Modeling and predicting affect in audio signals : perspectives from acoustics and chaotic dynamics

Pauline Mouawad

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Pauline Mouawad**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Modeling and Predicting Affect in Audio Signals: Perspectives from Acoustics and Chaotic Dynamics

Date de soutenance : 28 Juin 2017

Devant la commission d'examen composée de :

Myriam DESAINTE-CATHERINE	Pr., Bordeaux INP	Directrice de Thèse
Régine ANDRÉ-OBRECHT ..	Pr., Université Paul Sabatier	Rapporteur
Gérard ASSAYAG	DR HDR, Ircam	Rapporteur
Jenny BENOIS-PINEAU	Pr., Université de Bordeaux	Présidente
Shlomo DUBNOV	Pr., University of California San Diego	Examinateur
Marie BEURTON-AIMAR ...	MCF HDR, Université de Bordeaux .	Examinateur
Pascal DESBARATS	Pr., Université de Bordeaux	Invité

– 2017 –

Résumé

La présente thèse décrit un projet de recherche multidisciplinaire qui porte sur la reconnaissance de l'émotion dans les sons, couvrant les théories psychologiques, l'analyse du signal acoustique, l'apprentissage automatique et la dynamique chaotique.

Il est remarquable de constater que les interactions et les relations sociales des êtres humains s'appuient sur la communication des émotions. Souvent, la façon dont nous livrons nos pensées et nos opinions importe au moins autant que le message communiqué. Selon les perceptions affectives que nous avons pendant les échanges sociaux, nous pourrions ressentir une attraction ou un rejet envers les personnes ou les pensées exprimées. C'est parce que nous détectons souvent inconsciemment les vibrations provenant de ces échanges, et, inévitablement, cela affecte notre disposition.

En général, nous sentons les émotions dans des domaines tels que le visage, l'audition et le sensuel, et nous recueillons une foule d'informations sur les personnes, les situations ou les événements. Notre évaluation de ces informations peut être essentielle pour notre bien-être et notre survie: par exemple, dans les arts de la scène, des articulations authentiques d'émotions sont souvent exigées des acteurs professionnels, des chanteurs et des musiciens pour provoquer l'influence souhaitée sur le public; le son d'une querelle à proximité déclenche en nous une impulsion à vouloir fuir. Dans le premier exemple, nous sommes heureux, dans le second, nous avons peur. En fait, le son est peut-être le canal le plus important de la communication émotionnelle, dominant à cet égard d'autres canaux, qu'il s'agisse d'un son environnant, de la voix humaine ou de la musique. De toute évidence, même lorsque nous sommes impliqués dans une action et que notre attention est sollicitée, nous percevons constamment et inconsciemment les sons environnants, nous nous éloignons naturellement des sons désagréables et nous restons plus longtemps lorsque les sons sont agréables. Nous en déduisons également des associations: le chant des oiseaux nous rassure, le son des vagues de l'océan nous détend, instigue notre humeur pour les vacances. Moins fréquemment, il se peut qu'un tel son déclenche une rage: un cas qui s'est produit en 2011 lors de la projection du film Black Swan dans un cinéma multiplex Letton, un diplômé de l'école de police de 27 ans et ayant un doctorat en droit, a abattu un homme de 42 ans, parce qu'il mangeait son pop-corn trop fort[2].Il s'agit d'une maladie connue sous le nom de misophonie ou syndrome de sensibilité sonore sélective. Alternativement, la voix humaine est un instrument d'expression extraordinaire, qui transmet des informations telles que le genre du locuteur, l'âge, la santé ainsi que l'état émotionnel actuel. La musique est la forme la plus puissante du son en termes d'expression émotionnelle, car elle transmet des nuances émotionnelles distinctives qui affectent notre état émotionnel.

La prévalence de la composante affective peut être observée dans les interactions informatiques humaines (HCI). Il a été démontré que les processus impliqués dans nos échanges affectifs, tels que l'attention, l'apprentissage, la mémoire et la prise de décision, sont également déclenchés lorsque nous interagissons avec diverses applications multimédias. En particulier, une enquête sur un certain nombre de scénarios a montré que nous traitons les ordinateurs comme s'ils avaient des compétences sociales telles que l'intelligence et les sentiments. Cela peut expliquer pourquoi il est devenu de plus en plus important de développer des applications automatisées qui peuvent comprendre ou transmettre des émotions. Pourtant, il est important de noter que la composante affective ne devrait pas être considérée comme une fin en elle-même, mais elle devrait être incorporée dans le cadre de la conception des applications et des systèmes, ce qui les rendra significatifs et faciles à utiliser, de sorte que notre interaction avec ces systèmes devient une expérience positive.

[2] http://www.telegraph.co.uk/news/newstopics/howaboutthat/8337522/Man-shot-dead-for-eating-popcorntoo-loudly-during-Black-Swan.html

Dans son livre Affective Computing, Rosalind Picard le formule comme suit [153]:

> Les ordinateurs n'ont pas besoin de capacités affectives pour l'objectif fantaisiste de devenir humanoïdes, ils en ont besoin pour un objectif plus doux et plus pratique: fonctionner avec intelligence et sensibilité envers les humains.

Des exemples d'applications informatiques qui sont optimisées de manière affective peuvent être dans les conversations des centres d'appels, comme les systèmes de «miroirs affectifs» qui donnent aux opérateurs humains des commentaires sur l'émotion perçue dans leur voix, ce qui les aide à améliorer leurs compétences en interaction. Un système de dialogue peut détecter l'état émotionnel de l'utilisateur et fournir des commentaires sur les stratégies de conciliation, ou décider de transférer l'appel à un agent humain. Les systèmes automobiles qui génèrent des sons exprimant des alertes et des avertissements peuvent faire partie des systèmes numériques des véhicules à l'avenir.

Compte tenu de cela, et visant à comprendre les éléments des sons qui ont une signification affective, les études se sont principalement basées sur l'analyse de la relation entre les éléments acoustiques et les perceptions affectives. Une grande partie des connaissances que nous avons actuellement sur la reconnaissance de l'émotion (ER) dans la musique et la parole, nous le devons à l'approche d'analyse acoustique. Les exemples incluent des méthodes pour l'extraction de caractéristiques à partir de signaux et l'identification d'un vaste groupe de mesures acoustiques qui caractérisent les effets de la parole ou de la musique. En fait, les réalisations dans ces domaines ont rendu l'approche inévitable pour quiconque étudie ER en audio. Malgré ces faits, il n'y a pas de consensus sur un ensemble de caractéristiques acoustiques pour la caractérisation de l'émotion dans la voix ou la musique. De plus, des études ER sont souvent menées pour un type de son particulier, et peu de tâches ont abordé le problème de l'ER à partir d'une perspective holistique qui étudie l'émotion dans de multiples modalités sonores, comme la voix, la musique, ainsi que les sons environnementaux.

Récemment, de nouvelles approches d'autres domaines ont été utilisées pour compléter notre compréhension des mécanismes derrière l'expression de l'émotion dans les signaux audio. Un tel domaine provient du domaine de la dynamique chaotique. L'intérêt réside dans l'observation que les phénomènes non linéaires existent dans des sons comme: les instruments de musique, la voix, y compris les cris de nourrissons, les vocalisations de mammifères, les chants d'oiseaux, les voix pathologiques et le discours émotionnel. En outre, les attracteurs dynamiques inhérents au système dynamique du son et qui représentent ses trajectoires dans l'espace des états, portent une signification perceptive. Par conséquent, nous pouvons avoir une idée de l'expressivité affective des sons en inspectant les propriétés d'un tel comportement dynamique, ce qui peut être fait en mettant en œuvre des approches du domaine de la dynamique non linéaire. Ces informations clés ne peuvent être obtenues à partir de descripteurs acoustiques.

Le pouvoir émotionnel de la parole a été largement abordé dans la recherche. Une limitation principale des stimuli de la parole est qu'ils contiennent des informations sémantiques qui influencent le jugement émotionnel de l'auditeur. En outre, l'acoustique de la parole est modifiée quand il y a un contenu verbal, et comme nous comptons sur l'acoustique pour détecter les effets, la reconnaissance de l'émotion sera aussi affectée.

Moins d'études ont exploré le potentiel des communications vocales non verbales dans la transmission des émotions, telles que le chant non verbal ou les éclats affectifs non verbaux. De telles communications vocales peuvent être des expressions émotionnelles brèves, spontanées telles que des rires, des cris ou des soupirs, et leur valeur émotionnelle est contenue dans une syllabe.

L'avantage des vocalisations affectives est qu'elles sont des expressions spontanées d'émotions et sont donc sûres de transmettre une expression émotionnelle. En outre, les vocalisations affectives non verbales sont mieux reconnues que les stimuli de discours émotionnels exprimant la même émotion.

Le chant non verbal est une autre forme de communication vocale moins étudiée. Le fait qu'il n'y ait pas de paroles et pas de musique d'accompagnement, en fait une forme de sonorité intéressante pour les études ER, car l'information affective perçue ne sera liée qu'à l'expression chantée.

La communication affective dans la musique instrumentale non verbale est étudiée à partir d'une perspective d'analyse comparative avec des voix affectives non verbales. Le rôle des scènes auditives dans la transmission des émotions est également abordé dans cette thèse.

Portée de la thèse

La portée générale de cette thèse devrait être considérée comme un aperçu des perceptions émotionnelles des êtres humains dans de multiples types de sons, ce qui contribue à la psychologie, à l'acoustique et à la dynamique non linéaire. Au cœur de ce travail sont les sons non verbaux dans les domaines de la voix humaine, de la musique instrumentale et des scènes auditives.

Tout d'abord, nous voulons comprendre l'expressivité affective de la voix de chant non verbal quand aucune intention émotionnelle n'inspire la performance et en l'absence de musique et de paroles accompagnantes.

Deuxièmement, nous étudierons les vocalisations spontanées, qui sont intrinsèquement émotives par leur nature même. Ensuite, nous étudions l'efficacité des fonctionnalités acoustiques spécifiques à la voix pour capturer l'émotion dans le domaine de la musique. L'objectif de cette partie est d'apporter une nouvelle lumière de l'analyse acoustique et des algorithmes d'apprentissage supervisés, sur un débat en cours, concernant les origines communes de la musique et de la voix.

Troisièmement, au-delà des fonctionnalités couramment utilisées dans l'analyse acoustique des signaux audio, il est nécessaire de trouver de nouvelles fonctionnalités qui, idéalement, captureraient l'expressivité émotionnelle dans les sons tout en généralisant sur différents types de sons. Plus important encore, nous aimerions proposer un ensemble de fonctionnalités qui pourraient potentiellement faire l'objet d'un consensus parmi la communauté ER. Dans cette partie, nous étudions les phénomènes de dynamique non linéaire dans le son et proposons de nouvelles fonctionnalités dynamiques pour les études ER. Dans la partie expérimentale, nous donnons un aperçu de la dynamique non linéaire sur le débat soulevé au chapitre 3 concernant les origines communes de la voix et de la musique.

D’un point de vue de l’écoute humaine, une tâche d’annotation est menée pour construire un ground-truth de voix de chant non verbales, marquées par des descriptions catégoriques du modèle bidimensionnel d’émotions. Deux types de sons sont inclus dans l’étude: vocal et glottal.

D’un point de vue psychologique, la présente recherche porte sur un débat qui existe depuis longtemps parmi les scientifiques et les psychologues, concernant les origines communes de la musique et de la voix. La question est abordée à partir d’une analyse acoustique ainsi que d’une approche dynamique non linéaire.

D’un point de vue de la modélisation, ce travail propose une nouvelle approche dynamique non linéaire pour la reconnaissance de l’affect dans le son, basée sur la dynamique chaotique

et la symbolisation adaptative des séries temporelles. Tout au long de cette thèse, les contrastes clés dans l'expressivité de l'émotion sont illustrés parmi les différents types de sons, à travers l'analyse des propriétés acoustiques, les métriques de la dynamique non linéaire et les performances des prédictions.

Enfin, d'un point de vue progressif, nous suggérons que les travaux futurs étudient des caractéristiques motivées par les études cognitives. Nous suggérons également d'examiner dans quelle mesure nos caractéristiques reflètent les processus cognitifs. En outre, nous recommandons que nos fonctionnalités dynamiques soient testées dans des études à grande échelle de la reconnaissance d'émotions à travers la participation aux défis expérimentaux, dans le but de vérifier s'ils obtiennent un consensus.

Organisation de la thèse

Le travail dans cette thèse est organisé en six chapitres. Les chapitres 1 et 4 sont des examens de pointe sur les ER acoustiques basés sur la dynamique sonore et chaotique respectivement.

Le chapitre 1 est une revue de la littérature sur la recherche sur la reconnaissance de l'émotion dans les sons vocaux et musicaux. Il fournit des perspectives de la psychologie sur les émotions humaines, leur perception des sons et les différents modèles proposés pour les conceptualiser. Les différents défis rencontrés dans le domaine de la reconnaissance de l'émotion en audio sont discutés, l'approche dont ils ont été abordés et les questions ouvertes qui subsistent. Enfin, ce chapitre traite de divers algorithmes d'apprentissage supervisé dans le domaine et termine avec les travaux proposés pour les deux chapitres suivants.

Qu'est-ce qui peut être communiqué dans une voix de chant qui est dépouillée de musique et de paroles d'accompagnement? Il s'agit d'une question abordée au chapitre 2. D'abord, une tâche d'annotation humaine est menée où les auditeurs sont invités à écouter des voix de chant non verbal et à les annoter avec des descripteurs affectifs de la valence et de l'éveil. Les annotations correspondent à l'émotion que les auditeurs perçoivent dans le chant. Un aspect important du chant est qu'il est réalisé sans aucune intention émotionnelle. L'objectif est d'explorer l'expressivité affective de la voix chantante dans sa performance la plus fondamentale. Une analyse acoustique est faite de sons vocaux et glottaux. Ensuite, l'analyse statistique met en évidence la corrélation de chaque signal acoustique avec les étiquettes affectives.

L'émotion musicale a-t-elle évolué à partir de vocalisations affectives primitives? Est-il possible d'obtenir un modèle d'ER holistique qui prédit l'expression affective dans les deux canaux? Le chapitre 3 aborde un débat en cours entre les scientifiques et les psychologues sur les origines communes de l'expression musicale et vocale des émotions. Il présente une étude de la reconnaissance de l'émotion dans les formes primitives des expressions vocales et musicales. Grâce à l'analyse des attributs acoustiques spécifiques au domaine de la voix, elle étudie dans quelle mesure l'expression émotionnelle des sons vocaux et musicaux primitifs se révèle grâce à un code acoustique partagé. Nous appliquons une collection hybride de méthodes de sélection de fonctionnalités et testons divers algorithmes d'apprentissage supervisés. Afin de révéler le code acoustique de l'expression de l'émotion dans les deux domaines, une tâche d'apprentissage multi-domaine est réalisée.

Le chapitre 4 est une revue de la littérature sur les applications des méthodes de la dynamique chaotique à l'étude des phénomènes non linéaires dans les sons, dans le contexte de la reconnaissance des affects. Le chapitre traite du domaine de la théorie du chaos et des systèmes non linéaires déterministes, de l'analyse des séries temporelles non linéaires, de l'intégration temporelle (*time-delay embedding)*, des parcelles de récurrence *(recurrence plots) ainsi* que de l'analyse de la

quantification des récurrences. Ensuite, il élabore l'analyse symbolique des séries chronologiques et rapporte un modèle de symbolisation adaptatif récent appelé le Variable Markov Oracle, qui trouve les motifs significatifs dans la forme symbolisée d'une série temporelle. Enfin, le chapitre passe en revue des études récentes sur les applications de la dynamique non linéaire aux scènes vocales, musicales et auditives.

Grâce à l'application de la dynamique non linéaire à une méthode récemment développée de symbolisation des séries temporelles, le chapitre 5 aboutit à des mesures de complexité qui captent des modèles significatifs dans les séries temporelles ainsi que leur ordre temporel. Tout d'abord, le signal audio est transformé en vecteurs caractéristiques qui se rapprochent de l'analyse auditive humaine, puis les modèles sont quantifiés en utilisant des méthodes d'analyse statistique de la quantification des récurrences. Grâce à la mise en œuvre de méthodes d'apprentissage supervisé, nous explorons dans quelle mesure les phénomènes non linéaires portent une information affective perceptible et évaluent la performance des mesures de quantification caractérisant l'affect en utilisant des tâches d'apprentissage supervisé.

