], where a sonification of each segmented nucleus (parameterized by two discriminant statistical geometric feature-based parameters) is presented to the cytologist, in a complementary auditory form, to improve its diagnostic accuracy. Following the same principle, Ahmad *et al.* in [3] use a sonification of optical coherence tomography data, showing images of human breast adipose and tumor tissue, with the aim of distinguishing these tissue types based on the rendered audio signals.

Recent research efforts have been devoted to using low-level image sonification to produce a navigational system to assist and improve blind and visually impaired people's mobility in terms of safety and speed. A portable wearable headgear-based device with cameras located slightly above the position of the eye in [7] and a mobile tablet with two cameras in [17] have been used to build a vision assistance system that uses depth inference *via* real-time stereo matching and a depth-to-sound mapping to inform the user of the surroundings *via* sound. In the vOICe [52], the image is captured using a single video camera mounted on headgear, and the captured image, possibly a depth image as proposed in [13], is scanned from left to right for sound generation (with the sound loudness depending on the brightness of the pixel).

Exploiting disparity data with a sonification system is interesting to assist the mobility of visually impaired people to bypass obstacles and hazards when this depth information is available, and some efforts have been made in this regard, as mentioned above. Nevertheless, few works have been proposed to aid a visually impaired user to recognize

objects in a (synthetic or) natural image (or painting) or, more modestly, to help visually impaired persons perceive some characteristics of the main shapes, shown in an image, in terms of color, luminosity and texture, which would allow them then to draw some conclusions about the content of the image (or video frames).

In this context, Martins *et al.* [50] was the first to sonify a single texture pattern with a periodic audio signal. Then, chronologically, Yeo and Berger [96] used an image sonification technique based on simple raster scanning of the image to generate a sound whose loudness linearly fits the brightness value of the scanned grayscale image pixels. A similar approach was used in [51], in which pixel values are translated into a musical notes. Ivan and Radek [34] presented a simple sonification method for mapping color information to a frequency oscillator, where color information was mapped to the wave envelope, waveform and frequency of a sound. Yoshida *et al.* in [97], presented a sonification methodology based on edge gradients and distance-to-edge maps extracted from an image and *via* a mobile touch-screen device. More recently, in [8–10], the authors used the concept of color, color mixture with the combination of acoustical entities and the grade of roughness on pre-classified natural regions and edges with drum rhythms in their sonification system. A sonification tool proposed in [76] starts by scanning the loaded image from top to bottom and produces a sound for each row of the image that plays in sequence and that consists of other elementary sounds ; more precisely, the authors used the HSV color space, and in this space, the loudness is determined by the luminance or value (V) of the pixel, the signal'spectral envelope is controlled by the saturation (S) (from a sinusoid to a square wave for the lowest to the highest saturation value), and the hue (H) is mapped into the fundamental frequency of the synthesized sound. In [63], different sonification strategies for a guidance task were used to help participants to quickly find a vertical hidden target randomly placed on a virtual horizontal line on a pen tablet. Finally, in [69], the author proposed a mobile application that sonifies HSV color and greyscale images and permits blind children to recognize on an interactive screen a straight line and a curve.

In this work, we present a new image sonification system that strives to help visually impaired users access visual information *via* an audio signal. The visually impaired user

can explore the image on a touch screen and receives real-time auditory feedback about the image content at the current position. The proposed system works in real time and is intuitive. We use perceptually meaningful mappings, in which the properties of an image are directly reflected to the audio domain in a very predictable way and can be easily extracted by the listener that can draw conclusions about the image by decoding the sonified result. To this end, the audio signal that is generated tries to capture the most useful information of an image, such as low-level image processing cues (*i.e.*, color edges and texture) and mid-level cues (histogram of the gradient) at a low-level (at the pixel level) and high-level of information obtained from an efficient segmentation algorithm that represents the image content in different sub-parts or regions of coherent textural properties. The proposed method uses mainly musical notes at several octaves, the notion of timbre, and loudness but also the pitch, rhythm and distortion effect in an intuitive way to sonify locally and globally the image content. Such system could be useful to visually impaired persons by providing a special translation experience of paintings in a digital museum or an interpretation of the image of a live event received on a mobile phone.

2.2 Sonification Model

A well-designed real-time sonification system must be fast since it computes a sound for a position that varies in the image as a *probe* while presenting the result by means of either headphones or speakers. The sonified audio result must capture as much reliable information as possible about the image on many levels of abstraction, ranging from low-level to high-level and combining low-, mid- and high-level features at each location of the image. Above all, the sonified sound must be logical and consistent and make intuitive sense. The listener must be able to effectively and quickly extract (*i.e.*, interpret in real time) important and discriminant visual features of the image, easily and intuitively identifiable, from the sonified result in order to draw simple and reliable conclusions about the image content. To do that, we must propose perceptually meaningful mappings in which the properties of an image are directly reflected in the audio

domain in a very predictable way.

Our sonification model first relies on segmentation of the image. This allows us to obtain a high-level representation of the image to be sonified in which different subparts or homogeneous regions of coherent textural properties are located. To this end, we suggest using the reliable segmentation model proposed in [55], which is freely available on the Web (although any other model could be used). This method is based on the combination, in the Variation of Information (VoI) sense, of quickly and roughly estimated segmentation results obtained by the simple K-means procedure when the image is expressed in different and complementary color spaces. For each identified region or subpart of the image, a different audio sound, lasting one second, and repeating itself in a loop as long as the *pointer* (whose position is controlled by the mouse, the keyboard or a touchscreen device) remains in this region, will be generated with different characteristics, which we now explain.

2.2.1 Choice of the Color Space

First, we have to select a suitable color space in which we intend to extract our low- and mid-level discriminating visual features on each presegmented region of the image. In this sonification system, the HSL color model is used as an ideal intuitive color model, which better describes the human perception of color than the RGB model [58]. It was designed in the 1970s by computer graphics researchers to more closely align with the manner in which human vision perceives color-making attributes. In fact, the HSL color space can be easily understood, since it is the color space that can best be explained with words or simple concepts. Moreover, this is confirmed by the fact that this color space is also used by painting or drawing artists to naturally describe a color or to describe the manner in which paints of different colors mix together. In this HSL color space (see Fig. 2.1), the visual properties of a color can be described with words or simple concepts, such as Hue, Saturation and Luminance. Note that any color space based on the Munsell color system [58], like HSV, is a good candidate since it provides almost the same visual properties.

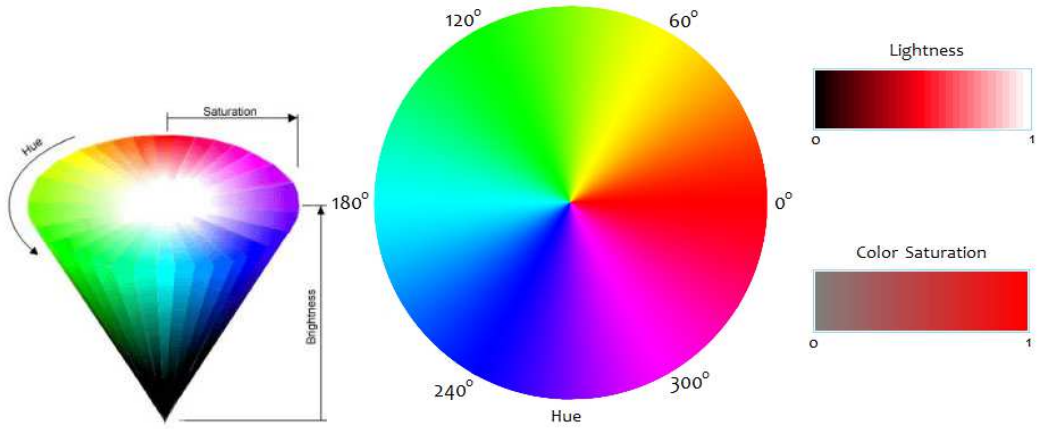


FIGURE 2.1 : The HSL color space with the hue component that is arranged in a radial slice (starting at the red hue at 0^o, passing through green at 120^o and blue at 240^o, and then wrapping back to red at 360^o) around a central axis of neutral colors, which ranges from black at the bottom to white at the top. The HSL representation models how colors mix together, with the saturation dimension resembling various shades of brightly colored paint and the luminance (or brightness) dimension resembling the mixtures of those paints with varying mounts of black or white paint.

2.2.1.1 Hue

is a term that one visually thinks of as an existing color that can be described with simple words, such as *'red'*, *'yellow'*, *'green'* or *'purple'*. Red (or green) is a distinct pure, primary hue, while the hue '*yellow*' is composed of equal quantities of (primary hues) red and green.

2.2.1.2 Saturation

is a measure of how intense or pale a color appears, and this concept is governed by the amount of white it contains. This term is generally used to describe the purity of a color. For example, *'pink'* is a tint of the color *'red'* to which between 10% and 70% of white has been added.

2.2.1.3 Luminance

Luminance can also be described by words such as bright or dark. This concept is dependent on the amount of energy that is being radiated. Thus, darkness is a lower-intensity shade of a bright red.

This color space is suitable for describing a gradient or color variation in space or time. For example, it can be seen that a cookie becomes more brown as it is baked : its hue remains relatively constant, but its luminance and saturation change during cooking [48].

2.2.2 Choice of the Frequency Sampling

In our application, we use a sampling frequency of $f_e = 16384$ samples/second, which allows us to model a maximum, the Nyquist frequency of 8kHz, that is used in most modern VoIP (Voice over Internet Protocol) communication products and which is sufficient to model complex audio signals. In addition, human sensitivity to frequency information above 8 kHz is rather limited, and conveying information above this high frequency remains perceptible by healthy human ears but is very sensitive to environmental external noise and thus difficult to decode [25].

Since we want to generate an audio signal that lasts $T = 1$ second with f_e, we have to generate a total of 16384 sound samples, which is also a power of two ($16384 = 2^{14}$) and which will allow us to then efficiently use an inverse FFT Fast Fourier Transform (since our sound mapping will be generated in the frequency domain) and to fully give 8 octaves with the highest frequency given by a 108-key grand piano. With this specification, let us note that the frequency resolution of our sonification model is $\Delta f = 1$Hz.

2.2.3 Sonification Mapping

The different steps of our sonification approach are the following :

2.2.3.1 Hue Translation

The core of our sonification mapping is based on the seven musical notes (*do-re-mi-fa-sol-la-si*) possibly played on several octaves at once (this will be explicit in Section 2.2.3.3). This choice comes from the fact that several blind or visually impaired persons naturally develop a *musical ear* (certainly due to the cortical plasticity [23][1]) and are able to easily identify every note immediately and in isolation from other played notes [29, 89][2].

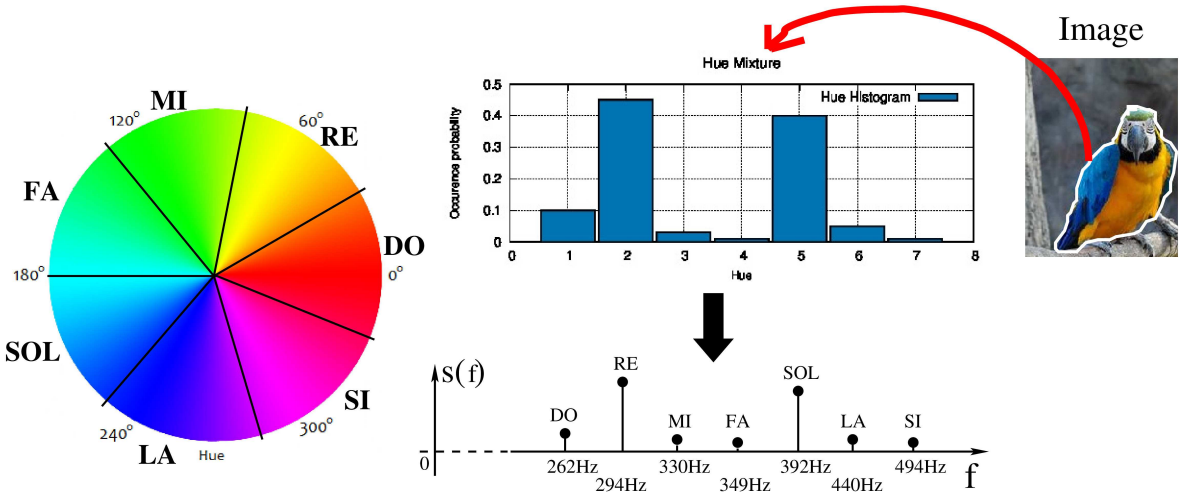


FIGURE 2.2 : Hue Translation : the re-quantized seven-bin hue histogram of each region is first estimated and then converted into an impulse function weighted by the corresponding value of the histogram, in the frequency domain, centered on the frequency corresponding to the different notes of the musical game.

Importantly, HSL offers the possibility to code a color using a precise note with only the H color channel (hue) and linearly (more precisely radially around the circle of the cone) with a semantic (well-understood) expression (*i.e.*, 0-Red, 60-Yellow, 120-Green,

[1] A part of the core area of the auditory cortex was found to be enlarged by a factor of 1.8 in the blind compared with sighted humans. Such cortical reorganization may be a consequence of the absence of visual input in combination with enhanced auditory activity.

[2] The authors show that blind people perform better than sighted individuals at tasks related to pitch discrimination and pitch-timbre categorization and on a range of auditory perception tasks. This advantage was observed only for individuals who became blind early in life.

180-Cyan, 240-Blue, 300-Magenta, 360-Red) (see Fig. 2.2). Additionaly, in this color space, colors with the same hue can then be distinguished semantically with adjectives referring to their saturation and lightness, just as our sonification system will be able to do but with different sound characteristics (such as octave and harmonics), as will be explained later.

More precisely, in our application, the hue information of each region is first modeled by a normalized re-quantized histogram with seven (one for each note) equal-width bin (in the hue interval) as a hue feature vector. In this simpler model, the texture of each pre-segmented region is herein characterized by a mixture of hues, or more precisely, by the values of the re-quantized hue histogram. This model is simple, quick to compute, and allows significant data reduction while being robust to noise and local image transformations.

Once the seven-bin histogram is computed, each of these seven re-quantized histogram values is converted into an impulse (or Dirac delta) function weighted by the corresponding value of the re-quantized histogram (see Fig. 2.2) in the frequency domain. More precisely, the first, second, …, seventh values of the histogram are converted to an impulse function centered on the frequency corresponding to the different notes of the musical game, *i.e.*, 262Hz (Do), 294Hz (Re), 330Hz (Mi), 349Hz (Fa), 392Hz (Sol), 440Hz (La), and 494Hz (Si), respectively (represented by the white keys of a piano keyboard for the octave C_4 Middle C [92]), in a manner such that in the temporal domain, the mixture of hue of each region will be translated by a mixture of pure tones or musical notes (easily identifiable by a visually impaired person) played together in a manner that sounds harmonious and determining the the so-called *pitch* of the generated sound.

2.2.3.2 Luminance Translation

Since this concept is dependent on the amount of energy that is being radiated (cf. Section 2.2.1.3), we thus bring this concept closer to the different octaves of vibration of a piano. To this end, the luminance value, initially in the interval $[0-255]$ is divided into 8 equal intervals, and each interval is assigned the name of an octave scale C_n [92],

in which is played the musical notes defined in Subsection 2.2.3.1 with :

$$C_n = \left\lceil \frac{l_R}{32} \right\rceil \tag{2.1}$$

where l_R is the mean luminance value of the region and $\lceil . \rceil$ the ceil function. For example, if $l_R = 100$, it implies the octave C_4 with the standard so-called *Middle C octave* [92] comprised within the range $[\text{Do} = 262\text{Hz} - \text{Si} = 494\text{Hz}]$. If $l_R = 180$, it implies the octave C_6 with the so-called *Soprano C octave* using the range $[\text{Do} = 1046\text{Hz} - \text{Si} = 1976\text{Hz}]$.

2.2.3.3 Saturation Translation

Since this concept describes the purity of the color (high-saturation colors look rich whereas full- and low- saturation colors look dull and grayish ; see Fig. 2.1), we translate this concept in audio space by adding to the sound previously generated more (for low-saturation colors) or fewer (for high-saturation colors) harmonics, thus making the sound more or less pure. More precisely, the saturation value, initially in the interval $[0-255]$, is divided into 8 equal intervals, and N_s is computed as follows :

$$N_s = 7 - \left\lfloor \frac{s_R}{32} \right\rfloor \tag{2.2}$$

where s_R is the mean saturation value of the region and $\lfloor . \rfloor$ the floor function. N_s represents the number of octaves ; in addition to the one in that is playing the sound previously generated (see Subsection 2.2.3.2), we duplicate this mixture of notes on several other (closest) octaves. For example, if $s_R > 224$ implying $N_s = 0$, the pure sound created on only one octave is generated. If $s_R = 100$ implying $N_s = 4$; 4 supplementary octaves (the closest to that estimated in Subsection 2.2.3.2) are added to the initial sound with a weighting amplitude of $1/N_s$ (*i.e.*, $1/4$ for our example (see also the example given in Fig. 2.3), making the generated sound less pure.

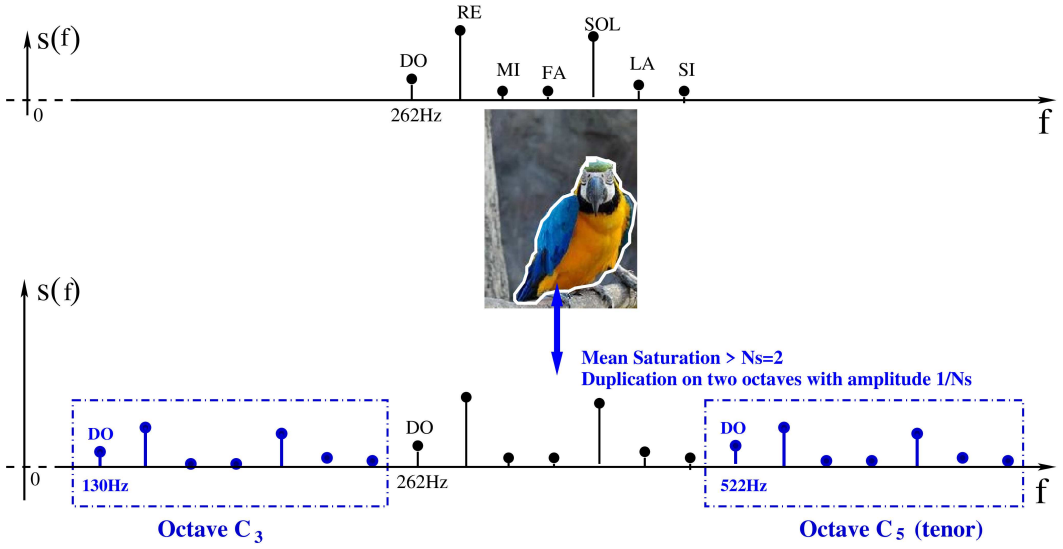


FIGURE 2.3 : Saturation Translation : In addition to the mixture of musical notes generated by the hue value (see Fig. 2.2), we duplicate it on several other (closest) octaves according to the saturation value. For example, if $s_R = 170$ implying $N_s = 2$ (see Eq. (2.2)), we duplicate this mixture on the two closest octaves of C_4 (with a weighting of 0.5).