Dans le chapitre 6, nous nous demandons si les invariants dynamiques estimés avec notre méthode sont compatibles avec ceux obtenus en utilisant d'autres méthodes. Nous calculons la dimension de corrélation, l'entropie de corrélation, l'entropie de Shannon et l'exposant de Lyapunov à partir des séries chronologiques symbolisées de la carte logistique obtenue avec notre modèle. Ensuite, nous les comparons avec les mêmes mesures obtenues avec différentes méthodes de la littérature. Dans une phase ultérieure, nous calculons ces mesures à partir des sons et évaluons leur performance dans la reconnaissance de l'émotion. Nous comparons davantage notre approche en utilisant des fonctionnalités de dynamique non linéaire avec une approche de base en utilisant des caractéristiques acoustiques standard. Enfin, nous examinons la performance d'un ensemble hybride d'acoustique ainsi que des fonctionnalités de complexité dans la reconnaissance des affects.

Keywords : non-verbal sounds, acoustic analysis, machine learning, neural networks, chaotic dynamics, embedding, Variable Markov Oracle, symbolization, dynamical invariants.

Mots-clés : sons non-verbaux, analyse acoustique, apprentissage automatique, réseau de neurones,

Laboratoire d'accueil Laboratoire Bordelais de Recherche en Informatique (LaBRI)-UMR 5800. Domaine universitaire, 351 Cours de la Libération, 33400 Talence

Contents

List of Figures

List of Tables

Dedication

To my Parents

In heartfelt recognition of your never-failing support and patient guidance through the years of my studies, for steadily strengthening me to follow my dreams, for your unequivocal generosity, and for always being there for me in difficult times, I dedicate this thesis to you, my dad Georges J. Mouawad and my mom Aurore Gebrie Mouawad, with my overflowing gratitude! My genuine thanks are for you my sister Reine for always standing by me, you are my best friend, and whom I lean on the most.

With all my affection, I earnestly thank you, I am forever indebted to you!

Acknowledgements

I would like to thank my supervisor Pr. Myriam Desainte-Catherine, for accepting me in her group and for giving me the freedom to explore and pursue my research interests.

I am ever grateful to Pr. Shlomo Dubnov for his valuable time, for his willingness to share his ideas and for giving me essential advices. His constructive criticism of my work, as well as his insightful and inspiring discussions, have helped me apprehend the pillars of research beyond the limitations of my field.

I would like to express my sincere thanks for the following people for their helpful collaboration: Pr. Pascal Desbarats, Dr. Marie Beurton-Aimar, Pr. Anne Gégout-Petit and Pr. Catherine Semal.

I further wish to earnestly thank the members of my thesis committee: the reporters, Pr. Régine André-Obrecht and Dr. Gérard Assayag, for generously offering their time to report my dissertation, and for their rigorous, insightful and encouraging reports. I also thank Pr. Jenny Benoit-Pineau, Dr. Marie Beurton-Aimar, Pr. Shlomo Dubnov, and Pr. Pascal Desbarats, for accepting to review my thesis and be on the jury. I also thank the jury members for letting my defense be a gratifying moment, for their brilliant and inspiring comments, as well as for the challenging questions that instigated new avenues for my future research.

Finally to my friends in the Informatics department, thanks for the laughs and tears you shared with me, I will always remember the lunches and coffee breaks we spent together, laughing, debating and analysing Freud and Jung!

Prologue

A play by William Shakespeare called *The Tempest*, is set on a remote island, where Prospero the sorcerer, takes under his wing the enigmatic Caliban, one of the most complex and most debated figures in all of Shakespeare. The son of a witch and a devil, he is *littered* on the island, described as a monster, and strenuously enslaved to Prospero after attempting to rape his daughter. In one scene, while Caliban is narrating his plight, he is interrupted by the music of the butler. The music warms Caliban's spirits, and his ensuing response is one of Shakespeare's most acclaimed and most memorable speeches:

> Be not afeard. The isle is full of noises,
> Sounds, and sweet airs that give delight and hurt not.
> Sometimes a thousand twangling instruments
> Will hum about mine ears, and sometime voices
> That, if I then had waked after long sleep,
> Will make me sleep again. And then, in dreaming,
> The clouds methought would open and show riches
> Ready to drop upon me, that when I waked
> I cried to dream again.[1]

The sound of music transforms the complaints of the monster into a tale that *cures deafness*, where a palette of sounds engages our senses and lures us into Caliban's momentous entry into civilization. Arguably, we wonder about this sudden swell of poetic descriptions of the sounds of the island. We can imagine that their pervasiveness has quietly, yet profoundly influenced Caliban's spirits and fashioned his episodic memories. Although no particulars are given about the nature of the sounds or the sweet airs, their duration, rhythm or complexity, what are the instruments or the voices heard, we know that they elicited delight, and compelled tears.

Resolutely inspired by the powerful influence of sounds on our emotions, this dissertation is an exploration of emotion recognition in sounds. It dissects a variety of signals, aiming to recognize the elements and the mechanisms behind the resounding sceneries that model our affective existence.

[1] The Tempest, Act 3, Scene 2

Abstract

The present thesis describes a multidisciplinary research project on emotion recognition in sounds, covering psychological theories, acoustic-based signal analysis, machine learning and chaotic dynamics.

In our social interactions and relationships, we rely greatly on the communication of information and on our perception of the messages transmitted. In fact communication happens when signals transmit information between a source and a destination. The signal can be verbal, and the information is then carried by sound patterns, such as words. In nonverbal vocal communication however, information can be perceptual patterns that convey affective cues, that we sense and appraise in the form of intentions, attitudes, moods and emotions.

The prevalence of the affective component can be seen in human computer interactions (HCI) where the development of automated applications that understand and express emotions has become crucial. Such systems need to be meaningful and friendly to the end user, so that our interaction with them becomes a positive experience. Although the automatic recognition of emotions in sounds has received increased attention in recent years, it is still a young field of research. Not only does it contribute to Affective Computing in general, but it also provides insight into the significance of sounds in our daily life.

In this thesis the problem of affect recognition is addressed from a dual perspective: we start by taking a standard approach of acoustic-based signal analysis, where we survey and experiment with existing features to determine their role in emotion communication. Then, we turn to chaotic dynamics and time series symbolization, to understand the role of the inherent dynamics of sounds in affective expressiveness. We conduct our studies in the context of nonverbal sounds, namely voice, music and environmental sounds.

From a human listening point of view, an annotation task is conducted to build a ground truth of nonverbal singing voices, labelled with categorical descriptions of the two-dimensional model of affect. Two types of sounds are included in the study: vocal and glottal.

From a psychological perspective, the present research addresses a debate that is of long standing among scientists and psychologists, concerning the common origins of music and voice. The question is addressed from an acoustic-based analysis as well as a nonlinear dynamics approach.

From a modeling viewpoint, this work proposes a novel nonlinear dynamics approach for the recognition of affect in sound, based on chaotic dynamics and adaptive time series symbolization.

Throughout this thesis, key contrasts in the expressiveness of affect are illustrated among the different types of sounds, through the analysis of acoustic properties, nonlinear dynamics metrics and predictions performances.

Finally from a progressive perspective, we suggest that future works investigate features that are motivated by cognitive studies. We also suggest to examine to what extent our features reflect cognitive processes. Additionally we recommend that our dynamic features be tested in large scale ER studies through the participation in ER challenges, with an aim to verify if they gain consensus.

Résumé

La présente thèse décrit un projet de recherche multidisciplinaire qui porte sur la reconnaissance de l'émotion dans les sons, couvrant les théories psychologiques, l'analyse du signal acoustique, l'apprentissage automatique et la dynamique chaotique.

Dans nos interactions et nos relations sociales, nous dépendons considérablement de la communication de l'information et de notre perception des messages transmis. En fait, la communication se produit lorsque les signaux transmettent des informations entre une source et une destination. Le signal peut être verbal, et l'information est ensuite portée par des motifs sonores, tels que des mots. Dans la communication vocale non verbale, cependant, l'information peut être des modèles perceptifs qui véhiculent des indices affectifs, que nous percevons et évaluons sous la forme d'intentions, d'attitudes, d'humeurs et d'émotions.

La prévalence de la composante affective peut être observée dans les interactions informatiques humaines (HCI) où le développement d'applications automatisées qui comprennent et expriment les émotions est devenu crucial. De tels systèmes doivent être significatifs et faciles à utiliser pour l'utilisateur final, de sorte que notre interaction avec eux devient une expérience positive. Bien que la reconnaissance automatique des émotions dans les sons ait reçu une attention accrue au cours des dernières années, il s'agit encore d'un jeune domaine de recherche. Non seulement cela contribue à l'informatique affective en général, mais il fournit également une compréhension approfondie de la signification des sons dans notre vie quotidienne.

Dans cette thèse, le problème de la reconnaissance des affects est abordé à partir d'une double perspective: nous commençons par adopter une approche standard de l'analyse acoustique du signal, où nous examinons et expérimentons les fonctionnalités existantes pour déterminer leur rôle dans la communication émotionnelle. Ensuite, nous nous tournons vers la dynamique chaotique et la symbolisation des séries temporelles, pour comprendre le rôle de la dynamique inhérente des sons dans l'expressivité affective. Nous menons nos études dans le contexte des sons non verbaux, à savoir les sons vocaux, musicaux et environnementaux.

D'un point de vue de l'écoute humaine, une tâche d'annotation est menée pour construire un ground-truth de voix de chant non verbales, marquées par des descriptions catégoriques du modèle bidimensionnel d'émotions. Deux types de sons sont inclus dans l'étude: vocal et glottal.

D'un point de vue psychologique, la présente recherche porte sur un débat qui existe depuis longtemps parmi les scientifiques et les psychologues, concernant les origines communes de la musique et de la voix. La question est abordée à partir d'une analyse acoustique ainsi que d'une approche dynamique non linéaire.

D'un point de vue de la modélisation, ce travail propose une nouvelle approche dynamique non linéaire pour la reconnaissance de l'affect dans le son, basée sur la dynamique chaotique et la symbolisation adaptative des séries temporelles. Tout au long de cette thèse, les contrastes clés dans l'expressivité de l'émotion sont illustrés parmi les différents types de sons, à travers l'analyse des propriétés acoustiques, les métriques de la dynamique non linéaire et les performances des prédictions.

Enfin, d'un point de vue progressif, nous suggérons que les travaux futurs étudient des car-

actéristiques motivées par les études cognitives. Nous suggérons également d'examiner dans quelle mesure nos caractéristiques reflètent les processus cognitifs. En outre, nous recommandons que nos fonctionnalités dynamiques soient testées dans des études à grande échelle de la reconnaissance d'émotions à travers la participation aux défis expérimentaux, dans le but de vérifier s'ils obtiennent un consensus.

Introduction

> Most major topics in psychology and every major problem faced by humanity involve emotion. Perhaps the same could be said of cognition. Yet, in the psychology of human beings, with passions as well as reasons, with feelings as well as thoughts, it is the emotional side that remains the more mysterious. Psychology and humanity can progress without considering emotion—about as fast as someone running on one leg.
>
> James Russell

It is conspicuous to observe that the humans' social interactions and relationships build upon the communication of emotions. Often, the manner with which we deliver our thoughts and opinions matters at least as much as the message being communicated. Depending on the affective perceptions we have during social interchanges, we might feel an attraction or a rejection towards persons or thoughts expressed. This is because we often unconsciously sense the vibes coming from such interchanges, and inevitably, this swings our disposition.

Generally we sense emotions in such channels as facial, auditory and gestural, and retrieve a wealth of information about people, situations or events. Our appraisal of such information can be essential for our well-being and survival: for example, in performing arts, authentic articulations of emotions are often required from professional actors, singers and musicians [94,182] to ensure the desired sway on audiences; the sound of a nearby escalating fight triggers in us an impulse to want to flee. In the first example, we are pleased, in the second, we are afraid. In fact, the sound is perhaps the most important channel of emotional communication, dominating in this respect other channels, be it a surrounding sound, the human voice, or music. Clearly, even when we are involved in some action and our attention is solicited, we constantly and unconsciously perceive the surrounding sounds, we naturally move away from annoying ones, and linger when sounds are pleasant. We also infer associations: the sound of singing birds makes us feel reassured, the sound of ocean waves makes us relaxed, sets our

mood for holidays. Less frequently, it may happen that a given sound triggers rage: a case that happened in 2011 during the screening of Black Swan movie in a Latvian multiplex cinema, a 27-year old police academy graduate and a PhD in Law, shot dead a 42-year old man, because he was eating his popcorn too loudly[2]. This is a condition known as misophonia, or selective sound sensitivity syndrome. Alternatively, the human voice is an extraordinary instrument of expression, that conveys information such as the gender of the speaker, age, health as well the current emotional state. Music, is the most powerful form of sound in terms of emotional expression, as it conveys distinctive shades of emotional nuances that affect our emotional state.

The prevalence of the affective component can be seen in human computer interactions (HCI). It was shown that the processes involved in our affective interchanges, such as attention, learning, memory and decision making, are also triggered when we interact with various media applications. Particularly, an investigation of a number of scenarios showed that we treat computers as if they had social skills such as intelligence and feelings [11, 163]. This may explain why it has become increasingly important to develop automated applications that can understand or convey emotions. Yet it is important to note that the affective component should not be seen as an end by itself, but it should be incorporated as part of the applications' and systems' design, which will make them look meaningful and friendly, so our interaction with such systems becomes a positive experience. In her book *Affective Computing*, Rosalind Picard formulates it as follows [153]:

> Computers do not need affective abilities for the fanciful goal of becoming humanoids, they need them for a meeker and more practical goal: to function with intelligence and sensitivity towards humans.

Examples of computer-based applications that are affectively optimized can be in call center conversations, such as 'affective mirrors' systems that provide human operators with feedback on the emotion perceived in their voice, which helps them improve their interaction skills. A dialogue system may detect the user's emotional state and provide feedback about conciliation strategies, or decide to transfer the call to a human agent. Automotive systems that generate sounds expressing alerts and warnings may become part of vehicles' digital systems in the future [207, 221].

Bearing this in mind, and aiming to understand the elements of sounds that carry affective meaning, studies have relied for the most part on analysing the relation of acoustical elements to affective perceptions. A great deal of the knowledge we currently have about emotion recognition (ER) in music and speech, we owe it to the acoustic-based analysis approach. Examples include methods for the extraction of features from signals and the identification of a vast group of acoustic measures that characterize affect in speech or music. In fact achievements in these domains have made the approach inevitable for anyone studying ER in audio. In spite of those facts, there is to date no consensus on a set of acoustic features for the characterization of emotion in voice or music. Additionally, ER studies are often conducted for a particular sound type, and few tasks have addressed the ER problem from a holistic perspective that studies emotion in multiple sound modalities, such as voice, music, as well as environmental

[2] http://www.telegraph.co.uk/news/newstopics/howaboutthat/8337522/Man-shot-dead-for-eating-popcorn-too-loudly-during-Black-Swan.html

sounds.

Recently new approaches from other domains have been employed to complement our comprehension of the mechanisms behind the expression of emotion in audio signals. One such domain comes from the field of chaotic dynamics. The interest arose from the observation that nonlinear phenomena exist in sounds such as: musical instruments [67,131], voice [77], including infant cries [138], mammals vocalizations [232], bird songs [59], pathological voices [4, 74, 78] and emotional speech [75, 76, 167]. Furthermore, dynamical attractors that are inherent to the sound's dynamical system and that represent its trajectories in state space, carry perceptual meaning. Therefore, we may gain insight on the affective expressiveness of sounds by inspecting the properties of such dynamical behaviour, and this can be done by implementing approaches from the field of nonlinear dynamics. Such key information cannot be obtained from acoustic descriptors.

The emotional power of speech has been extensively addressed in research. A main limitation in speech stimuli is that it contains semantic information that influences the listener's emotional judgement. Furthermore, the acoustics of speech are modified when there is verbal content, and since we rely on the acoustics to detect affect, the emotion recognition will be impacted as well.