2.2.3.4 Roughness Texture Translation

In order to aid the user's understanding of the possible roughness (due to the presence of gradients) of a textured region to be sonified, we can alter the purity of the sound signal, thanks to the concept of distortion. This can produce a *vibrant*, *rhythmic*, *growling*, or *gritty* tone depending on the type of distortion used. This effect can efficiently (and intuitively) model the grade and style of roughness of each pre-detected region.
More precisely, we can consider two different types of roughness properties of a region that can be quantified by two statistical features. The first one is the mean of the module of the first-order gradient within the region (defining the global roughness of the region), and the second is the variance of the orientation (following the four main directions) of the (first-order) gradient module (hence quantifying the presence of man-made geometric structure, such as a manufactured object or an un-natural texture).

• In our application, the first type is created by magnitude distortion by adding (noisy) randomized harmonics. This can be simply done by adding, to the frequency

bins with null amplitude, randomized values within the interval $[0\ ,\ \rho]$ with ρ proportional to the mean of the module of the first-order gradient within the region.

• The second type can be created by phase distortion by adding (noisy) randomized value $[-\beta\ ,\ \beta]$ to the phase spectrum, with β proportional to the variance of the oriented (in the four directions) module of the first-order gradient within the region.

2.2.3.5 Translation Into Temporal Space

Once the hue, saturation and luminance and the mean gradient module of each region have been mapped in the frequency domain and expressed in in terms of the magnitude spectrum vector[1] and the variance of the oriented gradient in the phase spectrum vector (thus defining the audio signal's spectral envelope (which is related to the perception of *timbre*), after Hermitian symmetry is imposed on the magnitude and on the phase spectrum, we return to the time domain, with a simple inverse fast Fourier transform (FFT), to get $s(t)$.

2.2.3.6 Histogram Gradient Translation

In Section 2.2.3.4, we have sonified, *via* the distortion effect of the audio signal, the mean and the orientation variance of the gradient magnitude as two different features related to the gradient-based region activity. Another important visual cue that remains to be expressed by sonification is the histogram or distribution of the amplitude of the gradient module. A way to express this visual cue and to give it a meaningful, interesting audio effect is through the notions of rhythm and loudness of the sonified sound. To this end, and in order to sonify this information to the user, such that it is easily decodable, we use the following strategy : we first compute a re-quantized histogram using ten equal-width bins of the first-order gradient module and use this histogram followed by

[1] Algorithmically, since $\Delta f = 1$Hz, in our application (with a sampling frequency of $f_e = 16384$ and 16384 sound samples ; see Section 2.2.2), it boils down to filling a 1D vector of length 16384 by simply putting an amplitude value in the n-th cell for the frequency nHz.

its mirror projection to weight the 1-second temporal envelope of the audio signal $s(t)$ (see Fig. 2.4).

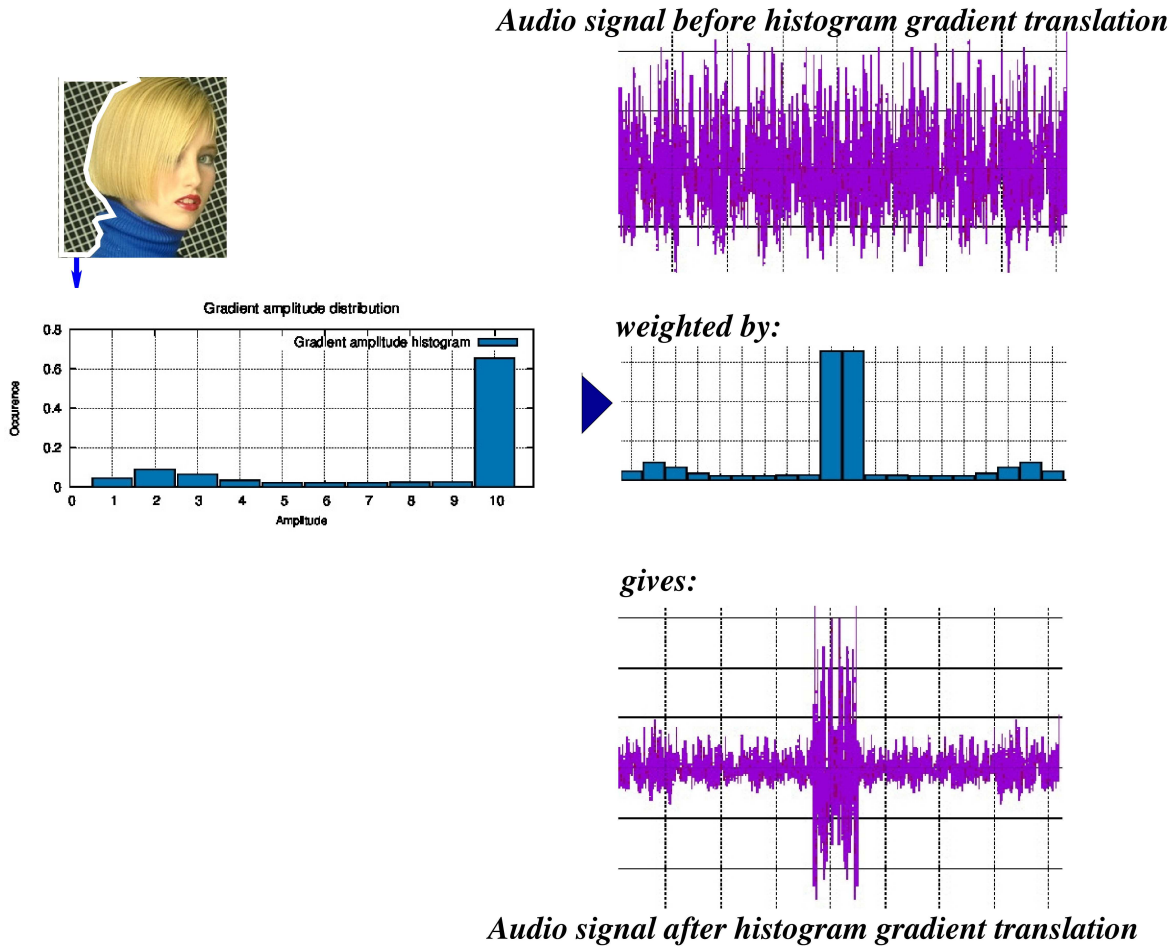


FIGURE 2.4 : The histogram or distribution of the amplitude of the gradient module within the considered region (to be sonified) is used to weight the temporal envelope of the audio signal.

More precisely, in our application, we keep a percentage p of the original signal (characterizing and encoding the image low-level visual features listed in previous points 1 to 5), and the weighting is applied for the other $100\% - p$ of the signal. This allows us to avoid generating signals of total silence, as in the example of the region given in Fig. 2.4. See Algorithm 1 for implementation details.

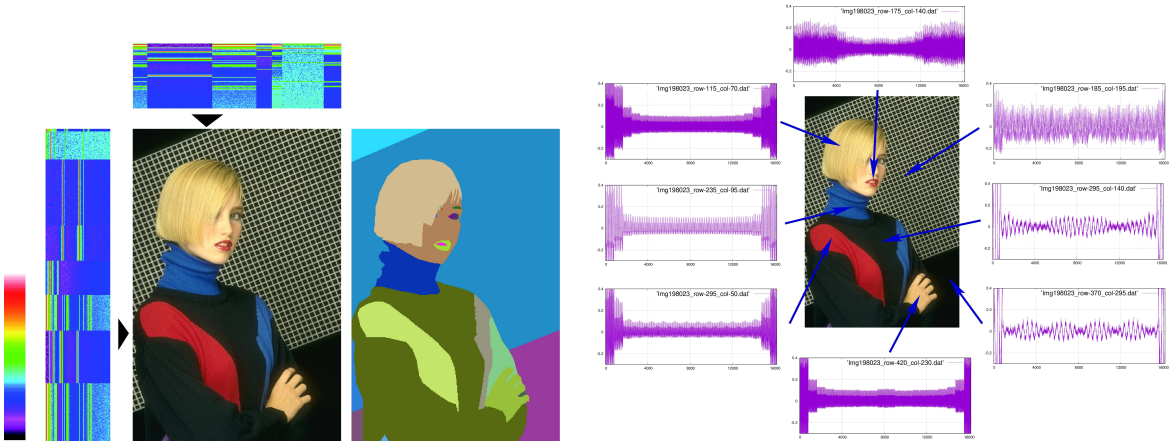
2.3 Experimental Results

In our experiments, we have tested our sonification algorithm on some images from the Berkeley segmentation database (BSD300) [49]. This image-base has both a great variability of naturally colored and textured images and a good (manually hand-segmented) segmentation for each image. This allows us to objectively analyze, discuss and highlight the pros and cons of just our sonification process. Nevertheless, in the absence of a segmentation map for each image, we can use any automatic segmentation model and especially the one proposed in [55] which obtains a segmentation score, in terms of the Rand Index equals to 0.81, meaning that on average, 81% of pairs of pixel labels are correctly classified compared to the segmentation maps of the BSD300, considered as ground-truths.

In our tests, we set ρ and β (Section 2.2.3.4), 5 times the mean of the module of the first order gradient and 5 times the variance of the oriented module of the first-order gradient, respectively. A high value of p (Section 2.2.3.6) reduces the effect of the gradient interpretation on the signal envelop, while a small one increases the risk of signal with silence. We then used $p = 10\%$ as a good trade off. Note that those values are empirical.

2.3.1 Discussion

When we position the pointer (controlled by the mouse or the keyboard) in the middle bottom and left border of the image shown in Fig. 2.5.a, we can easily recognize and thus localize two regions associated with a pure tone sound (lasting one second, and repeating itself in a loop) with a specific frequency corresponding to the musical note *Do* for the red part of the wool turtleneck sweater of the woman and the specific note *La* for its two blue parts (at the left arm and neck) (see Section 2.2.3.1). We can easily also guess the homogeneous black part of the sweater since the emitted sonified sound mainly vibrates at the bass tones (see Sect. 2.2.3.2 (low frequencies), but with several other harmonics (with smaller amplitudes) (cf. Sect. 2.2.3.3), indicating the presence



(a) **(b)**

FIGURE 2.5 : **(a)** : Image number 198023 from the BSD300 Berkeley database [49] (left) and its segmentation (right) with the spectrogram of the generated sound when the user examines the image from left to right starting from line 320 and from top to bottom starting from column 110. The frequency resolution of the spectrogram is $\Delta f =$ 20Hz, and the spectrogram ranges from 0Hz to 2 kHz (horizontally for the left one and vertically for the top one); the data are represented with the thermal (false-) color scale shown on the far left. **(b)** : Some audio samples generated at different locations of the image.

of a very low luminance and very saturated colors such as black. Let us note that this part can be distinguished from the background that is dark since the latter region is a uniform dark region without gradient (unlike the sweater region), and thus the temporal envelope of the sonified sound is different (see Fig. 2.5.b). We can also easily localize the blond hair of this person since the sonified sound is also a pure tone corresponding to the *Re* musical note, but with a slight magnitude distortion effect due to the presence of a mean gradient in this particular region (cf. Sect. 2.2.3.4) and thus sonifying the particular textural roughness of this region and also distinguishing the hair region from the facial area. Indeed, the lightness of this part is more important, the saturation is lower, and the distribution of the gradient is radically different (see Fig. 2.5.b), thus generating a very different temporal envelope for the sonified signal. We can easily draw the outline of the person without ambiguity. Finally, the background, whose area looks like a sort of fence, is associated with a sound that is a very complex sound (highly saturated) with a lot of

(amplitude and phase) saturation (cf. Sect. 2.2.3.4) and with a very peculiar temporal envelope, creating a kind of scratchy noisy sound that grumbles regularly every half second and thus representing appropriately the regular grilling.

Fig. 2.6 shows four images from the BSD300 Berkeley database with some audio samples generated at different locations of these images. Figures (a) and (d) share the same semantic concepts as do (b) and (c). We can notice that the audio results generated at the sky of the second and third images are very similar and very characteristic. Similarly, the sonified sound generated for the modern building in the second and third images are very similar (despite a slightly different texture, demonstrating that our proposed sonification system generalizes well across regions from the same semantic concept or label) and also characteristic of a man-made structure with geometric and regular patterns. The sound emitted by the by the starfish or its background is very rich and complex with amplitude and phase distortion characteristic of complex textures. Nevertheless, we can hear easily in them the musical notes Do-Re for the starfish and Mi-Fa for the background in relation to their respective hue.

2.3.2 Validation

To validate the sonification model, we obtained a certificate of ethics from the Université de Montréal and performed a pilot study (easily reproducible, from the BSD300, to facilitate eventual further comparisons with future methods) with 14 volunteers (students and nonstudents). Note that each test lasted approximatively 1h, and each participant was involved in a minimum of two tests. This puts constraints on a person´s availability and capacity to stay motivated along all the tests and reduced the number of different participants and the number of test samples. Due to some constraints, we were unable, unfortunately to include blind persons in the study. However, we believe that if sighted people are able to perform well on our system, blind people will do better since they demonstrate better ability with sounds [29, 89]. In all the experiments we masked the images to the user with a black screen.

The study consisted of multiple experiments : mapping properties recognition, scene

description, form detection, image categorization and a longitudinal study. In the experiments all subjects explored the very same images but with different orderings to avoid presentation order effects. Only one of the subjects had a *musical ear*, and none of them had a training session beforehand.

2.3.2.1 Experiment I (Mapping Properties' Recognition)

A calibration image (Fig. 2.7) containing five rows (pitch, octave, purity, distortion and loudness) was presented to the subject in order to train him with the system. For each row, the model related property was activated, and the subject scanned the row horizontally in order to learn the behavior of the sound based on the property. The learning session of the selected property lasted approximatively 5mn. Immediately after this session, the user was tested on the selected property using a dozen square images, masked by a black screen, from the calibration (Fig. 2.7) and mosaic (Fig. 2.8) images.

After all the properties had been separately tested, we activated all the properties in order to test the effects of the combination. The user was presented each square of the mosaic (Fig. 2.8) for identification.
Five volunteers participated in the experiment, and the results of the testing are reported in Table 2.1.

Property	Question	Single	All
Pitch	What is the color of the square ?	62%	51%
Octave	Is the square dark or light ?	96%	75%
Purity	Is the square pure or dirty ?	83%	66%
Texture	Does the square have a texture ?	92.0%	72%
Loudness	Is the image contrasted ?	90%	69%

TABLE 2.1 : Results of experiment I

As we can observe, the detection of the colors using the pitch, was difficult for the participants while detecting other properties was easy. This is explained by the fact that

distinguishing between low and high frequencies or the level of the volume is easier than detecting 7 musical notes. With more training time and practice, a better result could be achieved as we will demonstrate in Section 2.3.2.5.
We observed that combining all the properties slightly affected the precision of each property, especially, for a nonmusical ear. For example, a low-saturation image, introduces higher octaves (*via* the added harmonics), which makes the image sounds more acute, as does having a higher luminosity. Thus, it is sometimes difficult to identify whether the acuity of the generated sound is due to the luminosity or saturation of the source image (*i.e.*, the pitch of the musical note or the presence of harmonics). We also observed that for a nonmusical ear, it is difficult to detect the color (or the pitch) when the sound is mainly composed of very low or very high frequencies.

2.3.2.2 Experiment II (Scene Description)

In experiment II, three participants from experiment I were given 5 minutes to explore the images shown in Fig. 2.6, without further information. These images were selected to check whether their description fits the interpretation we performed in Section 2.3.1. At the end of the exploration of each image, the participants had to provide an oral description of the scene as they imagined it. A qualitative evaluation is reported in Table 2.2. We can notice that the subjects were able to easily recognize the shapes of the objects, but failed sometimes to recognize the color when the object was highly textured or too dark.

2.3.2.3 Experiment III (Form Detection)

During this experiment, eleven subjects had a training of 5 min with the mosaic image (Fig 2.8). Only two of them were also involved in experiments I and II. We selected from the database four images (Fig. 2.9) that contained different objects in terms of shape and number. The participants were given 5 min to explore each image. They had to then say, based on their exploration, if the image contained a long ostrich's neck in a

plain, three vertical sculptures in a garden, a vase laid against a wall or six umbrellas on a terrace. The goal of this experiment was to check if the information about the regions and edge detection was properly conveyed in the model.

The confusion matrix for the results of test III is shown in Table 2.3. Most of the subjects were able to associate the images to the correct content. The greatest confusion was for the vase and the ostrich's neck, since both have a similar vertical shape in the subject's imagination. The content that was easily identified was the three vertical sculptures in a garden (image (b)).

2.3.2.4 Experiment IV (Image Categorization)

During this experiment, three pairs of images (Fig. 2.10) that look similar were presented (5 min per image) to the eleven participants from the previous experiment. They had to group them into pairs based on their global similarity in terms of the sounds heard. This third test permitted validating the mapping of the whole hierarchical visual features (regions, edge, colors and texture) into sounds.