Fewer studies explored the potential of nonverbal vocal communications in transmitting emotions, such as nonverbal singing or nonverbal affective bursts. Such vocal communications can be brief, spontaneous emotional expressions such as, laughters, screams or sighs, and their emotional value is contained in one syllable. The advantage of affective vocalizations is that they are spontaneous expressions of emotions and therefore are sure to convey an emotional expression. Additionally, nonverbal affective vocalizations are better recognized than emotional speech stimuli expressing the same emotion [174].
Nonverbal singing is another form of vocal communication that is less researched. The fact that there are no lyrics and no accompanying music, makes it an interesting form of sound for ER studies, since the affective information perceived will be related to the sung utterance only. Affective communication in nonverbal instrumental music is investigated from a perspective of comparative analysis with nonverbal affective voices.
The role of auditory scenes in conveying emotions is also addressed in this thesis.

Thesis Scope

The overall scope of this dissertation should be seen as providing insight into humans' emotional perceptions in multiple types of sounds, making contributions spanning psychology, acoustics and nonlinear dynamics. At the core of this work are the nonverbal sounds in the domains of human voice, instrumental music and auditory scenes.

First we wish to understand the affective expressiveness of the nonverbal singing voice when no emotional intent inspires the performance, and in the absence of accompanying music and lyrics.

Second, we will investigate spontaneous vocalizations, that are intrinsically emotional by their very nature. Then we study the effectiveness of voice-specific acoustic features in capturing

emotion in the music domain. The objective in this part is to bring new light from acoustic analysis and supervised learning algorithms, on an ongoing debate pertaining to the common origins of music and voice.

Third, beyond the features commonly used in acoustic-based analysis of audio signals, it is necessary to find new features that ideally, would capture the emotional expressiveness in sounds while generalizing to different types of sounds. More importantly, we would like to propose a set of features that would potentially gain consensus among the ER community. In this part we investigate the nonlinear dynamics phenomena in sound, and propose new dynamical features for ER studies. In the experimental part, we provide insight from nonlinear dynamics on the debate raised in chapter 3 concerning the common origins of voice and music.

Thesis Overview and Organization

The work in this dissertation is organized in six chapters. Chapters 1 and 4 are state of the art reviews about acoustic-based ER in sound and chaotic dynamics respectively.

Chapter 1 is a literature review on emotion recognition research in vocal and musical sounds. It provides perspectives from psychology on human emotions, their perception in sounds, and the different models that have been proposed to conceptualize them. The various challenges faced in the field of emotion recognition in audio are discussed, how they have been addressed, and the open questions that remain. Finally this chapter discusses various supervised learning algorithms in the field and concludes with the proposed works for the following two chapters.

What can be communicated in a singing voice that is stripped of accompanying music and lyrics? This is a question addressed in chapter 2. First a human annotation task is conducted where listeners are asked to listen to nonverbal singing voices, and annotate them with affective descriptors of valence and arousal. The annotations correspond to the emotion that listeners perceive in the singing. An important aspect of the singing is that it is performed without any emotional intent. The goal is to explore the affective expressiveness of the singing voice in its most basic performance. An acoustic analysis is made of vocal and glottal sounds. Then statistical analysis highlights the correlation of each acoustic cue with the affective labels.

The contributions of this chapter include:

- A ground truth of emotionally annotated nonverbal singing voices.
- Confirms the role of acoustic cues in capturing affective information in the nonverbal singing voice performing without any emotional intent.
- Finds that the glottal sound of the singing voice conveys emotions to listeners, and that the underlying acoustic cues correlate with the emotions perceived.

Did musical emotion evolve from primitive affective vocalizations? Is it possible to obtain a holistic ER model that predicts the affective expression in both channels? Chapter 3 addresses an ongoing debate among scientists and psychologists about the common origins of musical and vocal expression of emotions. It presents a study of emotion recognition in primitive

forms of vocal and musical expressions. Through the analysis of acoustic attributes specific to the voice domain, it investigates to what extent emotional expression in primitive vocal and musical sounds is revealed through a shared acoustic code. We apply a hybrid collection of feature selection methods, and test various supervised learning algorithms. In order to reveal the acoustic code of emotion expression in both domains, a cross-domain learning task is made.

The contributions of this chapter include:

- Addresses a debate that is rarely studied in the ER community: is there a common origin of emotional expression in voice and music?
- Investigates the cross-domain capability of voice-specific features in capturing emotions in the music domain.
- Explores to what extent a holistic model is capable of predicting affective expression in voice and music.
- Makes recommendations of the learning models depending on the nature of the task conducted.
- Makes recommendations of the acoustic cues that capture affect per domain and across domains.

Chapter 4 is a literature review on the applications of chaotic dynamics methods to the study of nonlinear phenomena in sounds, in the context of affect recognition. The chapter covers the field of chaos theory and deterministic nonlinear systems, nonlinear time series analysis, time-delay embedding, recurrence plots as well as recurrence quantification analysis. Then it elaborates on symbolic time series analysis, and reports on a recent adaptive symbolization model called the Variable Markov Oracle, that finds the meaningful patterns in the symbolized for of a time series. Finally the chapter reviews recent studies on the applications of nonlinear dynamics to voice, music and auditory scenes.

Through the application of nonlinear dynamics to a recently developed method of time series symbolization, chapter 5 derives complexity metrics that capture meaningful patterns in the time series as well as their temporal order. First the audio signal is transformed into feature vectors that approximate human auditory analysis, then patterns are quantified using methods of dynamical statistical quantification analysis. Through the implementation of supervised learning methods, we explore to what extent nonlinear phenomena carry perceptible affective information, and evaluate the performance of the quantification measures characterizing affect using supervised learning tasks.

The contributions of this chapter include:

- We combine nonlinear dynamics approach with symbolic time series analysis to propose new features, namely the symbolic RQA_{S} measures from symbolic recurrence plots.
- We also compute the RQA_{S} estimates for the logistic map. We compare them with the same measures derived with embedding. The advantage of our method is that our RQA_{S} measures can be considered as dynamical invariants since they are obtained independently of the embedding parameters m and τ:

- The performance of the RQA_s measures in the recognition of affect is evaluated on three datasets: vocal, musical and auditory scenes.
- The generalization of the symbolic estimates is evaluated in a cross-domain classification task.

In chapter 6 we question whether dynamical invariants estimated with our method are consistent with those obtained using other methods. We compute the correlation dimension, correlation entropy, the Shannon entropy and the Lyapunov exponent from the symbolized time series of the logistic map obtained with our model. Then we compare them with the same measures obtained with different methods from literature. In a subsequent phase we calculate these measures from sounds and evaluate their performance in emotion recognition. We further compare our approach using nonlinear dynamics features with a baseline approach using standard acoustic features. Finally we examine the performance of a hybrid set of acoustics as well as complexity features in affect recognition.

The contributions of this work include:

- Proposes new complexity measures derived from a symbolization of time series.
- Proposes a hybrid ER model, that combines the invariants as well as the acoustics.

Dissemination

The work done in this dissertation appears in the following:

1. : Chapter 2:
   (a) Proceedings of the 10^{th} International Symposium on Computer Music Multidisciplinary Research, CMMR 2013.
   (b) Sound, Music and Motion, pp. 105—121, Springer LNCS 2014.
2. Chapter 3: the first part of the work appears in the following, and the second part is pending submission:
   (a) Proceedings of the 21^{st} International Symposium on Electronic Art, 8^{th} International Workshop on Machine Learning and Music (MML2015).
3. Chapter 5:
   (a) Pauline Mouawad and Shlomo Dubnov. Novel Method of Nonlinear Symbolic Dynamics for Semantic Analysis of Auditory Scenes. In proceedings of the First International Workshop on Semantic Computing for Entertainment, 11th International Conference on Semantic Computing, 2017.
   (b) Pauline Mouawad and Shlomo Dubnov. On Symbolic Dynamics and Recurrence Quantification Analysis for Affect Recognition in Voice and Music. In proceedings of the International Conference on Perspectives in Nonlinear Dynamics, 2016.

4. Chapter 6:

   (a) Pauline Mouawad and Shlomo Dubnov. Novel Feature Extraction Method of Dynamical Invariants for Emotion Recognition. Submitted to: Special Issue on Recent Advances in Engineering Systems, Advances in Science, Technology and Engineering Systems Journal, ASTESJ, 2017.

Chapter 1

Emotion Perception in Voice and Music State of the Art - Part I

> If you can change your perception, you can change your emotion and this can lead to new ideas.
>
> Edward de Bono

This chapter surveys literatures on emotion recognition (ER) in voice (VER) and music (MER). The first section provides theoretical perspectives on human emotions from psychology. The review includes the proposed models of emotions as well as notions of musical and vocal emotions. The second section introduces the general framework of content-based ER with its main components, that consist of human annotation experiments, acoustical analysis, and machine learning algorithms. This section discusses the challenges involved in ER studies and how the community addresses them. Then a review is made of the main literatures that identify acoustic measures as well as learning algorithms pertinent for ER studies. Next the evaluation metrics that are necessary to evaluate and compare the performances of different approaches are reviewed. The fourth section discusses the state of the art literatures on emotion recognition from voice and music. Finally in section five the objectives are narrowed down, and the work proposed is explained along the major contributions achieved.

Our objectives are first to identify the vocal acoustic cues that capture emotional communication in two realizations of the nonverbal voice: the spontaneous affective vocalizations and the singing voice. Second, we employ methods from computational ER to contribute to a research question that is a matter of speculation among scientists and philosophers: has music evolved from primitive vocal expressions? Or is music the product of human ingenuity and cultural creativity [213]? As will be elaborated throughout the chapter, prominent researchers from psychology have addressed this issue, however rare are the studies that have approached this question from an acoustical analysis angle. Choosing to investigate the nonverbal channels of voice and music, allows us to apprehend affect expression without having to address the challenges of semantics derived from words, sentences or lyrics.

1.1 Perspectives from Psychology

The nature and definition of emotion is an open research issue of long standing among psychologists. A 'working definition of emotion' on which there seems to be a consensus was proposed by Fontaine et al. 2007 [62] and referred to by Anagnostopoulos et al. 2015 [5] states that: "Emotions are episodes of coordinated changes in several components (including at least neurophysiologic activation, motor expression and subjective feeling but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism" [5]. From this definition, the general agreement is that emotions are episodes that have several components; these episodes can be grouped into categories; and emotional experiences, whether through verbal or nonverbal channels, can be represented in a lower-order dimensional space, such as valence and arousal [61].

Studies on human emotions in varying disciplines such as cognitive sciences, neuroscience or affective computing, frequently use the terms emotions, feelings and affect in an interchangeable fashion, although these emotions have distinct connotations in psychology. According to [136, 143, 200], affect is a non-conscious predecessor to feelings and emotions. Feelings are conscious, personal and biographical sensations. Emotions are expressions of affect, and the projections or displays of feelings, and can be genuine or feign.

In line with the usage adopted by Picard [153] in her book *Affective Computing*, the terms emotions and affect are mutually used in this thesis. Particularly, they both refer to the emotion we perceive in audio stimuli, through our auditory system. Emotion perceived in vocal stimuli is termed *vocal emotion*, and that which is perceived in musical stimuli is termed *musical emotion*.

Musical Emotions A classic debate regarding musical emotions concerns whether emotions are perceived or induced. The two positions are commonly referred to as the cognitivist and the emotivist perspectives, respectively. Arguably, often we might perceive a particular emotion in a music without it actually awakening that emotion in us [7]. Philosopher Peter Kivy considers that "we might recognize these emotions, but not necessarily feel them". In *The Music Instinct*, Philip Ball puts it this way: "Mozart's Jupiter Symphony sounds happy to me even if I am feeling rotten, and even if it doesn't make me feel any better" [7]. This is the cognitivist perspective, and it states that the listener by means of pure intellectual or cognitive appraisal perceives a musical emotion without actually feeling it [101, 140]. And the emotional appraisal is based on the evaluation of the audio features [217].

The emotivist position states that the listeners through a process of emotional contagion experiences physiological changes during music listening that are similar to the changes that happen with real emotions [95, 96, 217]. And the induced emotion is based on the emotional experience of the listener [217].

Vempala et al. 2013 provide evidence that a combination of both perspectives form the listeners' emotional judgements: listeners make the emotional judgement based on both, the audio features perceived in the musical stimuli as well as their personal experience of it [217].

In our work we do not attempt to provide supporting evidence to either of the perspectives. But we focus on the perceived emotions because they can be measured more objectively than induced emotions, which are subjective in nature and are very difficult to measure.

Vocal Emotions The human voice is a complex instrument of communication that has a great emotional power: it transmits various information such as gender, age, well-being, as well as the current emotional state of the speaker. The manifestations of vocal emotions can happen during speech, singing, or spontaneous affective outcries such as laughters, sobs or sighs. During our social interactions, we rely greatly on our perceptions of how things are expressed as much as what is being said. In other respects, there is a growing evidence that the way we express our emotions affects our physical health [40, 68, 94, 162].

The emotional power of speech has been extensively addressed in research [69, 94, 108, 150, 178, 182, 240]. However fewer studies explored the potential of nonverbal vocal communications in transmitting emotions, such as nonverbal singing or nonverbal affective bursts. A known limitation in speech stimuli is that speech contains semantic information that influences the listener's emotional judgement. The same applies for singing, where lyrics and accompanying music blur the listener's perceptions. In fact, the acoustics of speech and verbal singing are modified when there is verbal content, and since we rely on the acoustics to detect affect, the emotion recognition will be impacted as well. To address this problem, researchers have used the standard contents paradigm, where the same word or sentence are repeatedly uttered but with a different emotion each time. Other researchers have used nonverbal vocalizations of emotions [175, 177, 186] that contain no words, and hence it is sure that the listener's emotional judgement will be related to the utterance.

Non-verbal vocal communications can be brief, spontaneous nonverbal emotional expressions such as a laughters screams or signs, and their emotional value is contained in one syllable [174, 175, 177, 186]. The advantage of the affective vocalizations is that they are spontaneous expressions of emotions, and therefore are sure to convey an emotional expression to the listener [186, 201]. Furthermore, nonverbal affective vocalizations are better recognized than emotional speech stimuli expressing the same emotion [174].

Another form of nonverbal vocal communication is nonverbal singing. In this case, a singer performs a vowel, possibly on a musical scale. The first advantage of studying nonverbal singing voices lies in the fact that there are no lyrics and no music, so the affective information perceived will be related to the sung utterance only. Second, it has been shown that the sustained vowel 'ahh..' is sufficient for various voice assessment applications [212, 215]. Furthermore, spectral characteristics of vowels efficiently gauge emotional content in vocal sounds [114] and improve emotion classification performance [165].

1.1.1 Models of Emotions

There are three main conceptualizations of emotion modelling: categorical, dimensional and appraisal-based. In the context of ER research, the three models have been widely employed, with the appraisal model being less popular than the other two.

Categorical Account Mostly due to the work of Paul Ekman [48], the categorical approach defines a set of emotional categories that are *innate*, *basic*, and *universal* to mankind [50]. This model defines emotions in discrete terms or adjectives that are distinct from each other and that convey specific meaning. Furthermore, each basic emotion is associated with a distinct

physiological profile [49]. Although Ekman's work was primarily based on recognizing emotions in faces, the categorical approach has been widely used for ER in music and voice.

In her 1936 seminal work, Kate Hevner devised a set of eight clusters of affective terms to described musical emotions [79] perceived and reported by listeners. She arranged them in a circle that shows the relationships between the adjectives of neighbouring clusters (figure 1.1). The clusters were revisited by Farnsworth [56] and later by Schubert [187]. From those adjectives, the most commonly employed ones for ER in music and voice are: happy, sad, angry, disgust, surprise and fear [109, 112, 154].