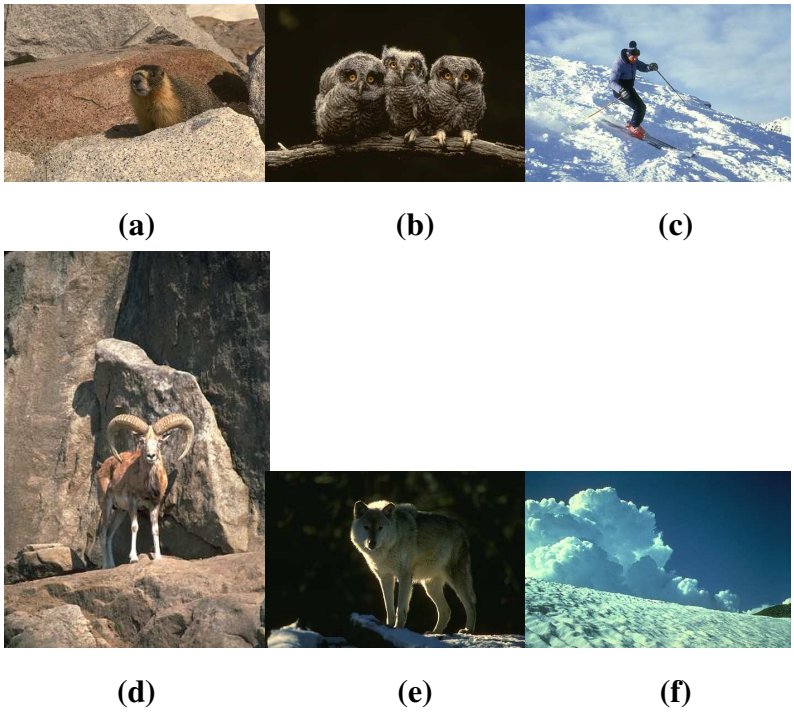


FIGURE 2.10 : Experiment IV : Image numbers **(a)** 41069, **(b)** 42044, **(c)** 61060, **(d)** 304074, **(e)** 42078 and **(f)** 176039 from the BSD300 Berkeley database [49] grouped vertically by visual similarity.

This test was more difficult than the previous one because it required the subjects to memorize many tones per image (3 on average). The association between the two images was not straightforward since several tones characterize an image, and the challenge was to use the dominant tones as a reference. The results presented in Table 2.4 indicated that some people associated a bright image with a dark one because they heard one low-frequency sound in some part of the first one. Others associated a blue tinted image with a red one because both contain high-pitch sounds, and it is not easy to distinguish musical notes at such high frequencies. However, the majority of people grouped together the red tinted (a) and (d) images (high-pitched Do sound), the dark (b) and (e) images (very saturated low-frequency sound), and the blue-dominant (c) and (f) images (high-pitched La sound).

2.3.2.5 Longitudinal Study

Two participants who were involved in all four previous experiments were allowed to use the system for a long time in order to observe their progression and learning curve. After they had used the system for two or three additional hours, they were able to easily distinguish the colors using the pitch. They improved their result on the combination testing (Section 2.3.2.1) by an average of 20%. They appreciated the research in the following terms :
Participant 1 : *I improved myself by taking the calibration test multiple times. I liked the mosaic testing, but it was with the scene description experiment that I better understood the usefulness of the research. This could be helpful for visually impaired people.*
Participant 2 : *I especially liked the scene exploration tasks (Experiments II to IV). It made me more imaginative and helped me see the utility of the project.*

2.4 Conclusion

In this paper, we have presented a new image sonification system that provides an intuitive mode of obtaining visual local spatial information and some context information about an image for visually impaired persons. The proposed system uses a set of hierarchical visual features about the image content at different levels of abstraction and perceptually meaningful mappings based on the additive synthesis technique, in the spectral domain ; it uses the concepts of timbre, loudness, pitch, rhythm and different distortion effects to translate the appearance of each individual pre-segmented region of the image into the audio domain. The proposed system allows us to easily localize different regions and classify regions into man-made and natural regions, sometimes with automated man-made object recognition. This system can be complementary to high-level image sonification (*i.e.*, an automatic verbal translation model), which is prone to errors and with which the listener can also miss all the richness, subtleties and complexities of the underlying visual information.

The validation results showed that although the subjects did not have a *musical ear*

and did not have any training session, in some cases, they were able to detect objects in the images and group images based on the visual features translated into sounds. This also showed that users were able to improve their performance on the system with more practice. These results are promising since people with visual impairments (*musical ears*) and some training sessions will surely be able to do better.

Variables

R_{max}	Maximum number of regions
N_{Hue}	Bin number of the hue
N_{Grad}	Bin Number of the gradient
N_{Samp}	Number of samples 16384
$R[][]$	Regions of Segmented Image
y	Current Region label
y_{Old}	Old Region Label
Pos_x, Pos_y	Cursor Position
$H_R[][]$	Mixture of Hue of size $R_{max} * N_{Hue}$
$G_R[][]$	Mixture of Gradient of size $R_{max} * NGrad$
$S_R[]$	Saturation table of length R_{max}
$L_R[]$	Luminance table of length R_{max}
$GM_R[]$	Gradient Mean table of length R_{max}
$OG_R[]$	Oriented gradient table of length R_{max}
$N_{Freq}[]$	Notes Frequencies of Octave C_4 of length 7
$M_{Freq}[], P_{Freq}[]$	Sound samples of length N_{Samp}

Initialization
Load Image and Convert to HSL
Compute or Load Image Segmentation to R
for *each region* $r < R_{max}$ **do**

- $H_R[r] \leftarrow ComputeHueHistogram()$
- $G_R[r] \leftarrow ComputeGradientHistogram()$
- $S_R[r] \leftarrow ComputeSaturation()$
- $L_R[r] \leftarrow ComputeLuminance()$
- $GM_R[r] \leftarrow ComputeGradientMean()$
- $OG_R[r] \leftarrow ComputeOrientedGradient()$

end

Wait For Event
while *user input and not exit* **do**
  $Pos_x, Pos_y \leftarrow GetCursorPosition()$ $y \leftarrow R[Pos_x][Pos_y]$
  if $y <> y_{Old}$ **then**
    $GenerateSound()$
    for $i < N_{Hue}$ **do**
      1. Translate hue and luminance
      $C_n \leftarrow \lceil \frac{L_R[y]}{32} \rceil$
      $M_{Freq}\lfloor N_{Freq}[i] * 2^{C_n-4} \rfloor \leftarrow H_R[y][i]$
      2. Translate saturation
      $N_s \leftarrow 7 - \lfloor \frac{S_R[y]}{32} \rfloor$
      for $l < N_s$ **do**
        $M_{Freq}\lfloor N_{Freq}[i] * 2^{Neighbors(C_n,l)-4} \rfloor \leftarrow H_R[y][i]/N_s$
      end
    end
    for $k < N_{Samp}/2$ **do**
      3. Magnitude Distortion
      if $M_{Freq}[k] = 0$ **then**
        $\alpha = 5 * rand() * GM_R[y]$ $M_{Freq}[k] \leftarrow \alpha/N_{Samp}$
      end
      4. Phase Distortion
      $\beta = 5 * rand() * OG_R[y]$ $P_{Freq}[k] \leftarrow \beta/N_{Samp}$
    end
    $HermitianSymmetry(M_{Freq}, P_{Freq})$
    $Sound \leftarrow IFFT(M_{Freq}, P_{Freq})$
    5. Translate Gradient Histogram
    $Sound \leftarrow Weight(Sound, G_R)$
    $PlaySound()$
    $y_{Old} \leftarrow y$
  else
    $PlaySoundIfNotPlaying()$
  end
end

Algorithm 1 : Image Sonification

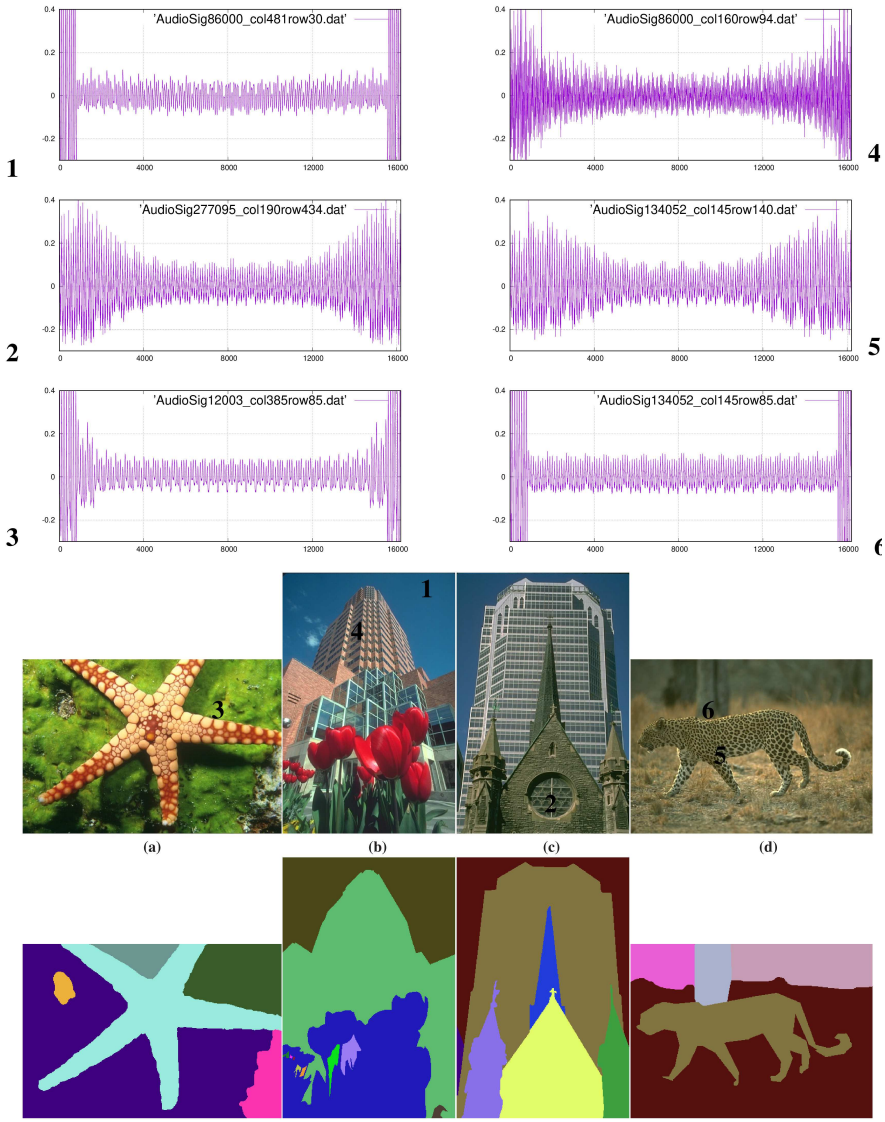


FIGURE 2.6 : Images number **(a)** 12003, **(b)** 86000, **(c)** 277095 and **(d)** 134052 from the BSD300 Berkeley database [49] with some audio samples generated at different locations of these images.

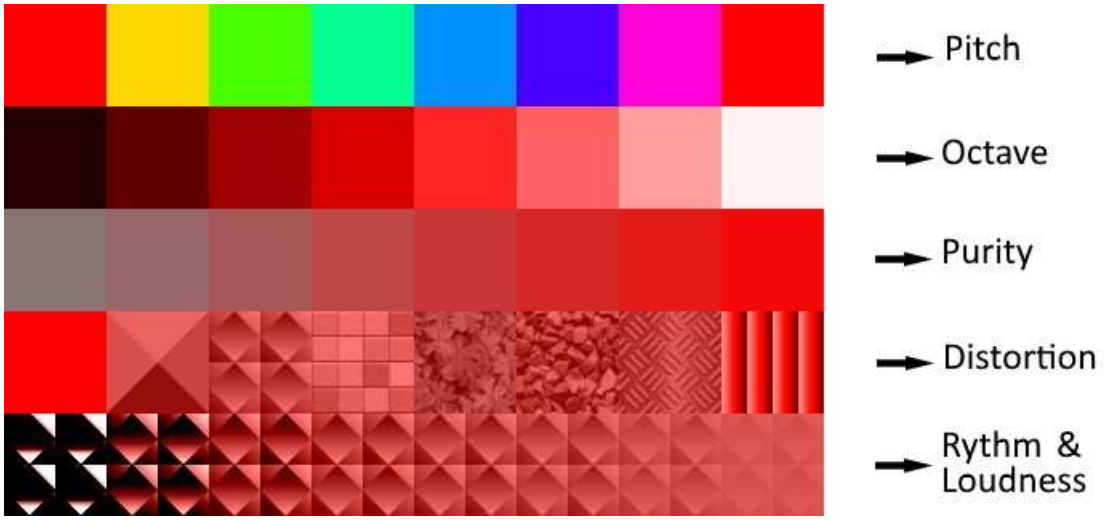


FIGURE 2.7 : Experiment I : Calibration and training image

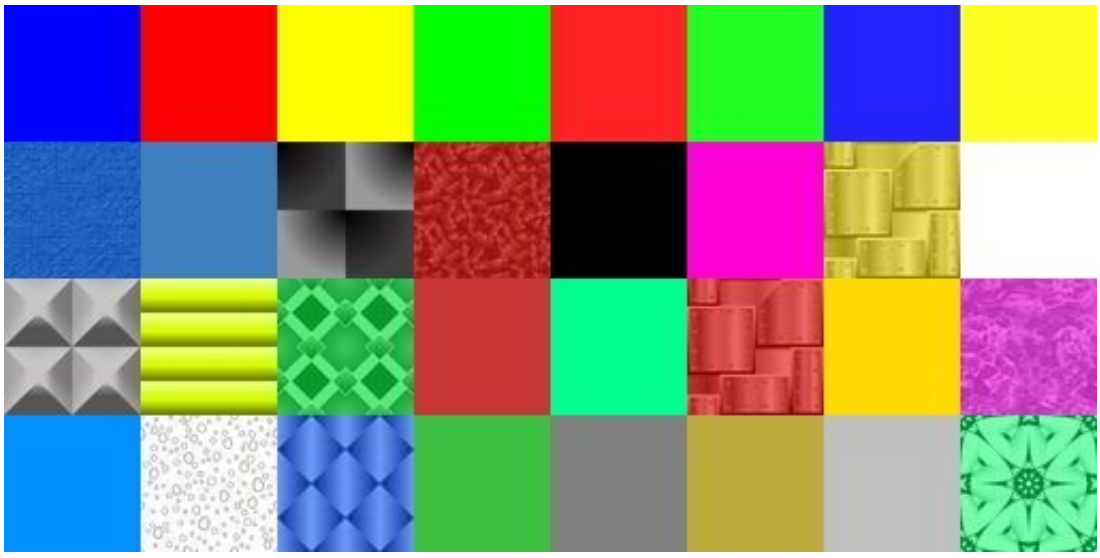


FIGURE 2.8 : Experiment I : Mosaic of testing images

Image	Subject	Description
(a)	1	A yellow textured object with spikes on a dark background. Could be a flower or a homehouse.
	2	A bright textured form with spikes on a dark background. No idea of what it could be.
	3	A very textured clear form in the shape of a star on a green background. Green sky ? No idea of what it could be.
(b)	1	A vertical pointed shape with a light red texture on the top and a dark one somewhere on the bottom. A blue background on the top. Could be a pyramid or a head with a pointed hat.
	2	A vertical textured shape with a blue colored background on top. Could be a bust of a man.
	3	A light red texture in the shape of a bottle. A clear background with no texture on the top. Could be a bottle or a flower.
(c)	1	A centered vertical shape textured in red. A clear blue background on the top. Could be a trunk.
	2	A standing elongated red form, very textured. Could be a tree trunk.
	3	A vertical dark textured shape. A clear background with no texture on the sides and top. No idea of what it could be.
(d)	1	A horizontal red textured shape on a green dark background. No idea...a landscape ?
	2	A horizontal bright textured shape centered on a green dark background. Could be a bird in the countryside.
	3	A light textured shape centered on a dark background. Could be a homehouse.

TABLE 2.2 : Results of experiment II

(a) **(b)** **(c)** **(d)**

FIGURE 2.9 : Experiment III : Image numbers **(a)** 66075 (a long ostrich's neck in a plain), **(b)** 101085 (three vertical sculptures in a garden), **(c)** 227092 (a vase laid against a wall) and **(d)** 242078 (six umbrellas on a terrace) from the BSD300 Berkeley database [49].

	(a)	(b)	(c)	(d)
(a)	**0.55**		0.27	0.18
(b)		**0.91**		0.09
(c)	0.36		**0.64**	
(d)	0.09	0.09	0.09	**0.73**

TABLE 2.3 : Confusion matrix of experiment III

	(a)	(b)	(c)	(d)	(e)	(f)
(a)	X		0.09	**0.73**	0.09	0.09
(b)		X		0.09	**0.73**	0.18
(c)			X	0.09	0.18	**0.64**
(d)				X		0.09
(e)					X	
(f)						X

TABLE 2.4 : Association Matrix of experiment IV

CHAPITRE 3

Dataset and Semantic Based-Approach For Image Sonification

Dans ce chapitre, nous présentons notre article publié dans le journal Springer Multimedia Tools and Application, intitulé : **O. Toffa and M. Mignotte, Dataset and Semantic Based-Approach For Image Sonification, in Multimedia Tools and Application, May 2022, doi : 10.1007/s11042-022-12914-z**.
Nous exposons ce dernier dans sa langue originale de publication.

Abstract

This paper presents an image-audio dataset and a mid-level image sonification system that strives to help visually impaired users understand the semantic content of an image and access visual information *via* a combination of semantic audio and an easily decodable audio generated in real time, both triggered by sliding, taping, holding actions when the users explore the image on a touch screen or with a pointer.

Firstly, we segmented the original image using a label fusion model and based on the user position in the image, a sonified signal is generated using musical notes and meaningful visual information within the active region like the color and the luminance, then the gradient and the texture.

Secondly, we integrated the semantic understanding of the image into our model using DeepLab semantic segmentation of the image and created a dataset of audio and images aligned on the 20 classes of the PASCAL VOC 2012 dataset.

The dataset of images are organized based on color, gradient, texture for low-level sonification and on semantic content with sounds for mid-level sonification.

Thirdly, in order to provide both types of information in a complementary way, the slide, tap and hold actions of a touch screen are incorporated in the model. The semantic audio providing a brief description of the visual object is played on slide action, the generated signal with color details of the object on the tap action, gradient and texture of the object on hold action. Finally, we validated our sonification model on the provided dataset during a pilot study and the subjects were generally able to identify the objects in the image, the color of the objects and even provide a general description of the scene of the image. Our system could be useful to visually impaired persons in a photo sharing application using a smartphone or for painting art description in a digital museum

Keywords

Sonification, visually impaired, touch screen, image accessibility, auditory feedback.

3.1 Introduction

Image sonification is the translation of image data, which describe shape, color and texture (sometimes depth information), objects, into sounds. It applies the sonification, the use of non-speech audio to convey information or perceptualize data, to image accessibility domain (see [59] for an interesting review of image accessibility). With an increased number of peoples who suffer from visual impairment (blind or low vision) due to the aging and growth of the world population and the increased presence of images on screens in daily activities through social media, image accessibility using sonification has gained much more attention. Sonification is already used in emergency services, aircraft cockpits, assistive technologies (Microsoft text to speech or iOS VoiceOver), climate sciences [28], elite sports [77], multimodal interactive environments [84], engineering analyses and simulations and interpretations based on the sonification of physical quantities [21]. More than ever, sonification has become more interactive [20] with the evolution of the technology of touch and tactile screen on top of smartphones, tablets and wearable devices, the increased computing capacity of CPU/GPU and advanced techniques of machine learning. Sonification applied to image is a recent field that has naturally emerged after the development of image and sound processing techniques and is used in assistive technologies for blind people through navigation [7, 52], art sonification in digital museum [11, 41], music composition [94], photo sharing [99], social media [93].

3.2 Background

Research in image sonification is usually classified in two categories : high-level sonification where a natural language is used to describe the visual content of the image and low level sonification where the same content is translated to a non-speech audio. High-level sonification aims to describe the semantic content of an image to a visually impaired people. To detect that semantic content some will use an object recognition engine like *Sudol et Al* [83] with LookTel a system that remotely captures the stream

video from the camera of a smartphone using a 3G signal, process it with object recognition engine and returns in real time the name of the object using text to speech engine. In [9], the authors used SVM detection and Bag of visual world available in OpenCV to detect the presence of objects in the scene and inform the user via a speech output. Though recent breakthroughs in machine learning using deep learning [42] have pushed the boundaries of semantic image analysis, it stays very complicated, to fully understand the semantic content of an image. That is why some researchers [56] prefer to use real-time crowdsourcing and image annotation technique to speak the alt text aloud to people who are visually impaired. Some methods will manually parse the content of the image, especially in the art domain, in order to create an exploration map or a 3D printed version of the original image, then guide the user using voice control and haptic feedback [11, 41, 68, 71]. Such methods have the advantage to describe not only the objects in the image but also the color, the luminosity and texture since it is manually made. Definitely, most of automated methods in high-level sonification will only describe the objects identified in the image without being able to provide full information related to the edges, the shape, the color variation, and texture. Such sonification is not able to interpret abstract drawings and painting.

Low-level image sonification aims to translate image features into a non-speech audio by mapping data between the visual and the audio domains, between the 2 dimensions time-independent of the image and 1 dimension time-dependent of the audio. This is not trivial and requires heuristic design to create a sonification system that is intuitive and easy to interpret by a listener who must hear a sound to visualize a content. In the domain of assistive technologies for blind people, low-level sonifications were initially used to develop navigation systems to improve visually impaired people's mobility using cameras [7, 13, 17, 52]. In terms of direct experience with a natural or synthetic image, peoples prefer to use the same approach as Braille alphabet with 3D printed tactile graphics to cover a tablet or haptic touchscreen which transform virtual image in a physical one or a tactile feedback for touch screens in the form of physical guides that are overlaid on the screen and recognized by the underlying application [27, 37]. Few

works have used sonification to help visually impaired people to recognize objects, feel colors, gradient and texture, perceive edges and shapes in a synthetic, natural image or painting, in order to draw some conclusions about the content of the image (or video frames). Simple conversion methods were dedicated to map different characteristics of the image to sound. The brightness of the pixel is converted to the volume of generated sound [96] or musical notes [51], texture pattern into a periodic signal [50], colors to the wave envelope, waveform and frequency of a sound of an oscillator [34], edge to frequency of a sound of an oscillator [97]. More sophisticated methods were developed like the one in [76] which exploit the richness of the color by mapping HSV color space to different characteristics of the sound : Hue (H) to the fundamental frequency, Saturation (S) to signal's spectral envelope, Value (V) to the loudness of the synthesized sound. The authors in [10], used the concept of color, color mixture with the combination of acoustical entities and the grade of roughness on pre-classified natural regions and edges with color distribution for each region of the image) features. The proposed system mainly uses musical notes at several octaves, the notion of timbre, and loudness but also uses pitch, drum rhythms in their sonification system. A more recent approach developed in our previous paper [85] captures the most useful and discriminant local information about the image content at different levels of abstraction, ranging from low-level (at the pixel level) to high-level (segmentation) and combining low-level(color edges and texture), mid-level and high-level (gradient or and the distortion effect in an intuitive way to sonify the image content both locally and globally. Though a low-level sonification can interpret the abstract content of an image it can not identify and name the objects present in that image.

In this work, we present a new image sonification system, called mid-level sonification, because it exploits low-level and high-level sonification techniques in a complementary way and use a semantic non-speech audio instead of a voice description of the objects. Such system offers to the end user a global experience that covers most of features available in an image : objects and semantic content, the edges, the shape, the color variation, the luminosity and texture. We segmented the original image using a label fu-

sion model [55] and used a low-level hierarchical-based sonification approach [85] to generate a sonified signal that exploits musical notes and the position of the user in the screen in order to convey meaningful visual information within the active region like the color and the luminance at first, then the gradient and the texture finally. In order to understand the semantic context of the image for high-level sonification, we integrated DeepLab [16, 75] semantic segmentation of the image that permits, not only to identify up to 20 classes of the PASCAL visual object class 2012 [24], but also their edge and shape instead of just their presence using a bounding box as in some previous papers [9]. We created a dataset of non-speech audio aligned with the 20 classes because sounds provide better description to a blind person than a vocal description of something they can never see but are used to hear. Since there is not any images dataset dedicated to image sonification and peoples usually struggle to find data on which testing their model we provided a dataset of images that contain different color variation, gradient, texture for low-level sonification and 20 different classes of objects for high-level sonification. To provide both types of information in a complementary way without any additional haptic or tactile accessory device, we incorporated the slide, tap and hold actions of a touch screen into the model. A semantic non-speech audio providing a brief description of the visual object is played on slide action, a generated signal with more detailed on the color on tap, gradient and texture of the object on hold action. Contrary to methods [11, 41, 68, 71] that offer an equivalent complete experience, our approach is automated, uses non-speech descriptive sound instead of a vocal description of content that a blind person has never seen and does not involve additional haptic or tactile material except a touch screen. Our system could be useful to visually impaired persons in a photo sharing application using a smartphone or for painting art description in a digital museum. The dataset and source code of the application are available at https ://github.com/ohinitoffa/ImgSonficiation2.

3.3 Sonification Model

To have a complete and detailed description of an image, a sonification model must convey as much as discriminant local information available at the position of the pointer in the image. Different levels of abstraction must be involved from pixel level to object level passing by a segmentation level. High-level sonification permits to tackle object and semantic content while low-level handle all other abstract information like the luminance, gradient, color, texture, edge and shape. The challenging part on touch screen is to combine both levels of sonification without any additional hap-tic or tactile device.

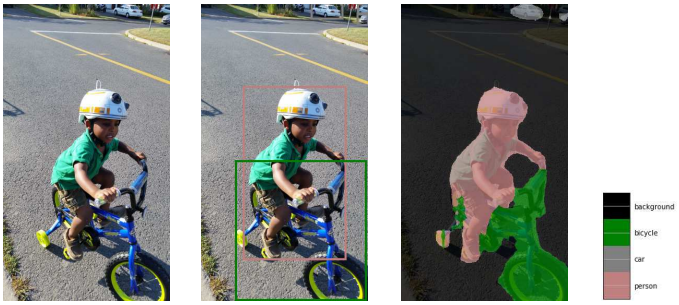
3.3.1 High-level Sonification

Understanding the semantic content of an image is one of the important points in high-level image sonification system. There are usually three ways to handle such concern : image description, object detection with bounding box and semantic segmentation. Image description permits to understand the global context of the image without going into details. The image description of Fig. 3.1.a would be for eg. *A person on bicycle in front of cars*. Such global description does not permit to the user to experience the details of the image as an object detection technique would do by indicating an approximative position and size of the objects identified using bounding boxes (Fig. 3.1.b). A semantic segmentation (Fig. 3.1.c) goes beyond the two previous techniques by labelling all the pixel of the image, then permits to detect the shape and contours of each object.

For that reason, we exploit one of the advanced semantic segmentation algorithms available in the domain and developed by DeepLab [16] using deep convolutional networks. More specifically, we integrate its mobile version based on MobileNet [75] into our sonification mobile application called TalkingImage 2 that we will use for experimentation in Section 2.3. See Fig. 3.1.c for example of segmentation result provided by DeepLab algorithm.

Most of methods [11, 41] translate the context information extracted from the image to voice using a Text to Speech engine but we think that in a pure sonification experience,

where eyes are replaced by ears, hearing a cow moo is better than hearing the voice of person indicating the presence of a cow. Based on this assumption, we downloaded from FreeSound[1] a list of audio under a creative commons license, representing the cry, the interaction or the manifestation of each of the 20 classes of object of PASCAL Visual Object Class (VOC) 2012 dataset. PASCAL VOC challenge is a benchmark in visual object category recognition and detection, which provides the research community with a standard dataset of images and annotation, and standard evaluation procedures [49]. The train/validation data of 2012 has 11,530 images containing 27,450 regions of interest annotated objects and 6,929 segmentations. To cover as much as possible cries for some animals, we concatenated or mixed multiples sounds. For eg. we will hear the dog bark then pant while the cat will purr and meow. The length of each audio clip is 4s, sufficient for a human to identify environmental sounds with accuracy [18]



(a) Original (b) Objects detection (c) Semantic segmentation

FIGURE 3.1 : Context of an image

3.3.2 Low-level Sonification

Low-level sonification strive to describe all other abstract visual content not covered by the high level sonification like the color, the gradient, the luminosity, the texture of an object. It is complementary to the high-level sonification since after the identification of

[1] https ://freesound.org/

an object, it permits to sonify the color and all other characteristic of the object. For this purpose, we will use the hierarchical feature-based approach developed in [85] which translate using musical notes, most of the properties of an image in the audio domain, in a very predictable way.

The hierarchical feature-based approach supposes that the original image is preliminary segmented into regions and since we do not have ground-truths for the new dataset we are proposing in Section 3.4, we use an automatic segmentation based on label fusion [55] which obtains a segmentation score, in terms of the Rand Index equals to 0.81 on BSD300 [49] dataset. BSD300 is a segmentation dataset of 300 color images of size 481*321 provided by the university of Berkeley and divided into a training set of 200 images, and a test set of 100 images. A set of benchmark segmentation results provided by human observers are available for each image and used as ground truth to quantify the reliability of the proposed segmentation algorithm. See the result of such segmentation on Fig. 3.2.

The second phase is to use the HSL color model for the mapping to audio domain because it easily describes the human perception of color with words or simple concepts such as Hue, Saturation and Luminance. [58]. Considering the HSL color space, the Hue of each segmented region is mapped to the pitch of the generated sound by a system of 7 bins quantization mapped to the 7 musical notes of a piano at the octave C4 Middle C. The first, second, ..., seventh values of the histogram are converted to an impulse function centered on the frequency corresponding to the different notes of the musical game, *i.e.*, 262Hz (Do), 294Hz (Re), 330Hz (Mi), 349Hz (Fa), 392Hz (Sol), 440Hz (La), and 494Hz (Si), respectively (represented by the white keys of a piano keyboard for the octave C_4 Middle C and determine the *pitch* of the generated sound.

The luminance or bright depends on the amount of energy that is being radiated thus was translated to the level of octave or vibration of a piano. The luminance value, initially in the interval $[0-255]$ is divided into 8 equal intervals, and each interval is assigned to the name of an octave scale C_n.

The saturation which represents the amount of white a color contains and generally

used to describe the purity of a color is converted to the purity of sound. More or fewer harmonics, depending on the saturation value are added to the octave of the original sound thus make the sound more or less pure.
The roughness of the texture is translated to the rhythm and the loudness of the sound by adding distortion in phase and in magnitude to the original signal. See the visual to audio mapping and calibration in Fig. 3.3) extracted from [85] where a complete description of the algorithm is available.

In this paper we will generate two types of audio : one that contains only the color information translated to pitch, octave and purity then the second one that contains the full visual features.

(a) Original (b) Segmentation

FIGURE 3.2 : Image segmentation

3.3.3 Touch screen interaction

Conveying the information produced by the low-level and high-level sonification using only touch screen action without any additional haptic or tactic device is one of the tricky parts of our sonification system. In order to develop a system that can be easily used on a touch screen mobile device as on a touch screen desktop and also on a desktop with a classic mouse, we find it simple to limit triggering actions to slide (mouse move), tap (mouse left-click), hold (mouse right-click).

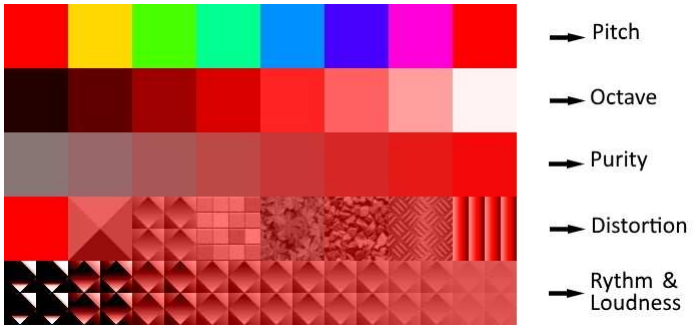


FIGURE 3.3 : Visual to audio mapping calibration and training image

In the final model, the semantic segmentation (from high-level) and segmentation (from low-level) are superimposed. On the slide action within a semantic-segmented zone tagged with an object label, the non-speech audio of the class of object is played. If labelled as background, the generated audio containing full visual features of the segmented region is played. On a hold action within a segmentation zone, a generated audio conveying the color information is played then on a tap, the audio conveying full visual features information is played. The system interaction flow is represented in Fig. 3.4. With this interaction flow, the user can identify an object, gets the color information of different parts of the object then going beyond with gradient and texture information.

3.4 Dataset

One of the first obstacles encountered during our research on image sonification is the availability of a dataset on which to evaluate our model. Most of research in the domain used set of images based on the specific feature they want to sonify. It is then difficult to have a dataset that cover multiple features from low-level to high-level sonification. In [85], a subset of the BSD300 [49], a dataset normally used for image segmentation and boundary detection study, is used but it is not suitable for semantic segmentation.

In this paper, we come up with a dataset[1] of images that covers different features used in an image sonification system like color, gradient, texture, segmentation, and

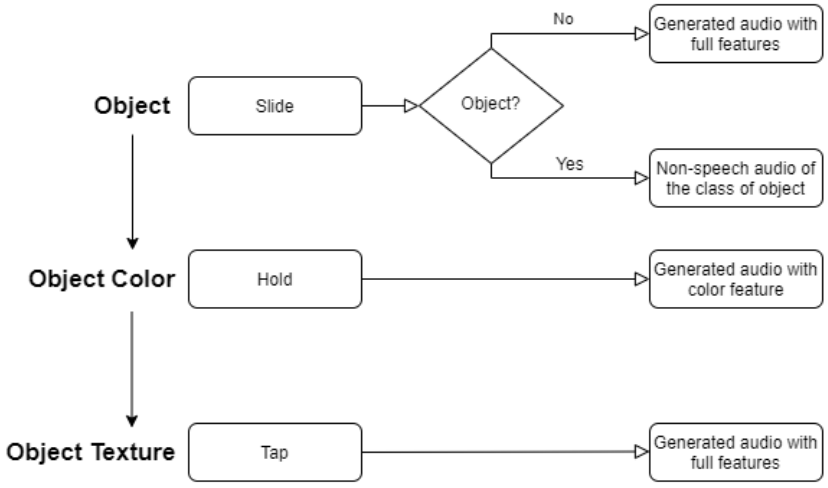


FIGURE 3.4 : User interaction flow

semantic segmentation. Some images come from our previous research on low-level sonification [85] while others come from personal library and Flick data under creative common licence for a total of 122 images. The height and width of the images were reduced to a maximum of 320 pixels because of the requirements of the segmentation algorithm used [55].

Primary colors and multiple colors variation in terms of hue, saturation and value are present in the dataset. Fig. 3.5 presents the mosaic of images of abstract content of the dataset in terms of colors variation but also luminosity, gradient, texture and roughness of the texture. In Fig. 3.6, a minimum of three images per class of object are presented with a variation of colors, texture and number of instances for a total of 20 classes of objects aligned on PASCAL VOC 2012 [49]. Such diversity in the datasets offers the possibility to evaluate the detection of different objects, the evaluation of the color, luminosity, and

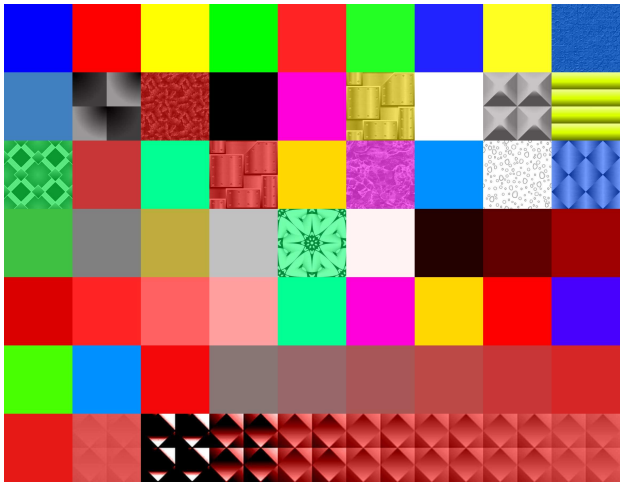


FIGURE 3.5 : Mosaic of colors and textures of the dataset

texture of different objects of the same class, number of instance of objects...

3.5 Experimental Results

We implemented the described model into an Android mobile application called TalkingImage 2 and made it tested by a group of persons who were part of the previous pilot study, then more experimented with the low-sonification part of our model [85]. All of them were volunteers composed of students and nonstudents and only one of them had a *musical ear*. Due to the Covid19 pandemic impact and the lack of availability of some persons, we were able to mobilize only six of the eleven people. Unlike the first phase of the study which took place in an office [85], this second phase was carried out remotely. Subjects were instructed on how to use the application and were recommended to use a headset. Three types of tests were proposed : object Identification, object's color Identification and scene exploration. In all the experiments the images were hidden to the user with a white or a blue screen and the images were displayed in the same order.

[1] https ://github.com/ohinitoffa/ImgSonficiation2

FIGURE 3.6 : Mosaic of objects of the dataset

3.5.1 Experiment I : Object Identification

The goal of this experiment was to test the semantic part of the sonification model by identifying different classes of objects present in the image. User was given 5 minutes to familiarize himself with the application by playing with the sonified version of the image of Fig. 3.1. Then he was given 5 minutes per image to explore the four images of Fig. 3.7 by sliding his finger and use the sound heard to name all the objects, part of the 20 classes, which are present. The results displayed in Table 3.1 shows that the users are able in most of the cases to identify the presence of the appropriate objects. However, the small resolution of the images (maximum of 320 pixels in width or height) negatively impacted the detection of a person's class because its instances cover a reduced number of pixels in the sample of images tested (3.7.b and 3.7.d). Some peoples confused the sound of sheep to cow in the image 3.7.d or the whistle of the person inside image 3.7.b to a bird song. This means that we need to improve the audio of some objects to make them more discriminative.