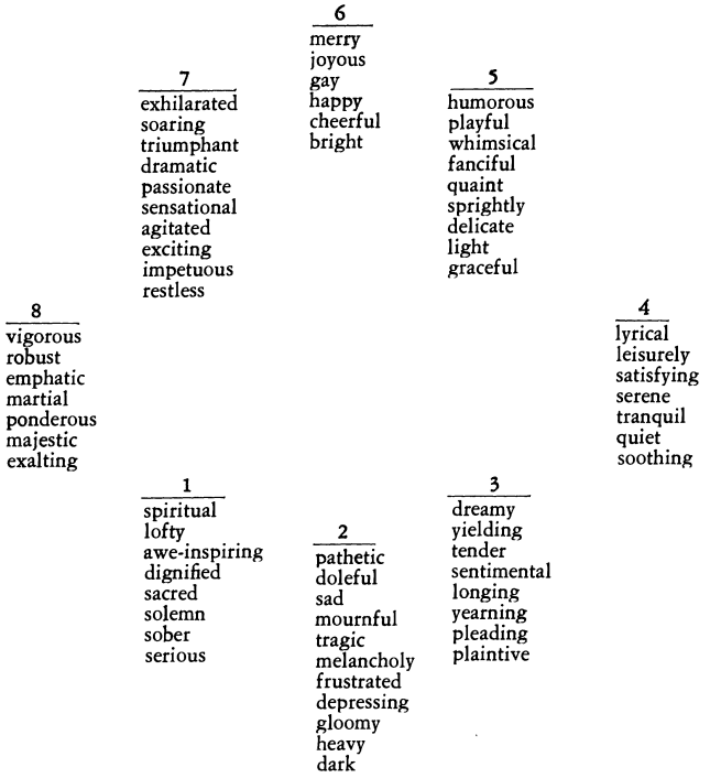

Figure 1.1: Hevner's eight clusters of musical emotions categories [79]

The advantage of the categorical approach lies in its objective description of the perceived emotions. However the set of discrete terms is too limited to capture the richness and nuances of musical and vocal emotions. And if a finer granularity of emotions is used, the ER analysis problem becomes difficult to perform because with a large set of terms, there will be a greater ambiguity in interpreting emotions since every term may have a different meaning for every person [234].

Numerous researchers provide support to the categorical model. Scherer et al. 2001 studied vocal expressions of five categorical emotions: joy, sadness, anger, fear and neutrality [180]. Their stimuli consisted of sequences of nonsense syllables [174], and achieved a recognition rate of 66%.

Laukka 2003 provided evidence for the suitability of the categorical model for ER in speech [107]. The stimuli consisted of concatenative synthesized speech that portrayed continua of emo-

tions: anger-fear, anger-sadness, fear-happiness, fear-sadness, happiness-anger and happiness-sadness [174]. Participants clearly perceived the boundaries in the emotions, which showed that the perception of emotions from vocal expressions is categorical.

The component process appraisal model of emotion The concept of a cognitive appraisal process of emotions dates back to Aristotle and was popularized in contemporary psychological research by psychologist Magda Arnold. The central notion is that emotions presuppose cognition, and emotions are object-oriented.

In [179] the component process appraisal model (CPM) describes affect as a process involving five functional emotion components: cognitive, peripheral efference, motivational, motor expression and subjective feeling. We limit the discussion of this model because it is outside the scope of our research, but further information can be found in [52, 179, 181].

Dimensional Account Another approach views musical and vocal emotions as a continua on some dimensions. Seminal works by James Russell [172] and Robert E. Thayer [211] have been the most influential in terms of affective dimensions and their applications to ER research.

Russell's model is called the circumplex model of core affect and consists of two dimensions, a horizontal axis of valence and an orthogonal axis of arousal. Emotional adjectives are distributed in a circular structure around the dimensions of valence and arousal (VA). Russell describes valence and arousal as being *core processes* of affect.

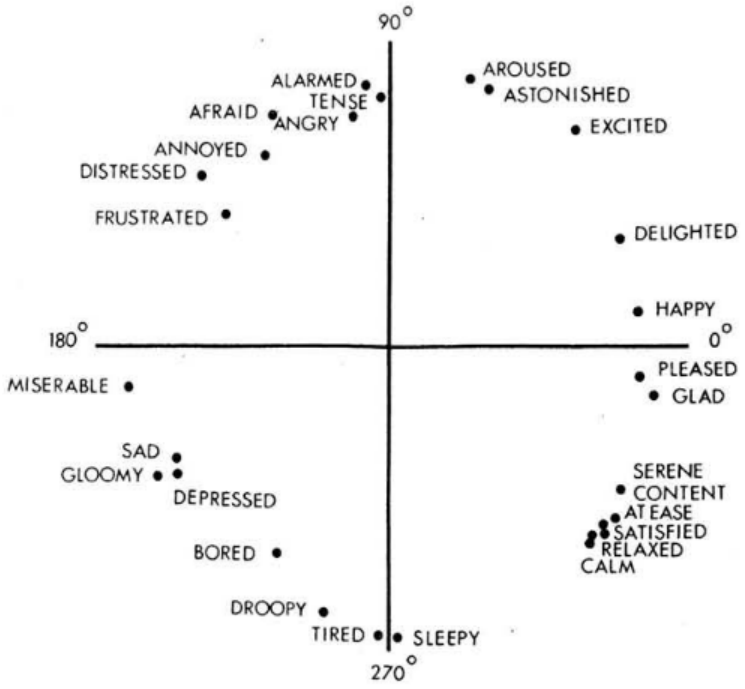

Figure 1.2: Russell's model of affect [172]

Many researchers provide evidence for the dimensional account of emotions [84]. Using the

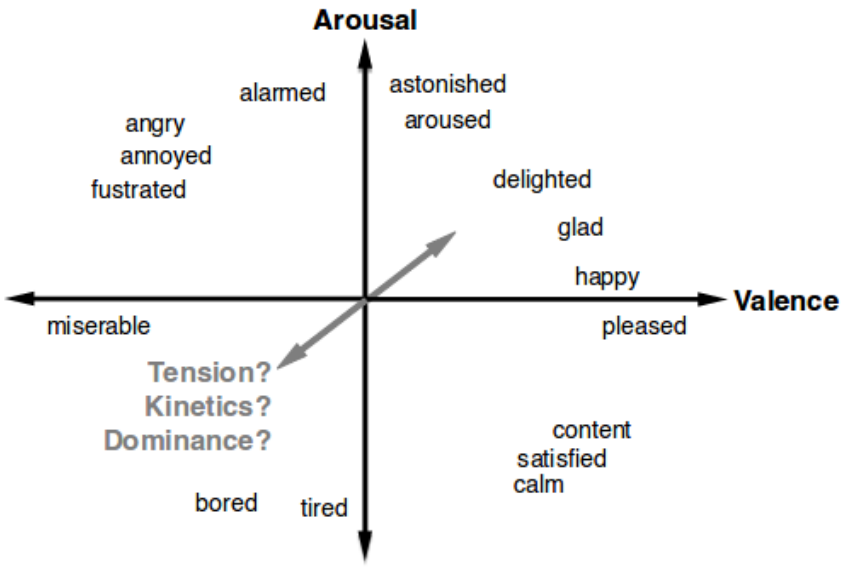

Figure 1.3: Valence-Arousal model [100]. Adjectives by Russell [172]. Third dimension: *tension* [46], *kinetics* [142], *dominance* [137]

dimensional model, ER studies are formulated as a regression problem [193] that predicts VA values on the 2D plane. Laurier et al. 2009 [110] as well as Levy et al. 2007 [118] studied the musical emotions derived from social tags, and both studies provided evidence to the pertinence of the two-dimensional model for emotion recognition studies.

A limitation of the two-dimensional representation of emotions is that it does not capture psychological distinctions between various emotions [111]. An example is the two emotions *anger* and *fear* which are placed close to each other in the upper left quadrant of the 2D model, although they have different physiological impacts. To address this problem some researchers proposed to add a third dimension *potency* having the two bipolar adjectives: dominant-submissive. However it is unclear to what extent it captures all of the relevant emotions, however it increases the complexity of the ER task.

Figure 1.3 shows the 2D model of affect (VA), along a third dimension proposed differently by different researchers: tension by [46], kinetics by [142] and dominance by [137].

Discussion Whether to model emotions in sounds using the categorical or dimensional model is a topic that's been debated for a long time among psychologists, and each model has its merits and limitations [234]. A comparison between the categorical and dimensional models is made in [47]. In this study, the authors systematically compare perceived emotions in music using five discrete emotions as well as three bipolar dimensional model of emotions (valence, energy-arousal and tension-arousal). They find that the three dimensions can be reduced to two without a significant loss of the goodness of fit. A main difference between the categorical and dimensional models is that the former has a poorer resolution [9], which indicates that it cannot distinguish between emotionally ambiguous examples.

Some researchers maintain that four dimensions are needed in order to portray emotions

from in a way that identifies similarities and differences between the channels studies [62]. The four dimensions are in decreasing order of importance: evaluation-pleasantness, potency-valence, activation-arousal and unpredictability [5]. However many studies have supported Russell's two-dimensional(2D) model of affect [44, 234]. For example [70, 172] show that two-dimensions are sufficient to represent human emotions communicated in voice and music.

Juslin and Sloboda 2001 claim that theoretically arousal is a major distinctive feature of emotion, and valence is directly related to behaviour [1].

The ER community appears to have favoured the 2D model of affect, because it provides a sufficient description of musical and vocal emotions, without increasing the complexity of the ER study, which the addition of dimensions would definitely cause [234]. The VA model can be described in categorical terms such as pleasant-unpleasant for valence, and awake-tired for arousal [183]. This provides a good tradeoff between using a limited set of discrete emotions, or an infinite set of continuous values on the VA model. Eerola and Vuoskoski 2011 made a comparison of the discrete and dimensional models and described the three-dimensional model of affect with the categories: high/low valence, high/low energy, high/low tension. They found that the two models provide highly compatible ratings of perceived emotion [47].

1.2 Emotion Recognition Framework

The ER framework is portrayed in figure 1.4. It consists of the following stages: given a signal, a human annotation task is conducted to label sounds with emotional labels/values and thus obtain a ground truth. The affective annotations can be numerical values of valence-arousal or discrete emotional adjectives. Then a bag of features are extracted from the signal, feature selection methods or dimension reduction algorithms are applied in order to reduce the initial feature set to its most relevant features. Then a supervised learning algorithm learns the relationship between the features and the affective labels: for numerical annotations the learning is a regression, and for categorical labels it is a classification. Finally the resulting learning model is used to predict the emotions of new unlabelled features [234]. The details of each stage are explained next.

Challenges Various challenges are involved when investigating the question of emotion recognition in human vocal communication, such as speech, affective bursts or singing.
First, real life recordings of emotional human vocal communication is a very difficult task to perform, so the vast majority of databases consist of enacted stimuli.
Second, it is very difficult to make a comparative analysis of the various methods employed in the literatures, because often very different stimuli are tested, using different feature vectors and different evaluation frameworks.
A third challenge particular to this thesis, is that most of ER research has been done on either speech, or music. Very little research have investigated the communication of emotion in the acoustics of nonverbal affective vocalizations. Except for [174], theoretical work on what acoustic measures are most suitable for this particular type of sounds is lacking, therefore we borrow concepts from the speech emotion recognition (SER) field to select a set of feature

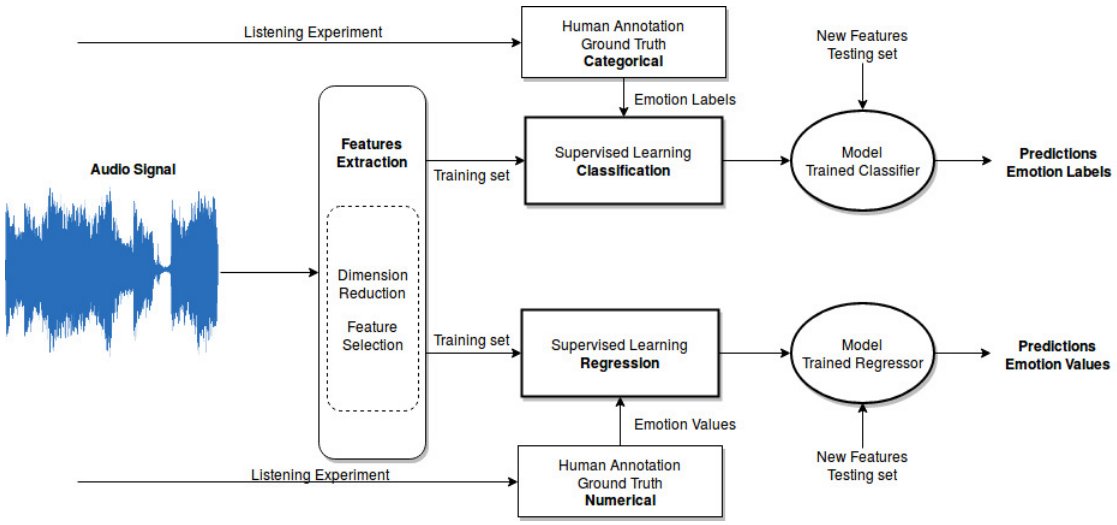

Figure 1.4: Framework of Audio Emotion Recognition: Top: the task is a classification if the ground truth is categorical emotions. Bottom: the task is a regression if ground truth is emotional values

vectors appropriate for ER in nonverbal vocalizations.
Fourth, given that we investigate the communication of emotion in nonverbal vocal and musical sounds, we are interested in the acoustic cues that are relevant to emotion studies in both modalities.

1.2.1 Databases and Ground Truth

A central challenge to the ER field is that researchers have to build their own ground truth, mainly because there is no consensus on the emotional model to use nor how many emotion categories to include in the studies [38,82] and also due to copyright reasons, the sounds cannot be made publicly available [234]. Generally the ground truth is obtained by conducting a human annotation task where subjects are asked to listen to the stimuli and assign an emotional label or value to the sound. A questionnaire collects various information about the participants such as the age, gender, or musical expertise, and explains what they are expected to annotate: the emotion perceived or induced.

In VER studies, most databases consist of enacted emotional speech or voices. Recording natural speech or affective vocalizations as they occur in real life is a very difficult task, in addition to the fact that the recordings will contain a great deal of noise. Arguably, the enacted emotional voices are not as natural as spontaneous ones. Yet most of vocal databases that are widely used consist of enacted sounds, and cover a wide range of emotions.

Singing Voice Database The Singing Voice Database[1] (SVDB) consists of nonverbal singing of a sung vowel 'ah' interpreted by professional singers (1 male and 1 female). The scale consists

[1] http://crel.calit2.net/projects/databases/svdb

of musical notes from A2 to E4 and A3 to A5 for male and female voice respectively. Vocal as well as glottal signals are recorded as mono files in WAVE PCM format, at 16 bits and 44 kHz. The sound files were trimmed using MIRToolbox [106] to remove the silence at the beginning, and were segmented using R statistical software [210] so that only the first note of the scale is retained. The final dataset consists of 44 sound samples, 22 vocal and 22 glottal of 1 second duration each. The database had no emotional annotations, so a human annotation task was performed to create the ground truth of emotionally annotated sounds. Details of this task are reported in chapter 1. The SVDB is considered pertinent to our studies, as it is a nonverbal expression of the singing voice.

Montreal Affective Voices The Montreal Affective Voices (MAV) [12] consist of 90 nonverbal affect bursts, enacted by 10 different actors in the following 8 categorical emotions: anger, disgust, fear, pain, sadness, surprise, happiness and pleasure, in addition to the neutral expression. The durations of the vocalizations vary between 0.385 to 2.229 seconds sampled at 44100 Hz. In order to perform a comparative analysis with different databases and to ensure consistency in the affective annotations, the affective categories of MAV sounds were mapped on the VA model described by: pleasant-unpleasant (V) and awake-tired (A). The MAV represent a primitive form of nonverbal, vocal expressions of affect.

Musical Emotional Bursts The Musical Emotional Bursts (MEB) [149] consist of 80 short instrumental musical clips, played using the clarinet and the violin, in the following 3 categorical emotions: happiness, sadness and fear, plus the neutral expression. The MEB was designed as a musical analogue of the MAV database, and as such, they represent a primitive form of nonverbal musical affective expression. The clips include imitations of the MAV stimuli as well as some improvisations on a given emotion. The mean duration of the MEB clips is of 1.6 seconds. In order to compare our study with the MAV database, the discrete emotions were similarly mapped on the VA model described by: pleasant-unpleasant (V) and awake-tired (A).