3.5.2 Experiment II : Object's Color Identification

The users were given a few minutes to refresh their memory with the calibration image. They were then given four images containing object with different colors (see

(a) (b) (c) (d)

FIGURE 3.7 : Experiment I : Object Identification

Image	Ground Truth	Answers
(a)	horse	horse (83.33%)
(b)	horse, person	horse (83.33%), person (50%)
(c)	cow	cow(100%)
(d)	sheep, person	sheep (60%), person (0%)

TABLE 3.1 : Results of experiment I

Fig. 3.8). They must use the slide action to identify the object, then hold their finger to identify the color of the object. The results of the experience reported in Table 3.2 shows that users are able to easily detect the color of a big object 3.2.a with homogeneous pixels and low variation than one which is small 3.2.d or contain too much color variation 3.2.c.

(a) (b) (c) (d)

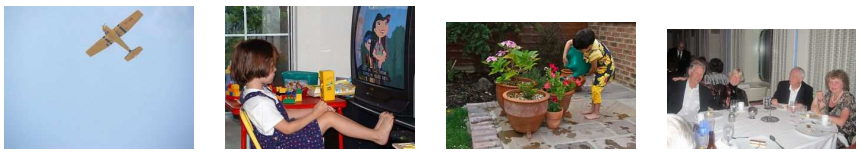
FIGURE 3.8 : Experiment II : Object's Color Identification

Image	Ground Truth	Answers
(a)	bottle red/white	83.33%
(b)	bird blue/orange	60%
(c)	bird green/dark	50%
(d)	bird red	50%

TABLE 3.2 : Results of experiment II

3.5.3 Experiment III : Scene Description

The users were given 5 minutes per image to explore four images shown in Fig. 3.9, without further information. They had then to provide a description of the image based on the audio feedback. The goal is to see if their description fits the content of the images. Five of the six subjects were able to complete this test and their results are displayed in Table 3.3. As we can observe, a simple image with an aeroplane in the air (Fig. 3.9.a) was easy to describe. The girl watching a television (Fig. 3.9.b) was ambiguous since the sound of the TV could be interpreted as a musical concert. Once again, the decision to use a person whistling sound to represent the non-speech sound of a person's class was a bad idea since some subjects will continue to confuse it with the song of a bird. While the scene of a person watering a potted plant (Fig. 3.9.c) was globally understood by most of the subjects, the image of persons around a dinning table (Fig. 3.9.b) was the most difficult to describe. Subjects claimed that the sound used to describe the dining table (a sound of dish placed on a table) was not clear enough to help them identifying the context of diner.



(a) (b) (c) (d)

FIGURE 3.9 : Experiment III : Scene Description

Image	Subject	Description
(a)	1	Airplane
	2	Sound of airplane flying in the air.
	3	Airplane dark yellow.
	4	An airplane in the air.
	5	An airplane in the air.
(b)	1	Television, I hear the music playing while exploring.
	2	A person playing music.
	3	Bird and musical instrument.
	4	A music concert with spectators.
	5	Bird in the water.
(c)	1	Sound of pouring water (probably someone watering) : potted plant.
	2	A person pouring water.
	3	Bird near a water
	4	Liquid that is poured into a glass.
	5	Water pouring.
(d)	1	Unable to understand the scene
	2	Group of persons, probably in meeting.
	3	Birds and plates on the table.
	4	Unable to understand the scene.
	5	Whistle of bird.

TABLE 3.3 : Results of experiment III

3.6 Conclusion

In this paper, we have presented a mid-level image sonification system that uses a non-speech audio dataset to describe the semantic content (20 classes of object) and a generated signal based on musical notes to describe the abstract content. We implemented our system in an Android application called Talking Image 2 and proposed an image dataset for evaluation.

The validation results showed that the subjects were generally able to identify the

objects in the image, the color of the objects and even provide a general description of the scene of the image. However, the non-speech sound used for some classes was confusing and need to be improved. We learned that the choice of sound that represents each class is highly important in a system where a vocal description is not used. Our system is a prototype that can be greatly improved using a hybrid deep-learning model on the image instead of convolutional neural network (CNN) model. Such method achieved higher detection accuracy in [95] where the bearing vibration signals were converted into time-frequency images using the continuous wavelet transform (CWT), then a CNN was used to extract intrinsic fault features from the images and feed them into a gcForest classifier.

Declarations

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no competing interests to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.
- The authors obtained a certificate of ethics from the Université de Montréal to perform the pilot study.

CHAPITRE 4

Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration

Dans ce chapitre, nous présentons notre article publié dans IEEE Transactions on MultiMedia, intitulé : **O. K. Toffa and M. Mignotte, "Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration," in IEEE Transactions on Multimedia, vol. 23, pp. 3978-3985, 2021, doi : 10.1109/TMM.2020.3035275**. Nous exposons ce dernier dans sa langue originale de publication.

Abstract

This paper presents a new approach to classify environmental sounds using a texture feature local binary pattern (LBP) and audio features collaboration. To our knowledge, this is the first time that the LBP (or its variants), which has a proven track record in the field of image recognition and classification, has been generalized for 1D and combined with audio features for an environmental sound classification task. To this end, we have generalized and defined LBP-1D and local phase quantization (LPQ)-1D on the 1-dimensional (1D) audio signal and have applied the original LBP, the variance LBP (VARLBP) and the extended LBP (ELBP) thus generated to the spectrogram of the audio signal in order to model the sound texture. We have also extensively compared these new LBP-based features to the classical audio descriptors commonly used in environmental sound classification, such as MFCC, GFCC, CQT, chromagram, STE and ZCR. We have evaluated our algorithm on ESC-10 and ESC-50 datasets using classical machine learning algorithms, such as support vector machines (SVM), random forest and k-nearest neighbor (kNN). The results showed that the LBP features outperform the classical audio features. We mix the LBP features with the audio descriptors, and our best mixed model achieves state-of-the-art results for environmental sound classification : 88.5% on ESC-10 and 64.6% on ESC-50. Those results outperform the results of methods that used handcrafted features with classical machine learning algorithms and are similar to some convolutional neural network-based methods. Although our method is not the cutting edge of the state-of-the-art methods, it is faster than any convolutional neural network methods and represents a better choice when there is data scarcity or minimal computing power.

Keywords

Environmental Sound Classification, Local Binary Pattern, Local Phase Quantization, Machine Learning, ESC-50, Audio Signal Spectrogram, SVM, Random Forest, kNN.

4.1 Introduction

Environmental sound classification (ESC) is the identification of daily sounds generated by the activities of humans or by nature, including a dog barking, fire crackling, baby crying, etc. Unlike music and speech, environmental sounds do not have a common structure since they actually have various origins and are very diverse. Their recognition and classification is one of the most important domains of audio signal processing, offering various applications : robot hearing, objectionable content detection, road surveillance, home automation and monitoring, and gunshot detection [5, 38, 70, 79, 90].

As with speech recognition, audio segmentation and other topics of audio signal processing, ESC relies on the extraction of specific and efficient audio features from time or frequency domains [14, 18, 98]. Some of those features are the short-time fundamental frequency (SFuF) [98], Gabor filters [5, 90], short-time energy (STE) [98], zero-crossing rate (ZCR) [98], constant-Q transform (CQT) [78], gammatone frequency cepstral coefficients (GFCC) [82, 88], chromagram [81], and mel-frequency cepstral coefficient (MFCC) [47]. The latter is the most commonly used for ESC. Though the spectrogram of the 1D audio signal is a 2D (time $\times$ frequency) image, it is not common to use image features to classify audio contents. With the growth and popularity of deep learning in image classification [43], numerous works have started using convolutional neural networks (CNNs), or a mix of spectrogram, MFCC and cross-recurrence plot (CRP) [12], to classify the spectrograms of sounds [64]. Nevertheless, to our knowledge, few research endeavors exploiting image descriptors have been published.

LBP is an operator or a texture analysis and characterization method based on the pixel neighborhood introduced by Ojala et *al.* [60, 61, 67]. Although this operator was originally developed for texture analysis and classification, it has become one of the most prominent and efficient texture descriptors used in many fields of image processing and computer vision, including image classification, biomedical image analysis, facial (and gender and family) recognition, and motion recognition [4, 30, 33, 35, 40, 44], to name a few. Despite its popularity, few works using LBP have been dedicated to the ESC task. Kobayashi et *al.* [39] and Ren et *al.* [70] are among the rare authors to classify

acoustic sounds using LBP. The authors in [39] applied LBP to the spectrogram, while others applied the LBP to the gammatone-like spectrogram. Our work falls within that category. We believe that we can proceed further than these previous works by directly applying the LBP to the 1D signal instead of the 2D spectrogram. First, we generalize the LBP for 1D by defining the local binary pattern-1D and local phase quantization-1D descriptors, which are directly computed from the 1D audio signal. Second, to achieve a strong characterization, we utilize a feature collaboration technique by combining the spectrogram-based 2D LBP descriptors (original LBP, variance LBP, and extended LBP) and audio features (MFCC, GFCC, ZCR, CHROMA, CQT, and STE) to successfully classify environmental sound. Third, we demonstrate that the proposed method offers the benefit of running faster than a deep neural network model that runs on a low-end GPU. The proposed method achieves interesting performance compared to the other methods.

The rest of this paper is organized as follows : Section 2 describes previous works on the ESC task. Section 3 details our method, which consists of 1D and 2D LBP descriptors as well as a combination with audio descriptors. Section 4 provides and analyses the results of our experiments, and Section 5 concludes the paper.

4.2 Related Work

ESC usually consists of the extraction of manually designed features that are then processed by a conventional classifier such as support vector machines (SVM), random forest or k-nearest neighbor (kNN) [5, 15, 38, 65, 70, 88, 90, 91, 98]. A good survey work [14] separates the methods into stationary (MFCC, STE, ZCR, MPEG-7, Gabor filters, Chroma, ZCR, STE, and linear prediction coefficients) and nonstationary (or wavelet-based) approaches such as continuous wavelet transform (CWT), fast wavelet transform (FWT), and Gaussian mixture models (GMM).

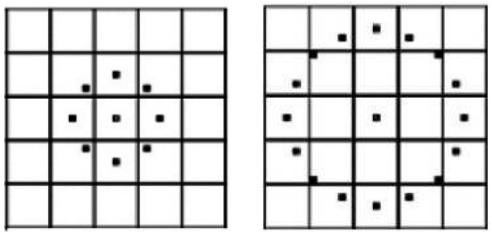
Most of the features listed in the previous paragraph are audio features. Kobayashi et *al.* [39] in 2014 were the first authors to use an image descriptor LBP to tackle a sound classification task. They enhanced the discriminative power of the LBP features with

L_2-Hellinger normalization and obtained good classification results using linear SVM on RWCP, a sound-event dataset. Their work was followed in 2017 by Ren et *al.*, [70], who applied a multichannel LBP to the gammatone spectrogram for robot hearing and demonstrated good performance on two sound-event datasets : RWCP and NTU-SEC.

With the success of deep learning in the field of image classification, many works have replaced the conventional classifiers with a CNN that is able to better learn the time-frequency features using weight-sharing and pooling. In this context, Huzaifah [32] compared CQT, CWT and short-time Fourier transform (STFT) on CNNs. Sharma et *al.* [80] implemented a deep CNN of multiple features channels composed of MFCC, GFCC, CQT and chromagram. On the other hand, the CNN is sometimes directly applied to the signal with end-to-end training [2, 86], or to its spectrogram without preliminary feature extraction [64]. The authors of SoundNet [6] trained their network by transferring discriminative knowledge from visual recognition networks into sound networks. Boddapati et *al.* [12] achieved good classification using transfer learning with image recognition networks GoogleLeNet and AlexNet. Also worth mentioning are the temporal attention mechanism [45], restricted Boltzmann machine (RBM) [72], very deep network [19], between-class technique [87], etc., to name only those among a long list of deep learning methods that have tackled ESC with success. While the CNN methods perform better than classical and conventional classifiers, they face issues of data scarcity or a lack of diversity in the datasets. These issues are usually resolved by data augmentation techniques such as time stretching, pitch shifting, and background noise and dynamic range compression [64, 73]. Despite the good accuracy achieved by CNN methods, they are time and resource consuming. For that reason, classical machine learning methods coupled with handcrafted features represents a good trade-off between accuracy and speed in the presence of a relatively small quantity of data.

4.3 Proposed Method

It is common in audio or image classification tasks to aggregate multiple features in order to achieve higher accuracy [14, 80]. This is due to the complementary structure of



P=8, R=1 P=16, R=2

FIGURE 4.1 : Neighborhood of *P* pixels and radius *R*

many features. The particularity of our model is to use two different types of features : image and audio. In this section, we will describe the image and audio features, the result of each feature when applied to audio from the ESC-50 dataset (see Figures 4.3 and 4.4) and the features collaboration technique.

4.3.1 Image Features

In this section, we present features that are commonly used in the image recognition and classification domain and their adaptation for audio classification.

4.3.1.1 LBP/VAR

LBP, as proposed by Ojala et *al.* [60, 61, 67], characterizes an image by a group of local patterns or microtexture. A local pattern is formed by encoding the difference in gray level between the pixel in the center and its neighbors, considering only the sign. The resulting binary codes of M-bits are concatenated to a decimal number. The histogram of different local patterns is used as a texture descriptor. Initially designed for a 3×3 pixel neighborhood, the LBP operator was quickly extended to a circular neighborhood of *P* pixels and radius *R* (see Fig. 4.1).

Given a center pixel c with gray level g_c, the LBP of the pixel is computed as follows :

$$LBP_c = \sum_{p-0}^{P-1} s(g_p - g_c)2^p \tag{4.1}$$

where

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{4.2}$$

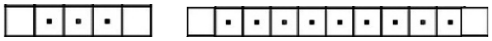
The authors in [67] identified a rotation invariant version of the LBP, but it is not useful in the context of a sound spectrogram. The LBP operator, as defined in the previous equation, is not affected by any monotonic transformation of the grayscale. It is a grayscale-invariant measure that unfortunately discards contrast. To detect the contrast, the authors in [67] proposed the *VAR* operator :

$$VAR_c = \frac{1}{P}\sum_{p-0}^{P-1}(g_p - \mu)^2 \text{ where } \mu = \frac{1}{P}\sum_{p-0}^{P-1} g_p \tag{4.3}$$

LBP and VAR are two complementary measures, and their combination or joint distribution is a very powerful measure of local image texture. The results of LBP and VAR on ESC-50 audio are provided in Figures 4.3.c and 4.3.d, respectively.

4.3.1.2 Extended LBP

Because of the popularity and simplicity of implementation of LBP, a large number of methods have been developed over the years to improve its performance : *opponent color* LBP (OCLBP) [53, 66], *multiscale color* LBP (MS-CLBP) [35], *completed* LBP (CLBP) [31], and *discriminative completed* LBP (disCLBP) [30], to name the best-known methods. The extended LBP [46], one of the most robust among these methods, extends the LBP with four complementary descriptors : the central pixel intensity (CI), the neighbor intensity (NI), the radial difference (RD) and the angular difference (AD).



P=2, R=1 P=8, R=4

FIGURE 4.2 : LBP1D neighborhood of P pixels and radius R

Their formulation is as follows :

$$CI - LBP_c = s(g_c - \mu_I) \tag{4.4}$$

relative to μ_I, the mean of the image is I.

$$NI - LBP_c = \sum_{p-0}^{P-1} s(g_p - \mu)2^p \text{ where } \mu = \frac{1}{P}\sum_{p-0}^{P-1} g_p \tag{4.5}$$

$$RD - LBP_c = \sum_{p-0}^{P-1} s(\Delta^{Rad})2^p \tag{4.6}$$

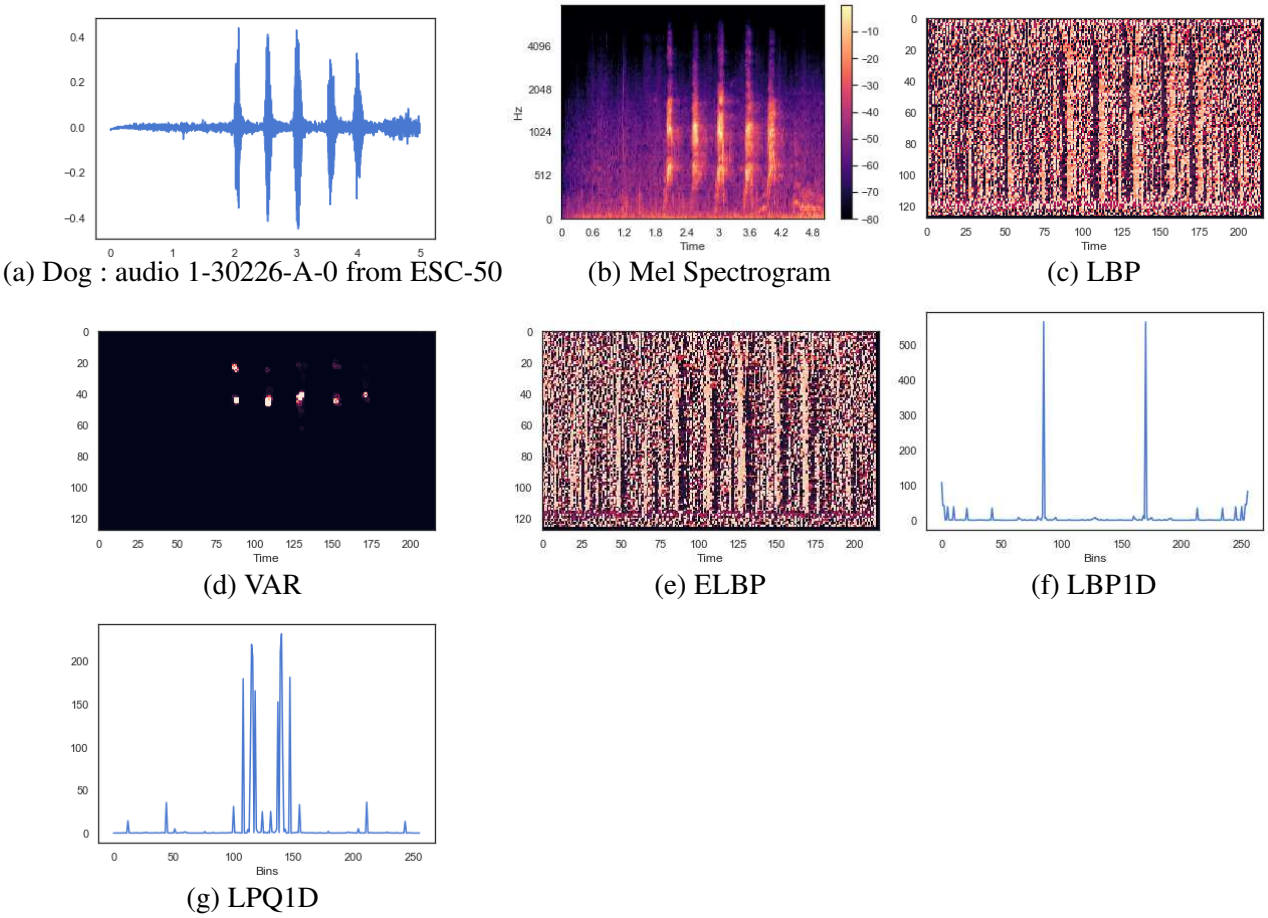
$$AD - LBP_c = \sum_{p-0}^{P-1} s(\Delta^{Ang})2^p \tag{4.7}$$

The AD-LBP is not used because it is too weak [46] and inadequate to provide a reliable and meaningful description of texture images. On the other hand, NI-RD and NI-RD-CI are strong descriptors. See the results of ELBP on ESC-50 audio in Figure 4.3.e.

4.3.1.3 LBP-1D/LPQ-1D

Unlike an image signal where the neighborhood is a circle covering an angle of 360^o, the audio signal has only two neighborhood angles : 0^o and 180^o. This allows us to define an audio texture in the time domain of a signal as a joint distribution of the pixel and its P (> 1 and even) neighbor points located in a horizontal radius R (Figure 4.2). We can then apply the equation 4.1 to the 1D signal. For example, a neighborhood of $P = 2$ and $R = 1$ presents 4 categories of patterns that cover the distinct characteristics of a signal, namely, growth, decay, minimum and maximum.

The LPQ [62], the frequency version of the LBP, consists of quantizing, in an eight-



(a) Dog : audio 1-30226-A-0 from ESC-50 (b) Mel Spectrogram (c) LBP

(d) VAR (e) ELBP (f) LBP1D

(g) LPQ1D

FIGURE 4.3 : Image Features

dimensional space, the phases of the four low-frequency coefficients of the STFT. The authors [62] showed that the low-frequency phase components are ideally invariant to centrally symmetric blur and that the LPQ is more efficient than the LBP in the presence of noise. The LPQ1D simply involves applying the same reasoning to the 1D signal. See the results of the histograms of LBP1D and LPQ1D on ESC-50 audio in Figures 4.3.f and 4.3.g, respectively.

4.3.2 Audio Features

For comparison reasons and to achieve image and audio feature collaboration, we consider the following descriptors that are usually used for audio classification : MFCC, GFCC, STE, ZCR, CQT, and chromagram [14, 32, 80, 98].

4.3.2.1 MFCC

The mel-frequency cepstral coefficients were developed to resemble the human auditory system and have been successfully used in music modeling, speech recognition and audio classification [47, 65]. This work employs a short-term spectral-based feature computed from the mel spectrogram of the sound using a defined number of filters. In our experiment, we used 128 bands. See Fig. 4.4.a for an example of MFCC.

4.3.2.2 GFCC

The gammatone filterbank attempts to approximate the human auditory system as the MFCC, using gamma distributions and sinusoidal tones [82, 88]. GFCC is known for its strong capability to represent impulsive signal classes such as transient sounds and offers complementary use with MFCC [14]. In our experiment, we used 128 bands for the GFCC ; an example is displayed in Fig. 4.4.b.

4.3.2.3 CQT

The constant-Q transform (CQT) is a technique that transforms a time-domain signal $x(n)$ into the time-frequency domain so that the center frequencies of the frequency bins are geometrically spaced and their Q-factors are all equal [32, 78]. CQT exhibits good results in the audio classification task [32]. In our experiment, we used 128 bins ; see Fig. 4.4.c for an example.

4.3.2.4 Chromagram

The chroma is used to characterize the sound by decomposing the signal into a number of pitch class profiles [81]. It captures the harmonic and melodic characteristics of the music. In our experiment, we used 128 pitch class profiles; see Fig. 4.4.d for an example.

4.3.2.5 STE

The short-time energy of the signal is the energy of the signal computed over a window of time. It enables a convenient representation of the amplitude variation over time. The STE also permits detection of the periodicity and the silence of a signal and characterizes the harmony of music [98]. It has the particularity of being able to be computed using time space or frequency space. See Fig. 4.4.e for an example of STE.

4.3.2.6 ZCR

Unlike most of the previous audio features that are frequency-based, the zero-crossing rate is a time-based feature. It permits determination of whether successive samples have different signs. The rate at which zero-crossing occurs is a simple measure of the frequency content of a signal. It permits separation of a vocal sound from a nonvocal sound [98]. An example of ZCR is shown in Fig. 4.4.f.

4.3.3 Features Collaboration

The features collaboration technique consists of the exploitation of multiple features in order to construct a strong discriminator. Each feature is usually designed to characterize one of the many temporal and spectral content properties of an audio signal, including the pitch, frequency, energy, loudness, timbre, amplitude, etc. While the ZCR, for example, can easily discriminate between a vocal and a nonvocal sound, the chromagram can only capture the harmony and melody of music. It very early became obvious for many studies to aggregate, concatenate or fuse multiple features in order to obtain a

majority of distinguishable and complementary features that can classify all categories of environmental sounds. Muhammad *et al.* [57] aggregated MPEG-7 audio low-level descriptors together with conventional MFCC and demonstrated a significant improvement in the recognition performance of the proposed system over MFCC or full MPEG-7. In [14], the authors showed that combining MFCC with GFCC is very successful since they complement each other. In [65], PicZak concatenated ZCR and MFCC to successfully classify the ESC-50 dataset. Sharma *et al.* [80] achieved state-of-the-art performance by combining MFCC, GFCC, chromagram and CQT in a multichannel neural network coupled with an attention network. In our case, we will concatenate the image and audio features in order to cover more characteristics available in all categories of environmental sound, and we demonstrate in the next section that such a combination achieves better performance than using a sole type of feature, namely, image or audio.

4.4 Experimental Results

4.4.1 Datasets

The ESC-50 dataset proposed by Piczak [65] in 2015 consists of 2000 labeled environmental sounds split equally among 50 classes (40 records per class). Each record has a duration of 5 seconds, with a sampling rate of 44.1 kHz. The 50 classes are divided into 5 categories (10 per category) : animal sounds, natural soundscapes and water sounds, human (nonspeech) sounds, interior/domestic sounds, and exterior/urban sounds. The ESC-50 dataset is popularly used in the ESC task [6, 12, 32, 45, 64, 65, 80, 86]. For rapid testing, Piczak also proposed an ESC-10 subset of 400 records and 10 classes, with good separability between classes.

4.4.2 Setup

In our experiments, we segment the audio into multiple windows of 2048 samples, each with an overlapping 1024 samples. A wideband (2048) segmentation has been proved to offer minor advantages over narrowband (1024) [32] and permits having descrip-

tors with reduced size in time/frequency dimensions (e.g., 216 for a signal of 5 seconds). A 3×3 neighborhood ($P = 8$ and $R = 1$) is used for the 2D image features. For LBP1D and LPQ1D, we use $P = 8$ and $R = 4$. In both 1D and 2D, a descriptor of 256 bins is obtained. The histograms of the image descriptors are computed for each window and then concatenated to form a final descriptor. The audio features are 128 bins, except for the STE and ZCR, which are both 1 bin. The implementation is performed using the librosa package[1] and the gammatone library[2].

Each descriptor ends with a size of $nbins \times 216$ for an audio signal of 5 seconds, which is too high. One way to reduce this size is to use multidimensional scaling (MDS), principal component analysis (PCA), independent component analysis (ICA) or any other dimension reduction technique. The drawback of those methods is time consumption, so we prefer to compute a simple mean and a standard deviation along the time axis in order to reduce the size of the descriptor to $nbins \times 2$. We then process the descriptors using conventional machine learning algorithms SVM, random forest and kNN with a 5-fold cross-validation regime. The extraction of all of the descriptors requires 20 seconds per audio segment on an Intel(R) Core(TM) i5-7440HQ 2.80 GHz CPU without any GPU acceleration. Each algorithm requires only a few seconds to run the 5-fold validation and display an average accuracy result. The results are presented and commented upon in the next sections of the paper. .

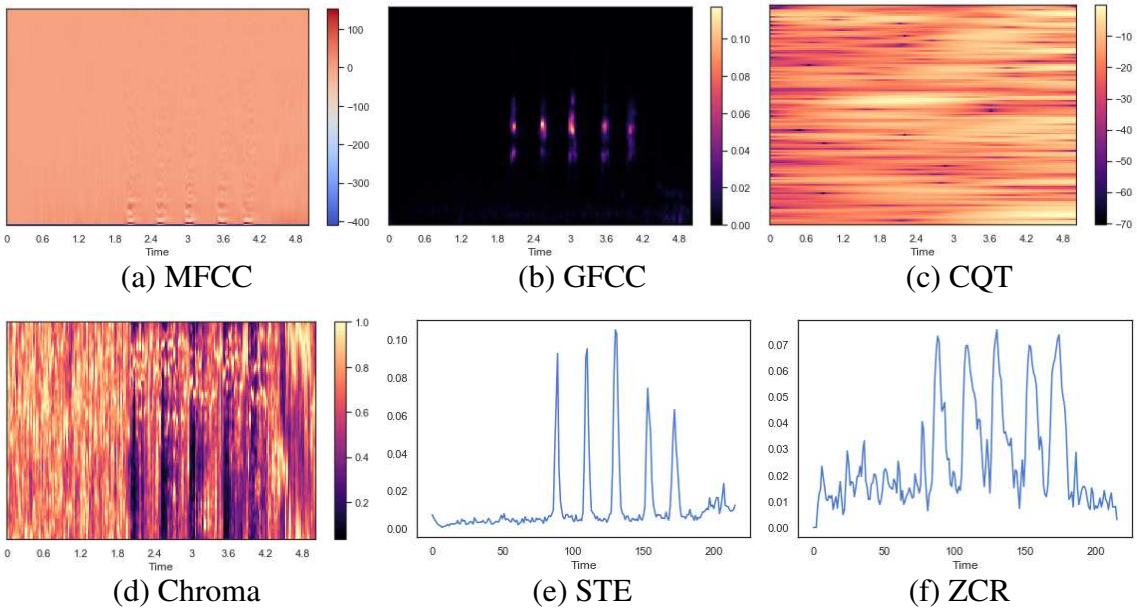
4.4.3 One Feature

In this section, we present the accuracy results of each descriptor on both the ESC-10 and ESC-50 datasets. Presented in bold are the best three accuracy values per algorithm.

The results on ESC-10 are presented in Table 4.1. The best two results using the kNN method are obtained for LBP (65.9%) and ELBP (65%). The MFCC, the first audio descriptor to perform well, shows a result of 64% at the third position. Using random

[1] librosa : v0.7.1 library by B. McFee et al., DOI : http ://dx.doi.org/10.5281/zenodo.12714 [Accessed Aug. 5, 2015]

[2] gammatone : https ://github.com/detly/gammatone.git.



(a) MFCC (b) GFCC (c) CQT

(d) Chroma (e) STE (f) ZCR

FIGURE 4.4 : Audio Features

forest, the GFCC presents the best accuracy of 77%, followed by LBP and ELBP, both at 73.5%, and MFCC at 72.2%. With the SVM method, the LBP (79.7%) is in the first position, followed by the MFCC (77%) and the ELBP (76.5%). Note that the accuracy of 79.7% for the LBP is the best classification score obtained here and so far. The LBP remains robust, regardless of the machine learning method, and then performs better than the rest of the descriptors. The ELBP performance is not very far from that of LBP, which is normal since the implementations of these two operators are only slightly different. MFCC is next, followed by the GFCC, which has unfortunately vanished with the SVM. The 1D descriptors LBP1D and LPQ1D, which we introduced in Section 4.3.1.3, turn out to be weak descriptors. This result is somehow expected because they only exploit the 1D variation of the signal, while the other features are based on the spectrogram, which is a more complete time/frequency representation of the signal. However, LBP1D and LPQ1D did perform better than other classic audio features such as STE, ZCR and CQT. LPQ1D outperforms LBP1D since it is more robust to noise [62]. Although the

Features	kNN	RF	SVM
LBP1D	50.7	61.5	58
LPQ1D	53	66.4	65.5
LBP	**65.9**	**73.5**	**79.7**
VAR	52	42.2	56.7
ELBP	**65**	**73.5**	**76.5**
STE	49.7	44.2	46.7
ZCR	44.7	38.2	22.9
MFCC	**64**	**72.2**	**77**
GFCC	61.7	**77**	30.7
CQT	29	29.2	20.2
CHROMA	53.7	61.2	53.7

TABLE 4.1 : ESC-10 : Results of classification with one feature

result of VAR is weak, we will not devote much attention to it because it is normally used conjointly with the LBP. We will study its impact in the next section.

Table 4.2 shows the accuracy results on ESC-50. The accuracy is reduced by approximately 20% for all features because the number of classes has increased. We can observe that the LBP and ELBP remain the best descriptors regardless of the machine learning algorithm. They are followed by MFCC, GFCC, LPQ1D, LBP1D and the rest of the descriptors. The overall best result of 54.5% on the ESC-50 dataset is obtained for a single LBP descriptor using SVM.

Those results outperform the top values of 72.7 on ESC-10 and 44.3 on ESC-50 that were originally obtained by Piczak [65] on his datasets using MFCC-ZCR with random forest and SVM.

4.4.4 Multiple Features

We showed in the previous section that LBP and ELBP features that are image descriptors offer better performance on the ESC-10 and ESC-50 datasets than the strongest commonly used classic audio descriptors such as MFCC or GFCC. Though the obtained results are very interesting, the LBP-based descriptors remain handcrafted features

Features	kNN	RF	SVM
LBP1D	17.6	23.6	26.7
LPQ1D	17.6	27.2	29.5
LBP	**35.7**	**45.9**	**54.5**
VAR	13.3	13.4	15.4
ELBP	**33.5**	**43.7**	**53.7**
STE	10.1	9.2	9.2
ZCR	9.4	8.1	6.1
MFCC	22.6	**43.7**	**46.2**
GFCC	**26.3**	**42.6**	10.3
CQT	5.1	7.3	5.8
CHROMA	16.2	22.5	17.4

TABLE 4.2 : ESC-50 : Results of classification with one feature

and are not able to fully extract all of the patterns that can fully characterize a signal. To correct this drawback, one of the solutions is to combine multiple features that will complementarily qualify the signal, which is called feature collaboration.

We start by combining the image descriptors together, then the audio descriptors together, and finally the image descriptors with the audio descriptors.

The results of the feature combination on ESC-10 are presented in Table 4.3. The best accuracy is obtained for the kNN algorithm by the aggregation of the four strongest features LBP-ELBP-MFCC-GFCC (68%), followed by LBP-VAR-ELBP-MFCC-GFCC (67.7%), LBP-VAR (67.5%) and LBP-VAR-ELBP (67.5%). We can notice the natural complementarity between LBP (strong descriptor) and VAR (weak descriptor), as demonstrated by the authors in [67], which permits us to obtain a good result near the best obtained with 4 strong descriptors. The LBP-VAR even presents the best result of 84.9% on SVM. The combination of LBP-VAR with ELBP does not make a great difference since both algorithms are similar. However, associating the LBP-based image descriptors with the 1D descriptors and the audio descriptors, LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC, with random forest provides the global best result of **88.2%**.

The results on ESC-50 presented in Table 4.4 show the same behaviors of the features as those observed on ESC-10. The combination of LBP, VAR and ELB provides the

best accuracies with the kNN and SVM. The topmost result of **64.6%** is obtained using random forest and the mix of the three types of descriptors LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA. Note that this result is similar to that obtained using a convolutional network in [64].

4.4.5 Analysis

Note that on both datasets, associating only image descriptors or audio descriptors does not yield the best accuracy. On ESC-50, for example, the accuracy values achieved with such combinations are all under 60%. However, as soon as we mix an image descriptor with an audio descriptor, breaking the 60% mark becomes possible. The audio and image descriptors can therefore be considered complementary.

We also observe that the random forest algorithm is more sensitive to the mixture of features. This is because SVM and kNN lose their power when the feature vector size is increased. The weak descriptors such as STE, ZCR and CQT perform poorly. They exert minor impacts, or sometimes negative impacts, when combined with the other descriptors.

Our best model performs well compared with the state of the art, as shown in Table 4.5. It outperforms any model based on conventional machine learning and offers similar performance to some deep learning methods. Most of the research in the ESC field is now oriented to deep learning methods, but our model represents a good alternative in the presence of limited computing power or a lack of data. It requires only a few hours to extract the descriptors and run a model, while a few days or weeks would be required for a deep learning method.

To prove this fact, we train on a low-end GPU the stream that processes the spectrogram in the three-stream network available in [45], which is the leader in Table 4.5. Such a network, presented in Table 4.6, is very deep and is composed of 18 levels of 2D convolution, max pooling and dense layers for a total of 88,143,882 trainable parameters. The input size is 512x384. Although it is not visible in Table 4.6, each convolution layer is followed by a batch normalization and a ReLU. We use a data augmentation

technique [73] (time stretch, time shift, noise, pitch shift) on ESC-10 that increases the number of samples from 400 to 2,000. We train the network on the ESC-10 dataset on an NVIDIA GeForce 930MX with 2 GB RAM, a low-end GPU, which has specifications that are comparable with low-power AI systems available on the market : NVIDIA Jetson Nano (Quad Cortex A57 @ 1.43 GHz, 4 GB RAM) and Google Edge TPU (Quad Cortex-A53, Cortex-M4F @ 2 GB RAM). The implementation is performed with the TensorFlow [1] library with a minimal batch size of 1, but the *resource exhausted error* is triggered because of the limited GPU memory of 2 GB. Since the test is only for performance measurement and not accuracy, we divide by 8 the number of filters at each level in order to obtain a reduced network of 2,700,922 trainable parameters. It takes 16 minutes to train the 5-fold ESC-10 for 1 epoch, *i.e.*, 26.6 hours for 100 epochs. In the paper [45], three networks were used, with an additional fourth network for the attention, and were trained for more than 100 epochs. Many weeks will be required to train such a system on a low-end GPU with sufficient memory, and months to train ESC-50. If we eliminate the data preparation time in both cases, this time is enormous compared with the 20 seconds required to run SVM or random forest on the same dataset. This fact shows that although low-power AI systems are available for the lowest price on the market, training a network that is slightly deep is not affordable for everyone. Those GPUs are not designed for training, but rather for inference and supporting limited transfer learning.