International Affective Digitized Sounds The International Affective Digitized Sounds (IADS-2) [20] consist of 167 normative and standardized affective sound stimuli at 44100 Hz sampling rate, designed for experimental studies on emotion and attention. The sounds cover a broad range of semantic categories describing auditory scenes such as the sound of a roller coaster, horse race, party, casino, human voices as well as music. The IADS-2 emotional ground truth is made by the authors on the three-dimensional model of valence, arousal and dominance. To perform the comparative analysis with the other databases, we used only the VA ground truth.

Film Music Excerpts The Film Music Excerpts (FME) [47] consist of 360 musical excerpts at 44100 Hz sampling rate, annotated on both the categorical as well as the dimensional model of emotions. The categorical emotions are: happy, sad, tender, fear, anger and surprise. The affective dimensions are described with: high valence, low valence, high energy, low energy, high tension, low tension. The FME is designed to provide musical stimuli that allow a systematic

comparison of the perceived emotions in music using two models of emotions: the categorical and the dimensional.

1.2.2 Acoustic Features for Emotion Recognition

When analysing complex signals such as vocal or musical sounds, a primary concern is the number of variables to extract as well as the appropriate type that is suitable for ER. If the number of the features is large, then the ER task will involve increased computation time and memory load. This can impact the performance of ER learning tasks where large databases are essential. Furthermore, a large feature set may impact the classifier's performance since it may overfit the data and generalize poorly to new data [159].

Feature extraction (figure 1.4) extracts a number of features from the signal. Then feature transformation techniques such as feature selection as well as dimensionality reduction reduce the number of the feature set and keep the most relevant ones. Feature selection methods consist of supervised algorithms that prune noisy or redundant features and keep only the most discriminative ones for for the emotions studied, and they can greatly impact/improve the performance of the prediction algorithm. Dimension reduction techniques such as principal component analysis (PCA) is an unsupervised method that retains the acoustic measures that are most significant with respect to ER.

Numerous acoustic features have been proposed in ER literatures with varying levels of success in discriminating emotions. Since surveying all the features employed in ER literature is outside the scope of this chapter, here we cover only those features that are relevant to our study. For a comprehensive list of acoustic measures related to emotions as well as their mathematical definitions, a good reference book is [152].

In section 1.4, we report on the acoustic estimates that are as well as on the ER learning paradigms that have motivated our features and methods in this work.

Feature Extraction

In this section we review the types of acoustic features that are pertinent for emotion studies in voice and hence motivate our choice of features [16, 175, 186].

The following issues must be given careful consideration in feature extraction: first, what the analysis region is: should local or global features be estimated. Second what types of features are best suited for ER. Third, what kind of processing should be made to the signal. And fourth, are acoustic features alone sufficient for ER, or should we combine them with other features from different modalities such as linguistic or facial features [51].

Given a vocal segment, a length of 20-30ms is sufficient to extract vocal tract system features [102]. The Fourier transform of a vocal segment gives the short time spectrum, from which the following features can be extracted: formants and their bandwidths, spectral energy and slope. The Fourier transform of the log magnitude spectrum gives the cepstrum, from which we extract the Mel Frequency Cepstral Coefficients (MFCC).

Local versus global features Generally both local and global features have been investigated in SER [51, 160]. When feature statistics are computed over the entire range of a vocal segment, they are termed *global* features. When the voice segment is divided into small segments called *frames*, and features are extracted from these frames, then they are *local* features. Although there has been much disagreement on whether local or global features are most suitable for ER tasks, the vast majority of researchers have agreed that the performance of global features outperforms that of the local features in classification accuracy and classification time [51, 81, 154, 199, 218]. Another advantage is that global features are less numerous than local features, therefore feature selection and cross validation execute faster [51].

Types of Features The features are estimated based on their pertinence for emotion studies. Features can be of two types: prosodic and spectral, and they are both found to be correlated to affective communication in audio stimuli.

Prosodic features Prosody refers to the melody of speech and captures the emotional content of an utterance [30, 51, 102, 240]. Although in our work we do not study speech segments but rather focus on nonverbal vocalizations, to ensure consistency with the description of acoustic features in literatures, we will refer to these features as *prosodic* throughout the chapters. Prosodic cues are:

1. Pitch of the voice: low or high. Estimated by the F0 measure also known as the fundamental frequency. Mean F0 is found to be an indicator of arousal [71].
2. Length of the sound: short or long. Estimated by the duration in time units.
3. Loudness: soft or loud. Estimated by the intensity or sound pressure level in decibels (dB). Mean intensity is positively correlated with arousal [71].
4. Timbre or voice quality. Estimated by the spectral characteristics. Jitter, shimmer, harmonics-to-noise ratio (HNR) reflect voice quality properties [108, 130, 178, 189]. Some challenges are associated with the estimation of voice quality parameters and their correlation to emotional terms [128, 129, 189]:
   (a) The role of voice quality in conveying emotions is not well defined, and there are disagreements between researchers regarding the interpretations of various descriptors of voice quality such as harsh, tense or breathy [51]. In [176] tense voice is associated with anger, joy and fear, whereas lax voice is associated with sadness. However, in [144] authors associate sadness with a resonant voice quality, and breathy voice with the two emotions anger and happiness.
   (b) The estimation methods as well as the decision about what voice quality parameters to extract are difficult to resolve. The first approach models the speech signal as the output of the vocal tract filter excited by a glottal source signal [157] such as removing the filtering effect of the vocal tract. The second approach estimates the parameters of voice quality directly from the speech signal, however different researchers extract

difference acoustics to describe voice quality. [119] quantifies voice quality using jitter and shimmer features [65]. [128] extracts voice quality properties by performing various operations on pitch, the first four formants and their bandwidths.

Spectral features Estimated from the voice spectrum, they are found to convey perceptible emotional information [91, 102, 240], since they are correlates of varying shapes of the vocal tract and the rate of change in the articulatory movements [102]. Sundberg 1990 states that the spectrum is the acoustic correlate of voice quality, and it shows the frequencies of the signal formants and intensity [206]. The spectral attributes that are mostly used in SER are the Mel Frequency Cepstral Coefficients (MFCC).

Nordenberg and Sundberg 2004 point out that the long-term average spectrum (LTAS) "reflects the contribution of the glottal source and the vocal tract for the voice quality" [146]. LTAS-derived acoustics have also been used for ER in the speaking and singing voice [182, 205].

Singing voice features Classical singers usually employ a technique in which they lower the larynx, creating an additional high-frequency resonance (around 4-5 kHz) not present in other types of vocal production. This resonance, known as the singer's formant and quantified by the singing power ratio (SPR), is especially important for being heard in the presence of other instruments, for example allowing an opera singer to be heard over an entire orchestra [206]. The singer's formant is a very important perceptual feature of the singing voice.

Millhouse and Clermont 2006 quantify the singer's formant by estimating formants F2, F3 and F4. The role of the singing power ratio (SPR) in perceptual studies is investigated in [69, 108, 130, 147, 204, 228]

Omori et al. 1996 computed the SPR from a singing vowel 'ah', as a quantitative metric for the evaluation of singing voice quality [130, 147] and found it was essential for professional singers in order to be heard over an orchestra. The SPR is quantified by computing the singing power ratio (SPR): the two highest spectrum peaks between 2 and 4 kHz and between 0 and 2 kHz are identified and the SPR is obtained by computing the 'amplitude difference in dB between the highest spectral peak within the $2 - 4$ kHz range and that within the $0 - 2$ kHz range' or by measuring the ratio of the peak intensities between 2–4-kHz and 0–2-kHz frequency bands for sustained vowels [228].

Eyben et al. 2015 studied ER in the singing voice. The stimuli consisted of sentences and vocalises of the vowel 'ah' on the ascending as well as descending scale, sung by eight professional opera singers. The singing was performed in ten different affective states in addition to the neutral state. Then the states were mapped on the 2D model of VA. The set of acoustic properties was based on [182] and included: LTAS features, loudness, spectral entropy, MFCC 1-4, and F0 [55].

Voice signal processing Prior to feature extraction, the signal may undergo a *pre-processing* operation such as the removal of the silence from the beginning or the end of the vocal signal. If the recording environments are different, an energy normalization is made for all the voice samples [51].

After feature extraction, a *post-processing* operation is generally made as well. It involves some or all of the following:

1. Normalization of the features is particularly useful since feature values may have different orders of magnitude. A known method is the z-score normalization such that column features are centered to have mean 0 and scaled to have standard deviation 1

2. Dimension reduction: referred to as a *feature extraction* in some literatures [113, 236, 237], is an unsupervised algorithm that reduces the feature set to a smaller number of most discriminative features with respect to the emotions studied. A common method is principal component analysis (PCA) [89], that projects the acoustic features on a factorial plane thus illustrating graphically their distribution on the affective dimensions.

Table 1.1 lists the acoustic properties used in this thesis as well as their references in literatures.

Table 1.1: Acoustic features for VER and related literature

Features	Literature
Formants: F1 to F5	[85, 93, 141, 204, 235]
Singing Power Ratio (SPR)	[69, 108, 130, 147, 204, 228]
Mean Intensity (MI), Mean Pitch (F0)	[87, 93, 103, 150, 235]
Jitter, Shimmer, Mean HNR, Mean Autocorrelation (MAC)	[93, 108, 178, 231]
Brightness (BR)	[63, 85]
Root Mean Square (RMS)	[184, 235]
Mean Roughness(MR)	[9, 184]
Zero Crossing Rate (ZCR)	[9, 103, 184]
LTAS: Slope, Mean LTAS (MLTAS), Local Peak Height (LPH), Linear Regression Slope (LRS), Sound Pressure Level (SPL)	[182, 205]
Spectral Centroid (SC), Spectral Spread (SS), Spectral Entropy (SE), Spectral Flatness (SF)	[9, 55, 90, 91, 93]
MFCC 1 to MFCC 13	[91, 103, 173, 184, 240]
Acoustic Measures Norms	http://www.sltinfo.com/acoustic-measures-norms

1.2.3 Feature Selection

Feature selection (FS) is a process of selecting from the original feature space a subset of features. Some evaluation criteria are applied in order to evaluate the feature subsets and

select the optimal subset that retains all the information content. Feature selection methods can be filter, wrapper or embedded methods.
For the classification problem, feature selection will select the features that discriminate samples belong to different affective classes. By doing so, selected features will improve the classification performance while minimizing the classification error [24, 72].

Wrapper algorithms such as the sequential feature selection (SFS) in forward (SFFS) or backward (SBFS) directions, use the learning machine (classifier) performance to estimate the goodness or saliency of feature subsets. Embedded methods select features in the training phase of the classifier. Both wrapper and embedded methods are classifier-dependent. Filter methods are classifier-independent algorithms that select subset of features as a preprocessing step, independently of the learning algorithm i.e. classifier. They employ measurements to rank the features according to a score, and then give to the classifier the most highly scored features [166]. Common filter methods are Relieff and the mutual information (MI). An advantage of the MI method is that it can describe non-linear relationships among the features. Various criteria have been used to evaluate the goodness of the selected features, examples are: distance measures, dependency measures, consistency measures, information measures and classification error measures.

In this work, both filter and wrapper methods are applied: classifier-independent Relieff and MI, as well as classifier-dependent SFFS and SBFS. We chose the set of features that achieves the highest classification performance on a given affective dimension. A review on feature selection methods for classification is [209].

The feature selection methods tested and evaluated in the context of this thesis are reported next.

Sequential Greedy Selection

There are two main categories of greedy sequential search methods: forward (SFS) and backward (SBS). In forward search (SFS), the algorithm starts with a null feature set and, for each step, the best feature that satisfies some criterion function is included with the current feature set. The SFS proceeds dynamically increasing the number of features until the desired dimension is reached.

The backward elimination search (SBS) works analogously, but starting with the full feature set and, for each step, a worst feature is eliminated from the feature set. SBS performs the search until the desired dimension is reached.

Relieff Algorithm

The Relieff algorithm for classification is a ranking filter approach for feature selection. It weights each feature according to its relevance to the class. Initially, all weights are set to zero and then updated iteratively. In each iteration, the algorithm chooses a random instance i in the dataset and estimates how well each feature value of this instance distinguishes between instances close to i. In this process two groups of instances are selected: some closest instances belonging to the same class and some belonging to a different class. With these instances, Relieff will iteratively update the weight of each feature and it differentiates data points from

different classes while, simultaneously, recognizing data points from the same class. At the end, a certain number of features with the highest weights is selected. In an alternative version, a threshold may be used in such a way that only the features with weights above this value are selected.

Mutual Information

A major weakness of the above feature selection methods is that they are not invariant under the transformation of the features. For example PCA results can vary considerably if the input variables are linearly scaled. Furthermore the methods that work well for simple distributions of the patterns that belong to different classes, can fail in classification tasks where there are complex decision boundaries. In addition, methods that are based on linear dependence of the features, such as correlation, cannot resolve arbitrary relations between the patterns coordinates and the different classes [10].

The mutual information (MI) method is a robust feature selection method that is classifier-independent [54, 115, 122, 151]. MI evaluates the *information content* of each individual feature with respect to the output class. The fact that it is independent of the classifier permits a robust estimation of the feature subset. The MI is "the amount by which the knowledge provided by the feature vector decreases the uncertainty about the class" [10]. It is defined as:

$$I(C,F) = H(C) - H(C|F) = \sum (P(C,F) log(\frac{P(C,F)}{P(C)P(F)} \tag{1.1}$$

where the entropy H(C):

$$H(C) = -\sum_{c=1}^{N_c} P(c) log P(c) \tag{1.2}$$

with $c = 1, ..., N_c$, is the initial uncertainty in the output class, and the conditional entropy of class C given feature vector F is:

$$H(C|F) = -\sum_{f=1}^{N_f} P(f)(\sum_{c=1}^{N_c} P(c|f) log P(c|f)) \tag{1.3}$$

where *P(c|f)* is the conditional probability for class *c* given the input vector *f*.

1.3 Classification Schemes

Our work is concerned with classification tasks, which correspond to the upper part of the ER framework in figure 1.4. Various pattern recognition algorithms are employed to construct different types of classifiers for emotion recognition tasks. There is no consensus on what classifier is best recommended for ER research, and the choice of the classifier is based on rule of thumb, on previous research or on experimental evaluation [102]. Classifiers for ER can be linear or nonlinear. A linear classifier performs the classification based on a linear combination

of the feature vectors by means of linear kernel function. A nonlinear classifier makes a nonlinear combination of the feature vectors using a nonlinear kernel function.

The most widely used ones are [5, 51, 189]:

1. Linear classifiers: Linear support vector machines (SVM), perceptron classifiers, linear discriminant classifier (LDC).

2. Nonlinear classifiers: artificial neural networks (ANN), Gaussian mixture models (GMM), Hidden Markov models (HMM), decision trees (DT), nonlinear SVM, K-nearest neighbour algorithm (KNN).

KNN classifiers are most successful for natural (non-acted) emotional speech [112, 227, 239] however they suffer from the curse of dimensionality.

SVM classifiers are promising and successful in SER as well as MER, they generalize well [127, 188, 191, 192, 220, 227] and can be employed for binary as well as multi-class classification where the output classes are multiple emotions [190]. SVM classifiers are based on the use of kernel functions to map the original features to a high-dimensional space, which allows the data to be classified using a linear classifier. Kernel functions can be linear, or nonlinear such as Gaussian (radial basis function kernel or rbf), polynomial, or other [44]. Although they are not the best classifiers [139], they outperform other classifiers [113] such as KNN, LDC or DT. However they treat non-separable cases heuristically, which means that the separability of the features is not always guaranteed [51].

ANN classifiers are very effective in modeling nonlinear mappings of data, and perform better than other classifiers such as GMM or HMM. The performance of the ANN strongly depends on the following parameters: the activation function to use, the number of input layers, the number of hidden layers and the number of neurons [51].

Different classifiers do not perform equally well on all emotions. For such cases, some approaches combine multiple classifiers in what is known as ensemble learning (EL). [5, 51, 189].

1.3.1 Evaluation Metrics

When performing a machine learning task, such as a classification, various evaluation metrics are computed to evaluate the performance of the classifier in learning the classes. Generally in a classification of discrete classes, as is the case in our work, the classifier's performance is summarized by the confusion matrix (CM), that determines the number of samples correctly or incorrectly classified [57, 83]. Then several evaluation measures can be obtained from the CM and used subsequently to compare different models.