4.5 Conclusion

In this paper, we presented a new ESC method that exploits LBP, a 2D texture classification descriptor. The LBP method was applied to the signal in one dimension as well as on the sound texture represented by the spectrogram. The results showed that LBP features outperform the audio descriptors and are more efficient on the two datasets. We showed that the combination of the audio descriptors with the LBP achieves state-of-the-art results using a simple machine learning classifier such as SVM or random forest. It also performs well compared with some CNN-based methods. This approach is faster

than deep learning and represents a good alternative when there is data scarcity or minimal computing power. Our method has many advantages but is not the leader of the state-of-the-art methods. It can be improved by using CNN with a multichannel descriptor consisting of a mixture of LBP and audio features. This improvement represents the topic of our future research.

Features	kNN	RF	SVM
LBP-VAR	**67.5**	77.2	**84.9**
LBP-VAR-ELBP	**67.5**	78.2	**84.2**
LBP-VAR-ELBP-LBP1D	52.5	80	72.7
LBP-VAR-ELBP-LBP1D-LPQ1D	57.2	82.4	68.7
MFCC-GFCC	64	83.5	77
MFCC-GFCC-CHROMA	64.2	84.5	77.5
MFCC-GFCC-CHROMA-STE	64.2	84.2	77.9
MFCC-GFCC-CHROMA-STE-ZCR	64.2	84.5	77.9
MFCC-GFCC-CHROMA-STE-ZCR-CQT	62.2	85	79.2
LBP-ELBP-MFCC-GFCC	**68**	86.9	**82.5**
LBP-VAR-ELBP-MFCC-GFCC	**67.7**	86.2	81.7
LBP-VAR-ELBP-LBP1D-MFCC-GFCC	64.5	87.2	81.2
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC	64.2	**88.2**	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA	64.2	86.9	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE	64.2	**87.5**	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR	64.2	**87.5**	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR-CQT	65	**87**	78.4

TABLE 4.3 : ESC-10 : Results of classification with multiple features

Features	kNN	RF	SVM
LBP-VAR	**33.3**	47.4	**54.5**
LBP-VAR-ELBP	**34.3**	49.6	**55.8**
LBP-VAR-LBP1D-ELBP	19	54.1	44.6
LBP-VAR-LBP1D-LPQ1D-ELBP	20.9	55.6	39.6
MFCC-GFCC	22.6	51.9	46.2
MFCC-GFCC-CHROMA	22.6	53.2	46.6
MFCC-GFCC-CHROMA-STE	22.7	53.6	46.6
MFCC-GFCC-CHROMA-STE-ZCR	22.7	52.9	46.6
MFCC-GFCC-CHROMA-STE-ZCR-CQT	20.7	54.1	43.6
LBP-ELBP-MFCC-GFCC	23.4	62.3	54.3
LBP-VAR-ELBP-MFCC-GFCC	23.5	61.4	**55**
LBP-VAR-ELBP-LBP1D-MFCC-GFCC	25	63.4	51.6
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC	24.7	63.2	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA	24.7	**64.6**	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE	24.7	**64.1**	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR	24.7	63.4	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR-CQT	**25.6**	**64.3**	47.2

TABLE 4.4 : ESC-50 : Results of classification with multiple features

Algo	ESC-10	ESC-50
Human [65]	95.7	81.3
Attention Network [45]	94.2	84
EnvNet [86]	86.8	66.4
EnvNet2 [87]	88.8	81.6
PiczakCNN [64]	90.2	64.5
AlexNet [12]	86	65
GoogleNet [12]	86	73
Piczak kNN [65]	66.7	32.2
Piczak RF [65]	72.7	44.3
Piczak SVM [65]	67.5	39.6
CNN+CQT+CWT+MEL [32]	-	56.4
SoundNet [6]	92.2	74.2
Our best Method	**88.5**	**64.5**

TABLE 4.5 : Best proposed model score and comparison with the state of the art

Layer	Filter Size	No. of Filters
Conv	3x3	128
Conv	3x3	128
MaxPooling	4x3	-
Conv	3x3	128
Conv	3x3	256
MaxPooling	4x4	-
Conv	3x3	256
Conv	3x3	512
MaxPooling	2x2	-
Conv	3x3	512
MaxPooling	2x2	-
Conv	3x3	1024
MaxPooling	2x2	-
Conv	3x3	1024
Conv	3x3	2048
MaxPooling	2x2	-
Dense	-	4096
Dense	-	4096

TABLE 4.6 : Spectrogram stream of the multistream with attention network [45]

CHAPITRE 5

Conclusion Générale et Perspectives

Dans cette thèse, nous avons présenté des approches de solutions à deux problèmes du domaine multimedia que sont la sonification d'image et la classification de sons environnementaux.

Dans un premier temps nous avons présenté un nouveau système de sonification d'image qui permet de traduire d'une manière intuitive l'information visuelle spatiale locale en sons. Le système proposé utilise un ensemble de caractéristiques visuelles sur le contenu de l'image à différents niveaux d'abstraction et perceptuellement significatifs. Il utilise les concepts de timbre, de volume, de hauteur, de rythme et différents effets de distorsion pour traduire l'apparence de chaque région pré-segmentée individuelle de l'image dans le domaine audio. Le système proposé nous permet de facilement localiser différentes régions et classer les régions dans les régions artificielles et naturelles, parfois avec reconnaissance d'objets. Les résultats de validation ont montré que bien que les sujets n'aient pas *l'oreille musicale* et n'aient eu aucune session d'entraînement, dans certains cas, ils ont pu détecter des objets dans les images et des groupes d'images en fonction des caractéristiques visuelles traduites en sons. Nous avons également montré que les utilisateurs pouvaient améliorer leurs performances sur le système avec plus de pratique. Les résultats sont prometteurs puisque les personnes déficientes visuellement (*oreilles musicales*) avec un certain entraînement pourront sûrement faire mieux. Contrairement aux méthodes existantes [10, 34, 50, 51, 76, 96, 97], qui traduisent seulement un petit nombre de caractéristiques de l'image en sons, notre méthode a l'avantage de traduire plusieurs caractéristiques visuelles à la fois, ce qui permet à l'utilisateur d'avoir plus de détails sur le contenu de l'image à travers une expérience sonore plus riche. Malheureusement, nous n'avons pas pu valider nos résultats sur des personnes malvoyantes, ce qui était le but premier de nos recherche.

Nous avons ensuite montré qu'il est possible de développer un modele de sonification complète automatique en complémentant la sonification de bas niveau proposée

précédemment avec un système de sonification de niveau intermédiaire qui utilise un ensemble de données audio non-vocal pour décrire le contenu sémantique (20 classes d'objets). Nous avons implémenté notre système dans une application Android appelée TalkingImage 2 (Annexe 1) et proposé une base de données pour l'évaluation. Les résultats de la validation ont montré que les sujets étaient généralement capables d'identifier les objets dans l'image, la couleur des objets et même fournir une description générale de la scène de l'image. Contrairement aux méthodes existantes offrant la même expérience complète de sonification, notre méthode n'est pas manuelle car elle ne requiert aucune analyse du contenu de l'image par un humain ou un support haptique. Cependant, le son non-vocal utilisé pour certaines classes était déroutant et devrait être amélioré. Nous avons appris que le choix du son qui représente chaque classe est très importante dans un système où une description vocale n'est pas utilisée. Une autre limite à cette contribution est le nombre peu élevé de participant et le fait que nous n'ayons pas pu conduire un test pilote avec des personnes malvoyantes.

Finallement, nous avons présenté une nouvelle méthode de classification de sons environnementaux (ESC) qui exploite les motifs locaux binaires (LBP), un descripteur de classification de texture ayant fait ses preuves dans le domaine de l'imagerie. La méthode LBP a été appliquée sur le signal à une dimension (1D) ainsi que sur la texture sonore représentée par le spectrogramme. Les résultats ont montré que les fonctionnalités LBP surpassent les descripteurs audios et sont plus efficaces sur les deux bases de données d'évaluation. Nous avons montré que la combinaison des descripteurs audio avec le LBP obtient de bons résultats à l'aide d'un simple apprentissage automatique avec un classificateur tel que les machines à vecteurs de support ou les forêts d'arbres décisionnels. Il se comporte également bien par rapport à certaines méthodes basées sur les réseaux de neurones convolutionnels. Cette approche est plus rapide que l'apprentissage profond et représente une bonne alternative en cas de rareté des données ou de puissance de calcul minimal.

Perspectives

La sonification d'image est un domaine encore très embryonnaire qui offre beaucoup de perspectives. Lors de nos travaux sur le modèle de sonification de niveau intermé-

diaire, nous avons remarqué que le choix des audios est très important. Il serait ainsi intéressant d'envisager de développer de nouvelles bases de données d'audios qui sont plus explicites et qui sont plus représentatives des objets qu'ils décrivent. On pourra par exemple intégrer de l'apprentissage profond pour déterminer les émotions dans l'image afin d'y mapper des sons de joies ou de tristesse. Il faudra augmenter le nombre de classes d'objets car le modèle actuel couvre seulement 20 catégories d'objets. De plus, il serait intéressant de complémenter les sons non-vocaux avec une description vocale. Il faudra conduire un test pilot avec des personnes malvoyantes et un nombre de participants élevé. L'application TalkingImage 2 demeure un prototype qui peut être grandement amélioré et intégré dans une application de partage de photos accessibles aux non-voyants.

Au niveau de la classification de sons environnementaux, notre méthodologie peut être améliorée en utilisant un réseau convolutionnel avec un descripteur multicanal composé d'un mélange de LBP et de caractéristiques audio.

Plus généralement, sans concerner directement et explicitement le problème de la sonification d'image pour non-voyants, les pistes de recherches intéressantes à explorer, que ce travail nous amène à considérer, pourraient concerner, par exemple :

La sonification d'images médicales ultrasonore, tomographie à émission de positrons (PET) ou fourni par une technique de résonance magnétique MRI possiblement fonctionnelle (IRMf) de patients pour la détection (segmentation, classification ou reconnaissance) de zones texturales complexes et difficile à discerner dans l'image visuellement (ou par traitement d'images) et associées à d'éventuelles tumeurs cancéreuses, précancéreuses ou pathologiques. En effet, le système visuel humain (SVH), même le plus exercé, ne sait distinguer qu'une centaine de niveaux de gris alors que son système auditif peut distinguer aisément plusieurs milliers de textures sonores différentes. Cette sonification d'image médicale permettrait de détecter plus facilement des tumeurs malignes cancéreuses ou des zones (difficilement discernables) dans les images IRMf associées à un trouble neurologique progressif (Parkinson) ou associées à un trouble cérébral (Alzheimer).

La généralisation de descripteurs de texture d'images pourrait être étendue au do-

maine de l'audio pour la détection, reconnaissance ou classification de sons complexes dans le domaine de la classification du son (ou de la parole). Inversement la generalisation de descripteur utilisé dans l'audio pourrait être généralisée en imagerie et aussi, bien sûr, la combinaison efficace de ceux-ci, issues des deux domaines (visuels et acoustiques), serait une stratégie qui serait intéressante à poursuivre pour d'autres applications.

Finalement, la combinaison de caractéristiques visuelles et audio permettrait de faire une meilleure catégorisation ou interprétation sémantique des vidéos en fonction de leur contenu multimédia.

BIBLIOGRAPHIE

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow : A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), USENIX Association*, pages 265–283, 2016.

[2] S. Abdoli, P. Cardinal, and A. L. Koerich. End-to-end environmental sound classification using a 1d convolutional neural network. *Expert Systems with Applications*, 136 :252–263, Dec 2019.

[3] A. Ahmad, S. G. Adie, M. Wang, and S. A. Boppart. Media 1 : Sonification of optical coherence tomography data and images. *Optics Express*, May 2010.

[4] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila. Recognition of blurred faces using local phase quantization. *Proceedings of International Conference on Pattern Recognition*, page 1–4, 2008.

[5] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi. Automatic detection and classification of audio events for road surveillance applications. *Sensors*, 18(6), Jun 2018.

[6] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet : Learning sound representations from unlabeled video. In *30th Conference on Neural Information Processing Systems*, 2016.

[7] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, and S. Yaacob. A stereo image processing system for visually impaired. *International Journal of Information, Control and Computer Sciences*, 2(9) :1–10, 2008.

[8] M. Banf and V. Blanz. A modular computer vision sonification model for the visually impaired. In *Proceedings of Eighteenth Meeting of the International Conference on Auditory Display (ICAD 2012)*, pages 121–128, 2012.

[9] M. Banf and V. Blanz. Sonification of images for the visually impaired using a multi-level approach. In *Proceedings of the 4th Augmented Human International Conference (AH '13)*, pages 162–169, 2013.

[10] M. Banf, R. Mikalay, B. Watzke, and V. Blanz. Picturesensation - a mobile application to help the blind explore the visual world through touch and sound. *Journal of Rehabilitation and Assistive Technologies Engineering*, 3, 2016.

[11] J.I. Bartolome, L.C. Quero, K. Sunhee, M.Y. Um, and J. Cho. Exploring art with a voice controlled multimodal guide for blind people. In *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '19, pages 383–390, New York, NY, USA, 2019. Association for Computing Machinery.

[12] V. Boddapatia, A. Petefb, J. Rasmussonb, and L. Lundberga. Classifying environmental sounds using image recognition networks. In *International Conference on Knowledge Based and Intelligent Information and Engineering Systems*, pages 6–8, Sep 2017.

[13] M. Capp and P. Picton. The optophone : an electronic blind aid. *Engineering Science and Education Journal*, 9(3) :137–143, 2000.

[14] S. Chachada and C.-C. Jay Kuo. Environmental sound recognition : A survey. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013.

[15] S. Chandrakala and S.L. Jayalakshmi. Generative model-driven representation learning in a hybrid framework for environmental audio scene and sound event recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2019.

[16] L-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 833–851, 2018.

[17] B. Chidester and Minh Do. Assisting the visually impaired using depth inference on mobile devices via stereo matching. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, July 2013.

[18] Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo. Environmental sound recognition with time–frequency audio features. *ieee Transactions on Audio, Speech and Language Processing*, 17, aug 2009.

[19] W. Dai, C. Dai, S. Qu, J. Li, and S. Das. Very deep convolutional neural networks for raw waveforms. In *ieee International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017.

[20] N. Degara, A. Hunt, and T. Hermann. Interactive sonification [guest editors' introduction]. *IEEE MultiMedia*, 22(1) :20–23, Jan 2015.

[21] G. Dubus and R. Bresin. A systematic review of mapping strategies for the sonification of physical quantities. *PLoS ONE*, 8(12) :e82491, december 2013.

[22] A. D. N. Edwards, G. Hines, and A. Hunt. Segmentation of biological cell images for sonification. In *2008 Congress on Image and Signal Processing*, volume 2, pages 128–132, May 2008.

[23] T. Elbert, A. Sterr, B. Rockstroh, C. Pantev, M.M. Müller, and E. Taub. Expansion of the tonotopic area in the auditory cortex of the blind. *Journal of Neuroscience*, 22(22) :9941, Nov 2002.

[24] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge : A retrospective. *International Journal of Computer Vision*, 111(1) :98–136, January 2015.

[25] S.A. Fausti, D.A. Erickson, R.H. Frey, B.Z. Rappaport, and M.A. Schechter. The effects of noise upon human hearing sensitivity from 8000 to 20 000 hz. *The Journal of the Acoustical Society of America*, 69(5) :1343, Jan 1981.

[26] K. Franklin and J.C. Roberts. Pie chart sonification. In Ebad Banissi and et al, editors, *Proceedings Information Visualization (IV03)*, pages 182–196. IEEE Computer Society, July 2003.

[27] T. Gotzelmann. Visually augmented audio-tactile graphics for visually impaired people. *ACM Trans. Access. Comput.*, 11(2), June 2018.

[28] V. Goudarzi. Designing an interactive audio interface for climate science. *IEEE MultiMedia*, 22(1) :41–47, Jan 2015.

[29] F. Gougoux, F. Lepore, M. Lassonde, P. Voss, R.J. Zatorre, and P. Belin. Pitch discrimination in the early blind. *Nature*, 430(6997) :309, Jul 2004.

[30] Y. Guo, G. Zhao, and M. Pietikäinen. Discriminative features for texture description. *Pattern Recognition*, 45(10) :3834–3843, Oct 2012.

[31] Z. Guo, L. Zhang, and D. Zhangr. A completed modeling of local binary pattern operator for texture classification. *ieee Transaction on Image Processing*, 19 :1657–1663, jun 2010.

[32] M. Huzaifah. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *arXiv :1706.07156*, Jun 2017.

[33] D. K. Iakovidis, E. G. Keramidas, and D. Maroulis. Fuzzy local binary patterns for ultrasound texture characterization. *Image Analysis and Recognition (Lecture Notes in Computer Science)*, 5112 :750–759, 2008.

[34] K. Ivan and O. Radek. Hybrid approach to sonification of color images. In *Proceedings of the International Conference on Convergence and Hybrid Information Technology*, 2008.

[35] A. Joshi and A. K. Gangwar. Color local phase quantization (clpq)- a new face representation approach using color texture cues. *International Conference on Biometrics*, may 2015.

[36] E. Jovanov, K. Wegner, V. Radivojevic, D. Starcevic, M. S. Quinn, and D. B. Karron. Tactical audio and acoustic rendering in biomedical applications. *IEEE Trans. Information Technology in Biomedicine*, 3(2) :109–118, 1999.

[37] S.K. Kane, M.R. Morris, and J.O. Wobbrock. Touchplates : Low-cost tactile overlays for visually impaired touch screen users. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, New York, NY, USA, 2013. Association for Computing Machinery.

[38] M. J. Kim and H. Kim. Audio-based objectionable content detection using discriminative transforms of time-frequency dynamics. *IEEE Transactions on Multimedia*, 14(5) :1390–1400, Oct 2012.

[39] T. Kobayashi and J. Ye. Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014.

[40] P. Král, A. Vrba, and L. Lenc. Enhanced local binary patterns for automatic face recognition. In *International Conference on Artificial Intelligence and Soft Computing*, volume 11509, pages 27–36. Lecture Notes in Computer Science, 2019.

[41] N. Kwon, Y. Koh, and U.Oh. Supporting object-level exploration of artworks by touch for people with visual impairments. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 600–602, New York, NY, USA, 2019. Association for Computing Machinery.

[42] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521 :436–444, 2015.

[43] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521 :436–444, May 2015.

[44] Z. Lei and S. Z. Li. Fast multi-scale local phase quantization histogram for face recognition. *Proceedings of International Conference on Pattern Recognition*, 33(13) :1761–1767, 2012.

[45] X. Li, V. Chebiyyam, and K. Kirchhoff. Multi-stream network with temporal attention for environmental sound classification. In *Interspeech*, pages 3604–3608, 2019.

[46] L. Liu, L. Zhao, Y. Long, G. Kuang, and P. Fieguth. Extended local binary patterns for texture classification. *Image and Vision Computing*, 30(2) :86–99, Feb 2012.

[47] B. Logan. Mel frequency cepstral coefficients for music modeling. *ISMIR*, 270 :1–11, Oct 2000.

[48] C. Loughlin. *Sensors for Industrial Inspection*. Springer Science & Business Media, December 2012.

[49] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. of the 8th International Conference on Computer Vision (ICCV)*, volume 2, pages 416–423, Vancouver, British Columbia, Canada, July 2001.

[50] A. C. G. Martins, R. M. Rangayyan, and R. A. Ruschioni. Audification and sonification of texture in images. *J. Electronic Imaging*, 10(3) :690–705, 2001.

[51] S. Matta, D. K. Kumar, X. Yu, and M. Burry. An approach for image sonification. In *First International Symposium on Control, Communications and Signal Processing, 2004.*, pages 431–434, 2004.

[52] P. B. L. Meijer. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering*, 39(2) :112–121, Feb 1992.

[53] T. Mäenpää and M. Pietikäinen. Computer vision using local binary patterns-classification with color and texture : jointly or separately. *Pattern Recognition*, 37 :1629–1640, 2004.

[54] M. Mignotte. *Segmentation d'images sonar par approche Markovienne hierarchique non supervisée et classification d'ombres portées par modèles statistiques*. PhD thesis, Ecole Navale (French Naval academy), Université de Brest, July 1998.

[55] M. Mignotte. A label field fusion model with a variation of information estimator for image segmentation. *Information Fusion*, 20 :7–20, 2014.

[56] M.R. Morris, J. Johnson, C.L. Bennett, and E.Cutrell. Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–11, New York, NY, USA, 2018. Association for Computing Machinery.

[57] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda. Environment recognition using selected mpeg-7 audio features and mel-frequency cepstral coefficients. In *Digital Telecommunications (ICDT), 2010 Fifth International Conference on. IEEE*, pages 11–16, June 2010.

[58] A.H. Munsell. A pigment color system and notation. *Journal of Psychology*, 23(2) :236–244, April 1912.

[59] U. Oh, H. Joh, and Y. Lee. Image accessibility for screen reader users : A systematic review and a road map. *Electronics*, 10(8), 2021.

[60] T. Ojala, M. Pietikäinen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. *Proceedings of International Conference on Pattern Recognition*, 1 :582–585, 1994.

[61] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1) :51–59, 1996.

[62] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. *International Conference on Image and Signal Processing*, page 236–243, 2008.

[63] G. Parseihian, C. Gondre, M. Aramaki, S. Ystad, and R. Kronland-Martinet. Comparison and evaluation of sonification strategies for guidance tasks. *IEEE Transactions on Multimedia*, 18(4) :674–686, April 2016.

[64] Karol J. Piczak. Environmental sound classification with convolutional neural networks. In *IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. ScienceDirect, 2015.

[65] Karol J. Piczak. ESC : Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, Oct 2015.

[66] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. Computer vision using local binary patterns. *Springer*, 2011.

[67] T. Ojala M. Pietikäinen and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *ieee Transactions on Pattern Analysis and Machine Intelligence*, 24(7) :971–987, jul 2002.

[68] L.C. Quero, J.I. Bartolome, S. Lee, E. Han, S. Kim, and J. Cho. An interactive multimodal guide to improve art accessibility for blind people. In *Proceedings of*

the 20th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '18, pages 346–348, New York, NY, USA, 2018. Association for Computing Machinery.

[69] A. Radecki, M. Bujacz, P. Skulimowski, and P. Strumillo. Interactive sonification of images on mobile devices for the visually impaired. In *2017 Signal Processing : Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 239–242, Sep. 2017.

[70] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann. Sound-event classification using robust texture features for robot hearing. *IEEE Transactions on Multimedia*, 19(3) :447–458, March 2017.

[71] J.B. Rodrigues, A.V.M. Ferreira, I.M.O. Maia, G.B. Junior, J.D.S. de Almeida, and A.C. de Paiva. Image processing of artworks for construction of 3d models accessible to the visually impaired. In *Advances in Manufacturing, Production Management and Process Control*, pages 243–253, Cham, 2019. Springer International Publishing.

[72] H.B. Sailor, D.M. Agrawal, and H.A. Patil. Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification). In *INTERSPEECH*, Aug 2017.

[73] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE SIGNAL PROCESSING LETTERS*, 24(3) :279–283, Mar 2017.

[74] Anabel S. Sánchez Sánchez and Maria Teresa Valderrama. Sonification of eeg signals based on musical structures. *2013 Pan American Health Care Exchanges (PAHCE)*, pages 1–1, 2013.

[75] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L-C Chen. Mobilenetv2 : Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[76] S. Scavaco, J. T. Henriques, M. Mengucci, N. Correia, and F. Medeiros. Color sonification for the visually impaired. In H. Krcmar M. M. Cruz-Cunha, J. Varajão and R. Martinho, editors, *Proceedings of International Conference on Health and Social Care Information Systems and Technologies (HCist)*, number 9 in Procedia Technology, pages 1048–1057. Elsevier, 2013.

[77] N. Schaffert and K. Mattes. Interactive sonification in rowing : Acoustic feedback for on-water training. *IEEE MultiMedia*, 22(1) :58–67, Jan 2015.

[78] C. Schörkhuber and A. Klapuri. Constant-q transform toolbox for music processing. In *7th Sound and Music Computing*, 2010.

[79] M. A. Sehili, D. Istrate, B. Dorizzi, and J. Boudy. Daily sound recognition using a combination of gmm and svm for home automation. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, Aug 2012.

[80] J. Sharma, O-C. Granmo, and M. Goodwin. Environment sound classification using multiple feature channels and deep convolutional neural networks. *arXiv :1908.11219*, Aug 2019.

[81] R. N. Shepard. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, 36(12) :2346–2353, 1964.

[82] M. Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer Technical Report*, 35, 1993.

[83] J. Sudol, O. Dialameh, C. Blanchard, and T. Dorcey. Looktel, a comprehensive platform for computer-aided visual assistance. In *2010 IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition - Workshops, pages 73–80, 2010.

[84] A. Tajadura-Jiménez, N. Bianchi-Berthouze, E. Furfaro, and F. Bevilacqua. Sonification of surface tapping changes behavior, surface perception, and emotion. *IEEE MultiMedia*, 22(1) :48–57, Jan 2015.

[85] O.K Toffa and M. Mignotte. A hierarchical visual feature-based approach for image sonification. *IEEE Transactions on Multimedia*, 23 :706–715, 2021.

[86] Y. Tokozume and T. Harada. Learning environmental sounds with end-to-end convolutional neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017.

[87] Y. Tokozume, Y. Ushiku, and T. Harada. Learning from between-class examples for deep sound recognition. In *ICLR*, Feb 2018.

[88] X. Valero and F. Alias. Gammatone cepstral coefficients : Biologically inspired features for non-speech audio classification. *IEEE Transactions on Multimedia*, 14(6) :1684–1689, Dec 2012.

[89] C.Y. Wan, A.G. Wood, D.C. Reutens, and S.J. Wilson. Early but not late-blindness leads to enhanced auditory perception. *Neuropsychologia*, 48(1) :344–348, Jan 2010.

[90] J-C. Wang, C.H. Lin, B-W. Chen, and M-K. Tsai. Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation. *IEEE Transactions on Automation Science and Engineering*, 11(2) :607–613, Apr 2014.

[91] J-C. Wang, J-F. Wang, K. Wai He, and C-S Hsu. Environmental sound classification using hybrid svm/knn classifier and mpeg-7 audio low-level descriptor. In *IEEE International Joint Conference on Neural Network Proceedings*, Jul 2006.

[92] Wikipedia. Piano key frequencies.

[93] R.M. Winters, N. Joshi, E. Cutrell, and M.R. Morris. Strategies for auditory display of social media. *Ergonomics in Design*, 27 :11–15, 2019.

[94] X. Wu and Z-N Li. A study of image-based music composition. In *2008 IEEE International Conference on Multimedia and Expo*, pages 1345–1348, 2008.

[95] Yang Xu, Zhixiong Li, Shuqing Wang, Weihua Li, Thompson Sarkodie-Gyan, and Shizhe Feng. A hybrid deep-learning model for fault diagnosis of rolling bearings. *Measurements*, 169 :108502, February 2021.

[96] W. S. Yeo and J. Berger. Application of raster scanning method to image sonification, sound visualization, sound analysis and synthesis. In *Proceedings of the Int. Conf. on Digital Audio Effects (DAFx-06)*, pages 309–314, Montreal, Quebec, Canada, September 2006.

[97] T. Yoshida, K. M. Kitani, H. Koike, S. Belongie, and K. Schlei. Edgesonic : Image feature sonification for the visually impaired. In *Proceedings of the 2Nd Augmented Human International Conference*, AH '11, pages 11 :1–11 :4, New York, NY, USA, 2011. ACM.

[98] T. Zhang and C.-C. Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *ieee Transactions on Speech and Audio Processing*, 9 :441–457, may 2001.

[99] Y. Zhao, S. Wu, L. Reynolds, and S. Azenkot. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), December 2017.

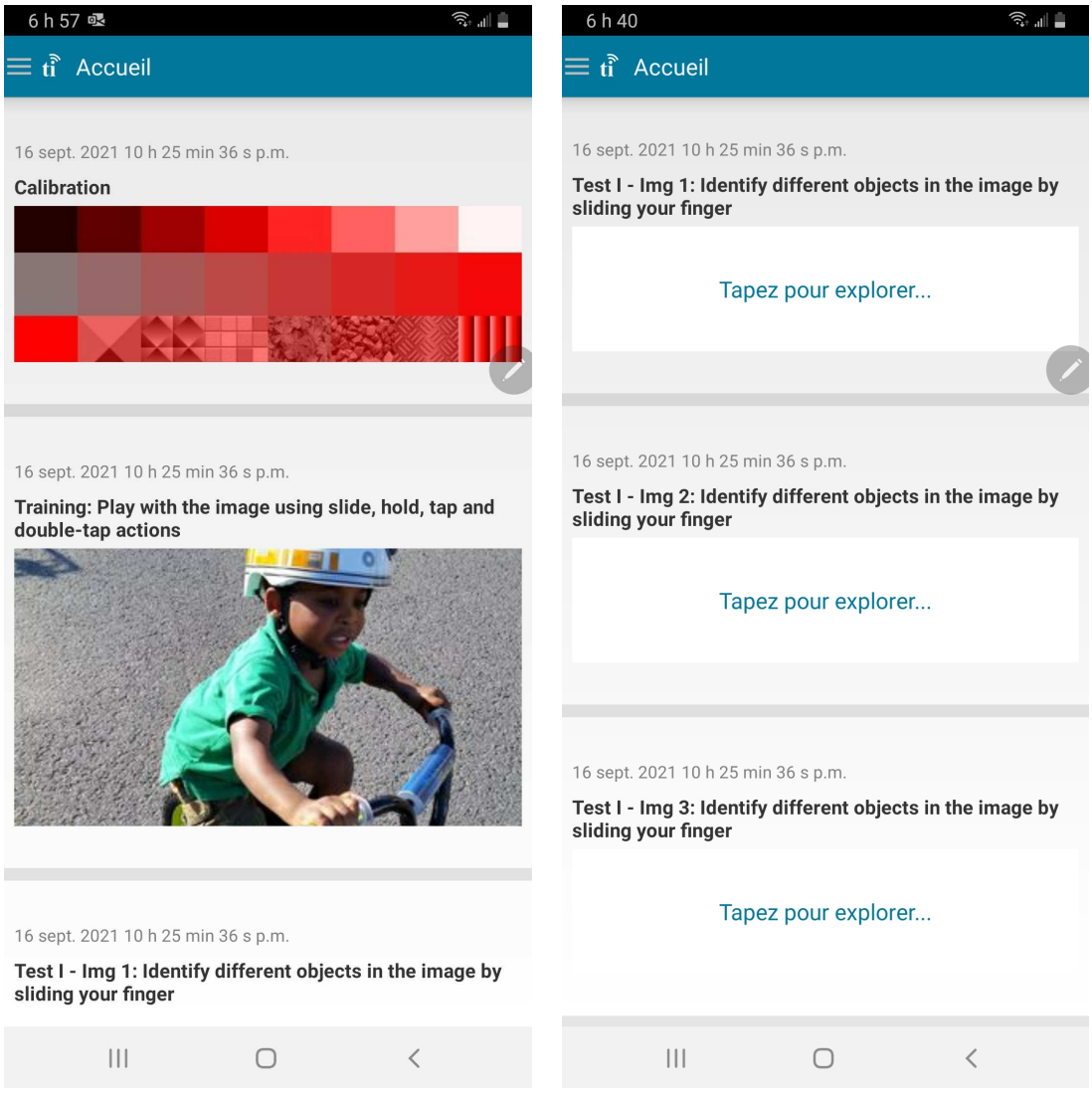
Annexe I

TalkingImage 2

L'application TalkingImage 2 est une application de sonification d'image qui est constituée de plusieurs parties.

- Une partie développée en python pour identifier les objets dans l'image
- Une partie en C++ pour segmenter l'image (http ://www.iro.umontreal.ca/ mignotte /ResearchMaterial/VOIBFM-SourceCode/ProgVOIBFM.tar.gz)
- Une partie en C++ pour générer les sons à partir du contenu abstrait (https ://github.com/ohinitoffa/ImgSonification)
- Une partie en java qui roule sur un téléphone Android et qui représente la partie client. Voir le fichier apk.
- Une partie serveur qui est développée en PHP et qui fournit les images et les sons
- Une base de donnees de sons et d'images

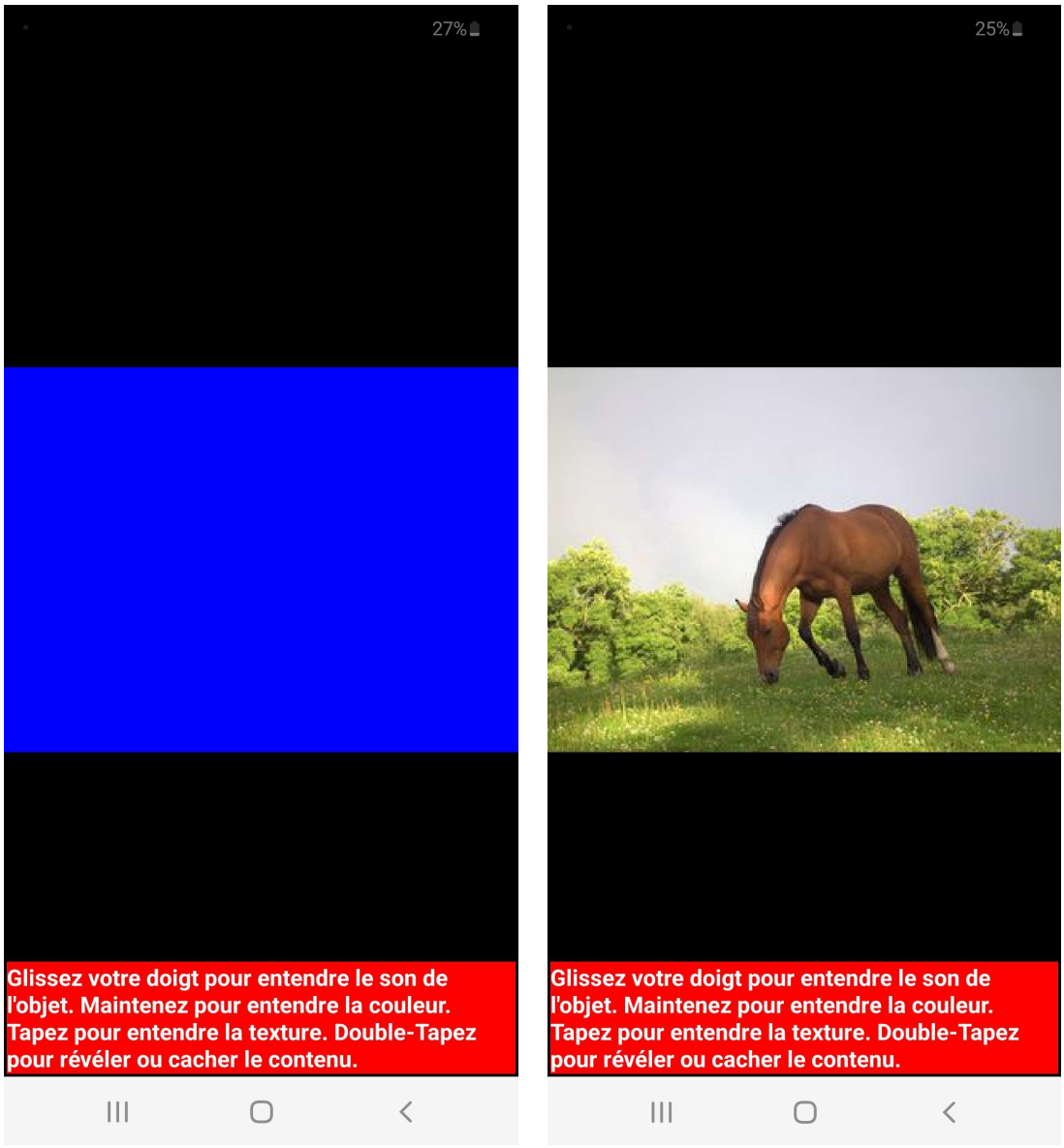
Le code source est disponible sur https ://github.com/ohinitoffa/ImgSonficiation2 L'application est téléchargeable au http ://www.talkingimage.ca/TalkingImage2.apk Nous en présentons quelques captures écrans dans les figures I.1 et I.2 .



(a) Training

(b) Test 1

FIGURE I.1 : Screenshots I of TalkingImage2 application



(a) Image hidden

(b) Image revealed

FIGURE I.2 : Screenshots II of TalkingImage2 application