Several metrics are proposed in literature, and in order to choose the right metric for the problem, it is important to determine whether the dataset is balanced or imbalanced. For a balanced dataset, the overall accuracy (ACC) obtained from the CM is a good metric and is generally used [83, 121], in addition to the CM derived measures: the *precision* or positive predictive value (PPV), the *recall* or true positive rate (TPR) [73, 121, 132]. However in the case of imbalanced datasets [73], the CM based metrics are not sufficient. Accuracy and precision are highly sensitive to data imbalance such that if the class labels of each sample are inverted,

i.e., positive samples become negative and vice versa, and a new confusion matrix is built, the results would be poor compared to the original confusion matrix. To cope with this problem four additional measures are proposed: Cohen's Kappa (κ) [27, 28, 121], $F1$-measure, $F2$-measure [73, 121, 132] and the area under the receiver operating characteristics graph (AUC) [58, 73, 121] that measures the discrimination ability of a classifier for various threshold settings.

Given a 2 x 2 confusion matrix of a binary classifier:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}$$

Accuracy (ACC) is a proportion of correct guesses, i.e., it is the ratio of the total number of correctly labelled examples (positive and negative classes) to the total number of examples:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1.4}$$

Precision or positive predictive value (PPV) is a measure of exactness and refers to the number of correctly labelled positive samples in the total set of examples that are labelled as positive:

$$Precision = \frac{TP}{TP + FP} \tag{1.5}$$

Recall or true positive rate (TPR) is a measure of completeness and refers to the number of samples who belong to the positive class and are labelled correctly as positive.:

$$Recall = \frac{TP}{TP + FN} \tag{1.6}$$

The $F1$-measure is the harmonic mean of precision and recall and tends towards the lowest of the two. The $F2$-measure sways recall more than precision, therefore emphasizing the false negative value which is the most critical element of the confusion matrix. The F_β measure is defined by:

$$F_\beta - measure = (1 + \beta^2).\frac{Precision.Recall}{(\beta^2.Precision) + Recall} \tag{1.7}$$

$$F_1 - measure : \beta = 1, F_2 - measure : \beta = 2$$

Cohen's Kappa (κ) measures how well a classifier performed as compared to how well it would have performed simply by chance:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \tag{1.8}$$

where p_0 is the relative observed agreement among raters and is identical to ACC, p_e is the probability of agreement by chance. If raters are in complete agreement, then $\kappa = 1$. And if there is no agreement among raters, then $\kappa \leq 0$.

1.4 Related Work

Sauter et al. 2010 investigated the correlation of a set of acoustic properties of nonverbal vocalizations with the affective dimensions VA. They computed a set of features based on [8,93]: amplitude features: intensity, envelope and duration; pitch features: mean pitch and mean variation; spectral cues: spectral centroid and variation of frequencies in the spectrum. They apply statistical analyses to find correlations between the acoustic properties and emotional judgements. However no machine learning algorithm is made to examine the generalizability of the features in ER [174].

Simon et al. 2009 investigated to what extent brief vocal bursts communicate 22 emotional categories including infrequently studied emotions such as embarrassment, guilt or compassion. They found that the nonverbal voice can communicate at least 14 distinct emotional states, and that the voice is a prime modality for emotion studies, with and without words. It would have been interesting to see how acoustic measures correlate with emotions such as embarrassment, guilt or compassion, and how they would perform in machine learning tasks [201].

Schröder studied the perceived emotional content of affect bursts for 10 categories of emotions: admiration, threat, disgust, elation, boredom, relief, startle, worry, contempt and anger. He finds that the recognition of emotion from the affect bursts is $> 90\%$. However there is no analysis of acoustic cues made to identify the elements of the bursts responsible for affective expressions. Furthermore no prediction or learning algorithm is employed [186].

Juslin and Laukka 2001 investigated the communication of emotions in brief verbal phrases read aloud with weak and strong emotion intensity. The emotions studied were: anger, disgust, fear, happiness, sadness and no expression. The acoustic features were computed from the speech rate, the fundamental frequency (F0), F0 contour, intensity, attack, formants, spectral cues, jitter and articulation. Results showed that the acoustic cues captured the emotional expressions, and that portrayals of the same emotion with different intensity resulted in different patterns of acoustic measures [93]. This highlighted the role of the acoustic measures in discriminating different intensities of emotion.

Yildrim et al. 2004 used a combination of spectral and prosodic features to classify four emotions in speech: anger, happy, neutral and sad, in vowel segments of speech. The features consist of global statistics of F0 (minimum, maximum, mean, median, range and standard deviation); the first three formants F1, F2, F3; root mean square (RMS) and spectral balance. They tested a various combinations of features in recognizing emotions using Fisher's linear discriminant analysis (LDA). They achieved a best performance of 67% by using all features [235].

Ishi et al. 2004 investigated how speakers control prosodic features in speech utterances to express various emotions, intentions and attitudes for speech synthesis [85]. They use formants F3 and F4 as well as root mean square (RMS). Formants F1-F5 are also used for ER in speech in [31, 112]

Bozkurt et al. 2009 used a combination of spectral and prosodic features from a speech corpus to classify five emotions: anger, emphatic, neutral, positive, rest. They employed GMM based classifiers and obtained a classification accuracy of 63% [18]

Ververidis et al. 2004 combined spectral and prosodic features from a Danish emotional

speech corpus to classify five emotional states: neutral, surprise, happiness, sadness and anger. Using a Bayes classifier, they achieve a correct classification rate of 52% [219].

Yu et al. 2001 tested the performance of prosodic features using SVM, KNN and NN classifiers on four categorical emotions: anger, happiness, neutral and sadness. They found that SVM outperformed the other two classifiers. As an example, SVM, KNN and NN achieved accuracies of 77.16%, 42.86% and 10% respectively for anger [239].

Schuller et al. 2004 combined acoustic as well as linguistic features for the automatic recognition of a speaker's emotion [190]. Different classifiers were tested such as GMM, ANN and SVM with seven emotions: anger, disgust, fear, joy, neutral, sadness and surprise. They found that SVM was more robust than the other classifiers for ER based on acoustic information.

Zhu and Luo 2007 investigated the performance of several classifiers for ER in speech using prosodic features for six emotions: happiness, sadness, anger, fear, surprise and disgust. Using a standard neural network, they achieved a recognition rate of 80.69%. Using KNN, the accuracy reached was of 79.31%. With their proposed modular neural network, they achieved a success rate of 83.31%.

For a comprehensive review of literatures of various classifiers for ER studies, good reference papers are [5, 102, 241].

For a comprehensive review on acoustic features ER in voice and music, good references are [94, 102, 241].

1.5 Contributions

The following contributions are made in chapters 2 and 3:

Chapter 2 provides:

1. A human listening experiment of non-verbal singing voices, using vocal and glottal sounds.
    - The ground truth is obtained using categorical descriptions of the two-dimensional model of valence-arousal.
2. An identification of acoustic properties that correlate with the affective labels in vocal versus glottal sound.

Chapter 3 presents an investigation of emotion communication in musical and vocal expressions:

1. Explores the effectiveness acoustic features specific to the voice domain in capture emotions in the music domain.
2. Performs cross-modal and cross-domain approaches in order to test for the generalizability of the acoustic feature set.
3. Addressed the debate about the common origins of musical and vocal expressions of emotions.

Chapter 2

The Singing Voice - Instrument of Affect Expression

The affective potential of the speaking voice has been widely researched for over a decade and various tasks of automatic emotion recognition (ER) in speech (SER) have been presented with an aim to develop learning models that automatically recognize emotions in speech [5, 51, 90, 102, 189]. In contrast, the emotional power of the singing voice has been rarely studied in an experimental fashion [182], and therefore the role of various acoustic features of the singing voice in communicating affect is less understood.

The singing voice is the oldest musical instrument, and it is endowed with an emotional power that is unequalled [238]. It can capture our attention, and sparkle our emotions. Like music, it can be a vehicle for communication of a rich variety of information, encompassing singing style, singer identity, gender, and emotional expression. The exquisite precision with which the voice communicates emotional information, is captured by our perceptions and can trigger an immediate response: we can be delighted, perturbed, intrigued or saddened. Indeed, in order to perform vocal music in Western music traditions, professional singers such as first-class opera singers, have to be able to genuinely produce a remarkable range of emotional shadings, and they are often judged in terms of their ability to 'inhabit' the emotional feeling they are performing and therefore to express it in their vocal performance [182].

Previous studies that have investigated emotion in the singing voice have identified acoustic features from a singing voice performing words and sentences, i.e. lyrics, and having accompanying music. However words carry emotional meaning that can blur the listener's perceptions, and so does the background music.

The purpose of this chapter is to investigate the intrinsic qualities of the singing voice in conveying affective meanings to the listener. To this end, the focus will be on the nonverbal singing voice, that performs a vowel 'ah' on a musical scale, with no lyrics, nor accompanying music.

We hypothesize that the singing voice is an instrument of emotional expression, that can communicate emotions with an accuracy that is perceived by listeners regardless of lyrics or music and in the absence of any emotional intent.

In our approach, we study two types of recordings of the singing voice. The first one is

a vocal waveform that contains the actual voiced singing. And the second one is a glottal waveform that contains the recording from the glottis.

While this work was underway, to the best of the author's knowledge, no prior work had so far addressed the emotional power of the nonverbal singing voice without lyrics nor accompanying music, from both the vocal as well as the glottal signals.

The contributions of this chapter are:

1. We annotate a database of nonverbal singing voices with emotional annotations.
2. We examine potential correlations between singing expressions and affective dimensions.
3. We identify a set of acoustic measures in the vocal signal that are correlated with the emotional meaning perceived by listeners, and that are intrinsic to the singing voice.
4. We show that the glottal features are correlated with the emotion expressed in the singing voice, and that the glottal signal communicates emotions that is clearly perceived by listeners.

The remaining of the chapter is organized as follows: Section 2.1 gives a brief overview of the anatomy of the voice production, how voiced sounds are generated, what is the role of the glottis and the importance of vowels for emotion studies in voice. Section 2.2 elaborates on the human annotation experiment made using a singing voice database, as well as the acoustic set extracted from the singing voices. Sections 2.3 and 2.4 report the statistical analysis done and the results obtained. Finally section 2.6 concludes the chapter with an emphasis on the perspectives for subsequent work, as well as the disseminations.

2.1 The Voice Production

2.1.1 Anatomy of the Voice Production

"The voice organ is an instrument consisting of a power supply (the lungs), an oscillator (the vocal folds) and a resonator (the larynx, pharynx and mouth). Singers adjust the resonator in special ways."- [203]

Our ability to perceive various information from singing voices such as singer's identity as well as emotions, relies on our perceptions, that is, our auditory system, as well as on the physiological aspects of the vocal apparatus. Therefore before we proceed to the quantification of the perceptual features and their extraction from the singing voice, it is useful to understand the mechanics of singing, since they form the basis for many current representations of the singing voice.

What follows is a brief description of the anatomy of the singing voice as well as the physiological process of singing. The description is taken after [238] and [26]. The voice organ produces speaking and singing, and it consists of: the lungs or air pressure system, the larynx or the vibratory system and the pharynx or the resonating system. An illustration is given in figure 2.1.

The air pressure system consists of the respiratory system composed of the diaphragm, chest muscles, ribs, abdominal muscles and the lungs, and its function is to provide and regulate air pressure to cause vocal folds to vibrate. The lungs generate an air stream by producing an excess of air pressure, and the air stream is then converted into a glottic signal by passing through the vocal folds. This signal is then filtered by the vocal tract and converted into different audible sounds by moving the articulators, like the tongue, lips and jaw. This sound can then be made into speech by various modifications of the supra-laryngeal vocal tract.

The vibratory system consists of the larynx (also called the voice box) and the vocal folds. When the vocal folds vibrate, they change air pressure into sound waves producing voiced sounds. Vocal folds are the primary source for the production of harmonic (pitched) vocal sounds. The key function of the voice box is to open and close the glottis, which is the space between the two vocal folds. When the folds are pulled apart, or abducted, the air is allowed to pass freely through, as is the case with breathing. When the folds are pulled together, or adducted, the airflow is constricted, which is the preparatory condition for vibration. The muscles of vocal folds can alter the shape and stiffness of the folds, resulting in corresponding changes to the acoustic sound generated. Vocal fold physiology is believed to be one of the key factors in establishing voice quality.

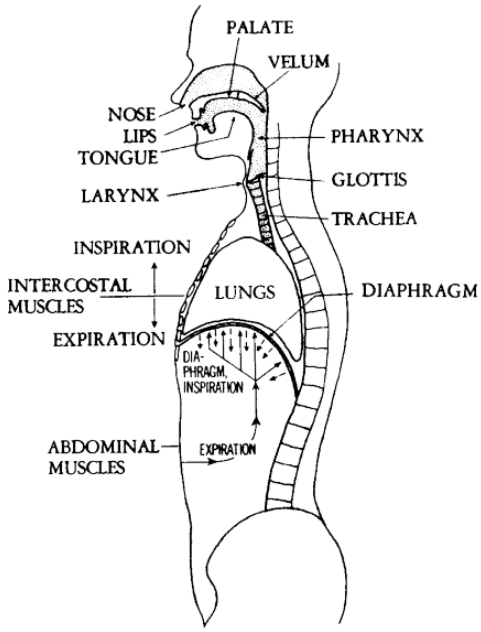

Figure 2.1: Anatomy of the voice production apparatus [26]

The resonating system is the vocal tract and consists of the throat (or pharynx), the oral cavity, and the nasal cavities. The personal quality of the voice is due to the vocal tract. An important property of the vocal tract is its ability to take a wide variety of shapes, and this has a direct impact on the acoustics of the voice. In fact this enclosed space defines the acoustic properties that describe it, therefore the physical flexibility of the vocal tract causes

an important acoustic flexibility. Vocal sounds can be voiced, unvoiced or mixed. All of the vowels are voiced.

In short, the lungs provide the energy source, the respiration; the vocal folds transform the energy into audible sound, the phonation, and the articulators transform the phonation into intelligible speech, the articulation.

2.1.2 The Singing Voice Production

During singing, the voice box brings both vocal folds to the midline to allow vocal fold vibration. It also adjusts vocal fold tension to vary pitch and changes the volume. This process is known as phonation. In voiced sounds, phonation results in a largely harmonic sound source. For unvoiced sounds, the vocal folds remain open and the breath pressure results in free airflow through the larynx into the mouth where it is impeded by a constriction (caused by the tongue, soft palate, teeth, or lips), generating a sound source resulting from air turbulence. Some vocal sounds require both phonation and turbulence as sound sources and are referred to as mixed sounds [1].

In all three cases, the source (phonation, turbulence, or both) is modified by the shape of the vocal tract (throat, mouth, nose, tongue, teeth, and lips). Each shape creates a different acoustic filter, further colouring the overall sound. This description of the human voice is the basis of the source-filter model which is the foundation of the majority of modern voice research.

2.1.3 Consonants versus Vowels

Consonants have a greater degree of constriction than vowels, and in acoustic studies, they are less prominent than vowels because vowels are more intense than the consonants that surround them. Although some consonants can have a total intensity that is greater than that of the adjacent vowels, vowels are almost always more intense at low frequencies than adjacent consonants. Furthermore, research on emotion recognition shows that because vowels have a longer utterance time and a larger amplitude compared to a consonant, they are the most efficient and audible sounds, therefore vowels have a more dominant impact on the listener's impression than does a consonant.

In the most common classical singing technique, known as bel canto, singers learn to sustain vowels long enough between other phonemes which can make it easier to automatically determine the vowel from analysis of the signal. Classical singers usually employ a technique in which they lower the larynx, creating an additional high-frequency resonance (around 4-5 kHz) not present in other types of vocal production. This resonance, known as the singer's formant or singing power ratio (SPR) is especially important for being heard in the presence of other instruments, for example allowing an opera singer to be heard over an entire orchestra [206].

[1] https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/

2.1.4 Waveforms of the Singing Voice

There are two types of waveform that can be extracted from voice: vocal and glottal. As mentioned earlier, the source for the production of voiced-sounds is the airflow passing through the glottis, modulated by the oscillations of the vocal folds. The glottal airflow (thereof glottal signal) is a low-frequency signal, that can have many variations due to the degree of tension on the vocal folds that the speaker control.

Though the vocal waveform is still the primary form used in research on affect recognition from voice, the glottal signal carries information about gender as well as age. Furthermore, there is evidence for a strong correlation between the features derived from the glottal waveform and emotional states. For example, in [33–35], researchers found that the glottal waveform is considerably affected by the excessive tension in the laryngeal musculature, created by various emotional states. This modification of the glottal waveform will be reflected in the acoustic parameters extracted from it.

A graphical representation of the glottal and vocal signals is given in figure 2.2 and figure 2.3 respectively. Figure 2.2 shows the waveform and the power spectral density (PSD) of a glottal signal. As can be noticed from the PSD, the energy of the glottal signal is mostly located in the lower part of the spectrum.

In this chapter, various acoustic features are extracted from the vowel of a glottal and vocal waveform of a singing voice. The singing voice is nonverbal, performing the vowel 'ah' on a musical scale. Only the first note of the musical scale is considered. First, the recordings of the singing voice are segmented to retain the first note, then acoustic features are extracted. A user study is carried to annotate the segments with valence-arousal, described with the four adjectives: pleasant-unpleasant, and awake-tired, as explained in the Chapter 1 of this work. The method is detailed in the following sections.

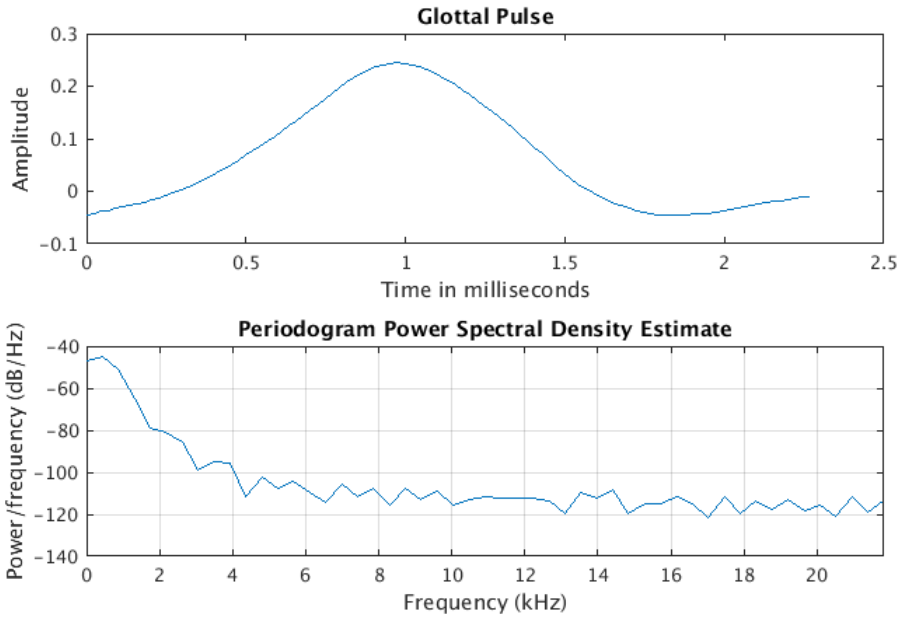

Figure 2.2: Waveform and Power Spectral Density of a Glottal Signal

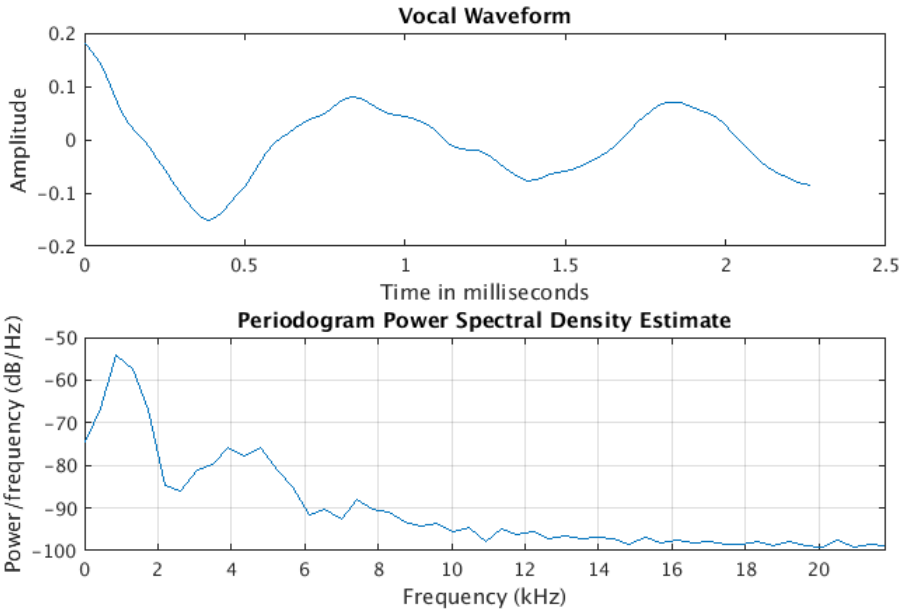

Figure 2.3: Waveform and Power Spectral Density of a Vocal Signal

2.2 Our Approach

Our approach follows the framework depicted in figure 2.4

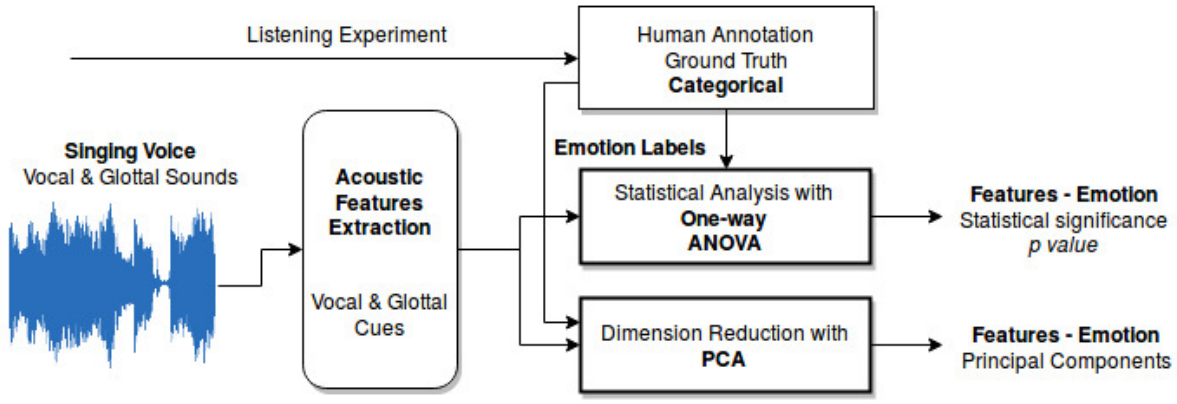

Figure 2.4: Acoustic features extraction and statistical analysis framework

2.2.1 Human Affective Annotations

In order to collect listeners' emotional judgements on vocal as well as glottal sounds, an experiment was conducted, and the details elaborated next.

2.2.2 Stimuli

The stimuli is taken from the Singing Voice Database [2] (SVDB) and consists of scale recordings of vocal as well as glottal sounds of a sung vowel 'ah' interpreted by professional singers (1 male and 1 female). The musical notes range from A2 to E4 and A3 to A5 for male and female voice respectively. The recordings are mono files in WAVE PCM format, at 16 bits and 44100 Hz sampling rate. As a first processing of the signals, a trimming was done to remove the silence at the beginning of each sound. The trimming was done using MIRToolbox in MATLAB [106]. Then the recordings were segmented using R Statistical software [210] in order to retain the first 'C' note of the ascending musical scale. The final stimuli consists of 44 sounds, of which 22 are vocal, and 22 are glottal. Each sound is of 1 second duration. The singing was performed in different singing expressions and these are: bounce, hallow, light, soft, sweet, flat, mature, sharp, clear, husky and no expression. The female singing did not include the flat singing expression.

[2] http://crel.calit2.net/projects/databases/svdb

The stimuli was relevant to the current study for the following reasons: first, the singing is nonverbal and consists of only one vowel, hence the perceived emotion is not influenced by lyrics; second the singer did not perform any specific target emotion, which is advantageous to understand the role of the singing voice in conveying emotion even when there is no emotional intent in the performance; and third there was no accompanying music to influence the emotional judgements made by the listeners.

2.2.3 Participants

A questionnaire was given to the participants to collect the following information: age, gender, whether the participants had some kind of formal singing training, whether they enjoyed singing, whether they thought that music expresses emotion and whether the voice of the singer was important to their personal enjoyment of a song. There was a total of 15 participants that consisted of 9 males and 6 females (age Mean = 26.1, σ = 9.6), of whom 1 is a professional singer, and 4 have had some kind of formal singing training. 7 reported enjoying singing, 14 agreed that music expresses emotions, and that the voice of the singer was important to their personal enjoyment of a song. All participants reported that the voice of the singer is important in expressing emotions in singing.

2.2.4 Procedure

Participants were asked to rate the perceived affect dimension for each of the voice samples on a 5-point Likert scale ranging from using the two-dimensional model of affect represented by the four broad affect terms: pleasant-unpleasant for valence, and awake-tired for arousal. Today's internet bandwidth and sound technologies have made it possible to conduct psychoacoustic tests over the internet [32], therefore for the purpose of practicality the experiment was distributed through email with instructions explaining the objectives. The participants could play the sound file as many times as they desired, and they had the choice to save their answers and come back complete the experiment later. Each sample occurred 3 times and the order of the sample was randomized. The duration of the experiment was of 45 minutes.

2.2.5 Affect Responses

Considering that the number of responses for each of the 5 categories on the Likert scale was slight, therefore it was difficult to meet the assumptions of statistical validity, and therefore the responses were grouped under 'pleasant', 'unpleasant', 'neutral' for valence, and 'awake', 'tired' and 'neutral' for arousal. For example responses for 'awake' and 'extremely awake' were grouped under 'awake'. Then we computed the mean of the ratings for the 3 occurrences of each sound and then the sound was annotated with the emotion that had the highest total number of votes. On the valence dimension, 16 were annotated as pleasant (13 vocal, 3 glottal), 22 were labelled as unpleasant (7 vocal, 15 glottal) and 6 were labelled as neutral (2 vocal, 4 glottal). On the arousal dimension: 22 were labelled as awake (17 vocal, 5 glottal), 17 were labelled as tired (3 vocal, 14 glottal) and 5 were rated as neutral (2 vocal, 3 glottal).

2.2.6 Acoustic Parameters

The physical phenomenon of producing speech is traditionally represented by a source-filter model. This model remains a reference point for identifying the main acoustic properties of a vocal signal.

Acoustic descriptors used for characterizing affect in voice are intended for modeling modifications in acoustic signals related to physical changes in either the glottis or the vocal tract. Prosodic and voice quality parameters describe changes in the acoustic signal of the glottis, and formants as well as cepstral coefficients convey changes in the acoustic signal of the vocal tract.

An initial set of 11 acoustic features were extracted. The choice of the features was made according to their relevance for emotion studies as established in the literature (see table 1.1 in chapter 1). The features were extracted using Praat software [15]. The settings are the following: pitch information was obtained using a cross-correlation method for voice research optimization, with pitch floor and ceiling set to 75 Hz and 300 Hz respectively for male singing and to 100 Hz and 500 Hz respectively for female singing. The spectrum was obtained from the waveform using the Fast Fourier Transform method (FFT) setting a dynamic range of 70 dB, a window length of 5 ms and a view range from 0 to 5000 Hz for male and from 0 to 5500 Hz for female singing. The singer's formant was quantified by computing the singing power ratio (SPR): the two highest spectrum peaks between 2 and 4 kHz and between 0 and 2 kHz are identified and the SPR is obtained by computing the 'amplitude difference in dB between the highest spectral peak within the 2 – 4 kHz range and that within the 0 – 2 kHz range'. Furthermore, considering that the 'perceptual singer's formant' is contributed by the underlying acoustic formants F2, F3 and F4' [141], the means of F2, F3 and F4 were measured individually for each sound file. The mean intensity of the sound was obtained using energy averaging method. The voice quality was quantified with measures of pitch perturbation (jitter), amplitude perturbation (shimmer), mean autocorrelation and mean harmonics-to-noise ratio (HNR). Brightness was obtained using the MIRToolbox.

2.3 Analysis I

The entire analysis was done in R statistical software [210].

2.3.1 Singing expressions and affect dimensions

In order to determine the relationship of various singing styles to the affective dimensions, a cross-tabulation was made between the singing expressions and the emotional annotations given by the participants, and the counts of the combination of each factor level was determined.

The results are reported for the vocal as well as glottal recordings of the singing voices.

On the valence dimension, the singing voices perceived as pleasant are: all the vocal sounds in light, soft, sweet expressions as well as 67% of the voices performing with no specific expression. All the soft glottal sounds. The singing voices perceived as unpleasant are: mature

and sharp vocal sounds, and all the glottal sounds in bounce, husky, mature, sharp and no expression.

On the arousal dimension, the singing voices perceived as awake are: all the vocal sounds in bounce, clear, mature, sharp and no expression. The singing voices perceived as tired are: the vocal sounds in soft and sweet expressions. All the glottal sounds in clear, hallow, soft, sweet as well as 67% of the sounds having no singing expression.

2.3.2 Acoustics and Emotion

Considering the size of the acoustic set as well as the size of the stimuli was small, a one-way ANOVA was performed for each acoustic measure with valence and arousal as factors with three levels each. The analysis results were verified using Tukey's multiple comparisons of means and were Bonferroni corrected. Furthermore, statistical values of mean (M) and standard deviation (σ) are computed for the features that have a significant p value on the affective dimensions.

Results: On the valence dimensions, the acoustic features whose means are statistically different on the pleasant-unpleasant factors for the vocal signals are: SPR, mean intensity, jitter, shimmer, mean autocorrelation and mean HNR for the vocal signals (Table 2.1), and for the glottal signals they are: brightness, mean intensity, shimmer and mean autocorrelation, with brightness being significant for the pleasant-neutral factors as well (Table 2.2). It can be noticed that the mean values of shimmer and mean intensity are higher for glottal signals and lower for vocal signals on the pleasant dimension. And mean autocorrelation's mean value is lower for glottal signals and higher for vocal signals.

On the arousal dimension, the acoustic features whose means are statistically different on the awake-tired factors for the vocal signals are: SPR and mean intensity (Table 2.3). For the glottal files the are: mean intensity, jitter, shimmer, mean HNR and mean pitch, with mean HNR being also significant for the neutral-awake factors (Table 2.4). For the glottal sounds the mean intensity is higher than for the vocal sounds.

Table 2.1: p values, mean and stdev of vocal features on pleasant-unpleasant

		Pleasant		Unpleasant	
Acoustic Features	p	M	σ	M	σ
SPR	0.002	28.58	7.475	15.90	5.998
Mean Intensity	0.001	53.94	5.026	62.67	1.834
Jitter	0.023	0.653	0.334	1.166	0.471
Shimmer	0.032	3.591	0.857	5.123	1.71
Mean Autocorrelation	0.004	0.985	0.006	0.952	0.031
Mean HNR	0.001	21.44	2.131	16.20	3.305

Table 2.2: p values, mean and stdev of glottal features on pleasant-unpleasant

		Pleasant		Unpleasant	
Acoustic Features	p	M	σ	M	σ
Brightness	0.032*	0.163	0.105	0.069	0.018
	0.003				
Mean Intensity	0.027	59.42	7.172	65.89	2.601
Shimmer	0.011	7.583	3.865	3.75	1.528
Mean Autocorrelation	0.042	0.97	0.026	0.988	0.007

*pleasant-neutral

Table 2.3: p values, mean and stdev of vocal features on awake-tired

		Awake		Tired	
Acoustic Features	p	M	σ	M	σ
SPR	0.002	19.13	6.689	32.430	7.275
Mean Intensity	0.002	60.31	3.898	52.50	5.394

Table 2.4: p values, mean and stdev of glottal features on awake-tired

		Awake		Tired	
Acoustic Features	p	M	σ	M	σ
Mean Intensity	0.02	68.84	2.335	63.51	3.911
Jitter	0.007	0.56	0.136	1.285	0.423
Shimmer	0.023	2.314	0.704	5.257	2.263
Mean HNR	0.006	29.32	3.019	23.57	3.328
	0.031*				
Mean Pitch	0.014	218.70	3.937	138.10	51.543

*awake-neutral

2.3.3 Features Transform

A dimension reduction with PCA is performed. Figures 2.5 and 2.6 show the two components that were retained that explained 78.1% of the total variance of the vocal measures. The first component (PC1) obtained from the analysis of the vocal features explains 57.7% of the original variance and accounts mainly for variations in SPR, F4, mean pitch, opposed to jitter and mean intensity; and the second component (PC2) explains a further 20.4% of the original variance and accounts mainly for variations in brightness.

Figures 2.5 and 2.6 show that the vocal features projected onto the PC1-PC2 planes appear to cluster reasonably well according to valence and arousal, although a bit weaker for arousal. For example, pleasant vocal sounds are those having higher values of SPR, F3, F4 and mean pitch and lower values for jitter and mean intensity and/or lower values for brightness. Unpleasant sounds are those having higher values for jitter, shimmer and mean intensity and lower values for SPR, mean autocorrelation and mean pitch, and/or higher values for brightness. High energy sounds (awake) are those having higher values for jitter and mean intensity and rather lower values for brightness, and low energy sounds are those having higher values for SPR, mean pitch and F4 and/or lower values for brightness.

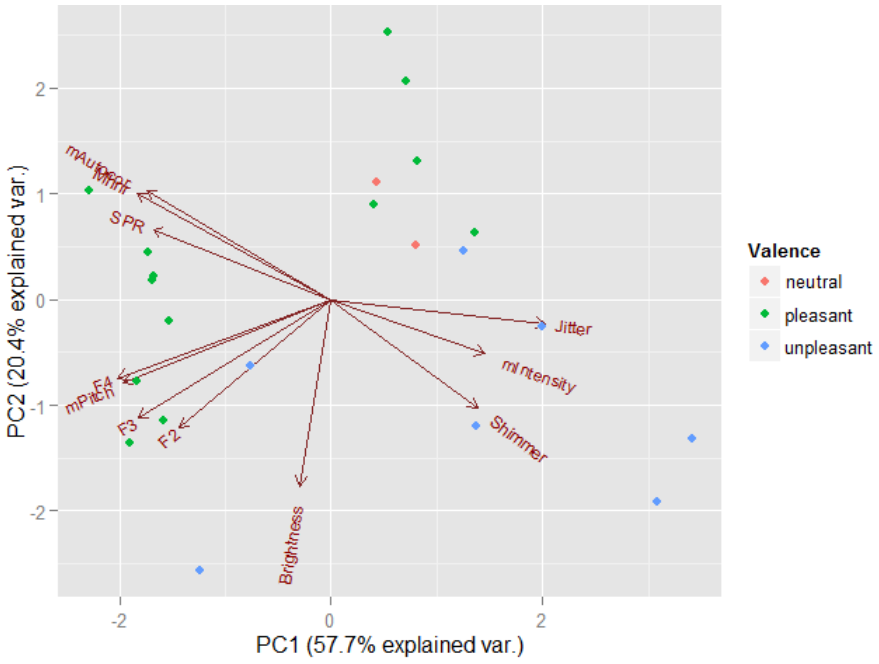

Figure 2.5: First two components of vocal features clustered by Valence

Figure 2.7 and figure 2.8 show the first two components that were retained and that explained 73.5% of the total variance of the glottal measures. The first component explains 53.4% of the original variance and accounts for variations in shimmer, F2, F3, opposed to mean intensity, mean autocorrelation and mean HNR; the second component explains a further 20.1% of the original variance and accounts for variations in mean pitch and F4.

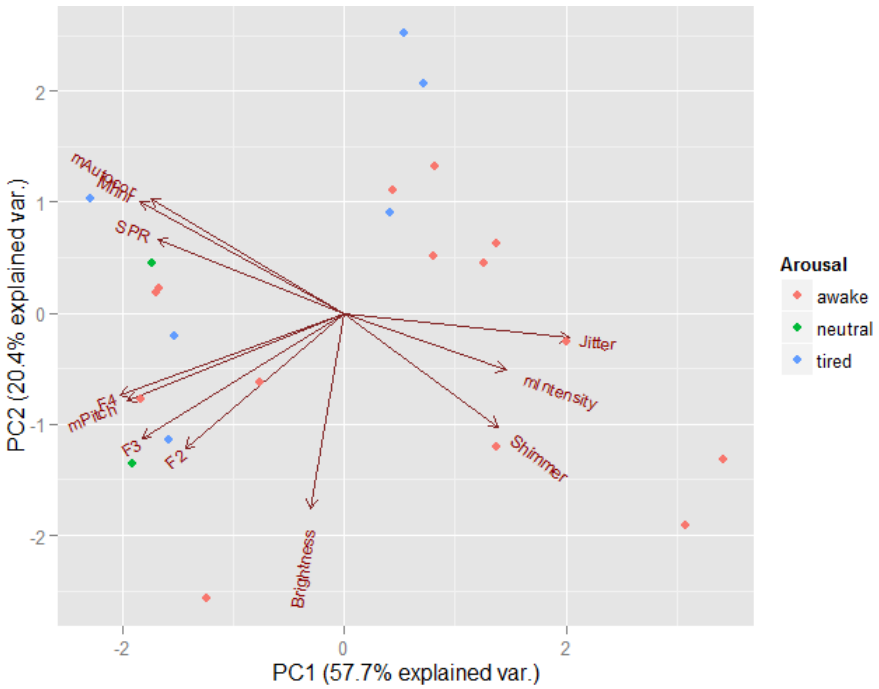

Figure 2.6: First two components of vocal features clustered by Arousal

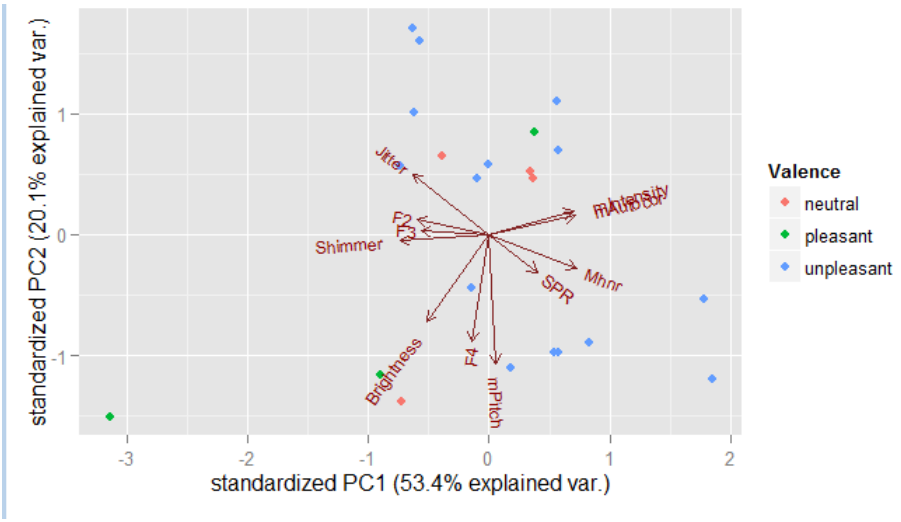

Figure 2.7: First two components of glottal features clustered by Valence

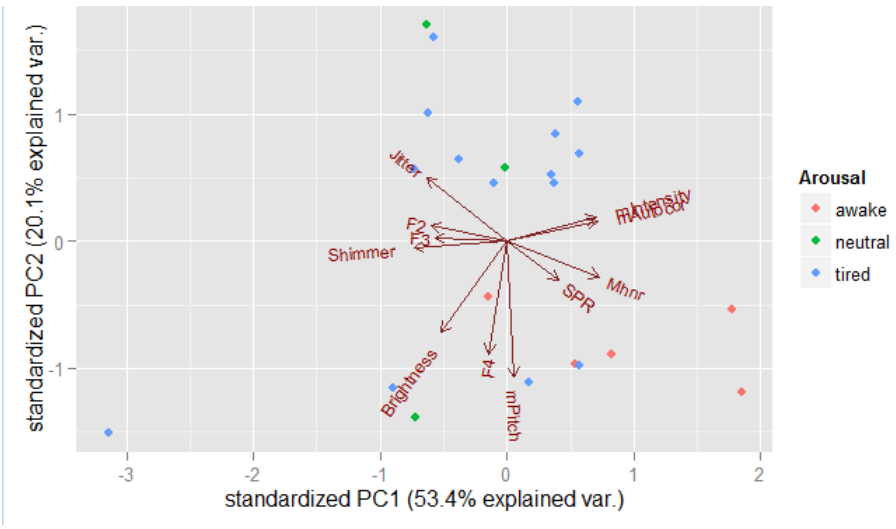

Figure 2.8: First two components of glottal features clustered by Arousal

2.4 Analysis II

In this part, prosodic and spectral acoustic features are extracted from the vocal and glottal signals. Research on emotion recognition from voice has shown that prosodic and spectral features contain emotional information and are therefore suitable for voice emotion recognition (VER) tasks. A total of 22 features were extracted, and their validity with respect to VER is established in the literature (table 1.1 in chapter 1). The new acoustic measures that were added to the previous set of 11 features are: spectral features extracted from the pitch-corrected long-term average spectrum (LTAS) with an analysis bandwidth of 100 Hz, a maximum frequency of 5000 Hz for male and 5500 Hz for female voice with pitch floor and ceiling values adjusted for male and female voice samples. The LTAS features are: sound pressure level (SPL), mean LTAS, local peak height (LPH), slope, spectral tilt (also called linear regression slope (LRS).

2.4.1 Extended Acoustic Features

New ANOVA tests were made for each acoustic measure with the affective dimensions as factor with three levels each. Again the analysis results were verified using Tukey's multiple comparisons of means and were Bonferroni corrected.

On the valence dimension and for pleasant-unpleasant factors, the acoustic features whose means were statistically different are SPR, mean intensity, jitter, shimmer, mean autocorrelation, mean HNR, mean LTAS, RMS, SPL and LPH for the vocal signals (Table 2.5), and brightness, mean intensity, shimmer, mean autocorrelation, B1 and F5 for glottal signals, with brightness being significant for the pleasant-neutral factors as well (Table 2.6). Contrasting vocal and glottal mean values, the shimmer's mean value in vocal sounds is lower for the pleasant dimension and higher for the unpleasant dimension. This is consistent with the thresholds of pathology[3]whereby a shimmer value $> 3.810\%$ is considered to be a sign of potential pathol-

ogy, which may explain why listeners judged files with a mean shimmer value of 5.120% as unpleasant. The same comment can be made for the jitter means of the vocal sounds, being <=1.040% (threshold[3]) for pleasant files and >1.040 for unpleasant sounds, as well as for mean HNR means being <201 for unpleasant sounds, potentially indicating a noticeable hoarseness. The mean values of the mean intensity are higher for glottal signals, on both pleasant and unpleasant dimensions.

On the arousal dimension and for the awake-tired factors, the acoustic features whose means are statistically different are SPR, mean intensity, mean LTAS, RMS, LTAS slope for the vocal signals (Table 2.7), in addition to B1, B4 for awake-neutral and tired-neutral, and LPH for awake-neutral (Tables 2.9 and 2.10). For the glottal sounds, the features are mean intensity, jitter, shimmer, mean HNR, mean pitch, F5, mean LTAS, RMS and SPL, with mean HNR and F5 being also significant for th awake-neutral and tired-neutral factors respectively (Table 2.8). Contrasting the mean values of both types of signals mean intensity, RMS and mean LTAS have higher values in glottal files for both affect dimensions.

In summary, the features whose means are statistically different for both affect dimensions are SPR, mean intensity, mean LTAS, RMS and LPH for vocal signals, and mean intensity and shimmer for glottal signals. Those that are significant for both vocal and glottal signals are mean intensity, shimmer and mean autocorrelation for valence, and mean intensity and shimmer for arousal.

Table 2.5: p values, mean and stdev of extended vocal set on pleasant-unpleasant

		Pleasant		Unpleasant	
Acoustic Features	p	M	σ	M	σ
SPR	0.002	28.58	7.475	15.90	5.998
Mean Intensity	0.0006	53.94	5.026	62.67	1.834
Jitter	0.03	0.653	0.334	1.166	0.471
Shimmer	0.032	3.591	0.857	5.123	1.71
Mean Autocorrelation	0.004	0.985	0.006	0.952	0.031
Mean HNR	0.001	21.44	2.131	16.20	3.305
Mean LTAS	0.080*	16.86	4.95	25.34	3.742
	0.0007				
RMS	0.06*	0.109	0.006	0.026	0.008
	0.0001				
SPL	0.09*	54.10	4.80	62.45	3.66
	0.0007				
LPH	0.091	0.38	5.61	6.46	6.25

*pleasant-neutral

[3]www.sltinfo.com/acoustic-measures-norms/

Table 2.6: p values, mean and stdev of extended glottal set on pleasant-unpleasant

		Pleasant		Unpleasant	
Acoustic Features	p	M	σ	M	σ
Brightness	0.03*	0.163	0.10	0.069	0.01
	0.003				
Mean Intensity	0.02	59.42	7.17	65.89	2.60
Shimmer	0.011	7.58	3.86	3.75	1.52
Mean Autocorrelation	0.042	0.97	0.02	0.98	0.007
B1	0.022	187	83.06	533.40	421.78
F5	0.130*	4727	206.61	4518	142.54
	0.074				

*pleasant-neutral

Table 2.7: p values, mean and stdev of extended vocal set on awake-tired

		Awake		Tired	
Acoustic Features	p	M	σ	M	σ
SPR	0.002	19.13	6.68	15.90	5.998
Mean Intensity	0.002	60.31	3.89	52.50	5.39
Mean LTAS	0.064*	22.90	4.07	15.93	5.12
	0.016				
RMS	0.082*	0.021	0.008	0.009	0.005
	0.009				
LTAS Slope	0.006	−7.340	4.60	−14.390	2.38

*pleasant-neutral

Table 2.8: p values, mean and stdev of extended glottal set on awake-tired

		Awake		Tired	
Acoustic Features	p	M	σ	M	σ
Mean Intensity	0.02	68.84	2.33	63.51	3.91
Jitter	0.007	0.56	0.13	1.28	0.42
Shimmer	0.02	2.31	0.70	5.25	2.26
Mean HNR	0.006 0.03*	29.32	3.01	23.57	3.32
Mean Pitch	0.014	218.70	3.93	138.10	51.54
F5	0.120*	4571.00	184.26	469.00	233.60
Mean LTAS	0.095	29.42	1.88	25.05	5.71
RMS	0.08+ 0.19	0.039	0.008	0.025	0.012
SPL	0.183+ 0.095	66.56	1.92	62.24	5.53

*awake-neutral
+tired-neutral

Table 2.9: p values, mean and stdev of new vocal cues on awake-neutral

		Awake		Neutral	
Acoustic Features	p	M	σ	M	σ
B1	0.037	285.20	185.78	931.305	1109.203
B4	0.0001	293.631	162.66	1074	490.675
LPH	0.15	4.64	5.56	−4.408	8.75

Table 2.10: p values, mean and stdev of new vocal cues on tired-neutral

		Tired		Neutral	
Acoustic Features	p	M	σ	M	σ
B1	0.0530	275.058	164.886	931.305	1109.203
B4	0.0002	325.63	91.283	1074.03	490.675

2.4.2 Dimension Reduction II

A new PCA was made on the acoustic correlates in order to reduce the acoustic features set to a few number of features that are most discriminative with respect to valence-arousal. A series of tests were carried on various combinations of acoustic correlates.

First, the entire set of 24 features was analysed and the first two components were retained. Figures 2.9 and 2.10 show the PCA analysis for the vocal descriptors. PC1 and PC2 accounted for a mere 47.5% and 21.5% respectively of the total variance. Figures 2.11 and 2.12 show the results for the glottal descriptors. PC1 and PC2 accounted for 40.63% and 20.62% respectively of the total variance. It is clear from the figures that both the vocal and glottal measures were quite dispersed on both affective dimensions.

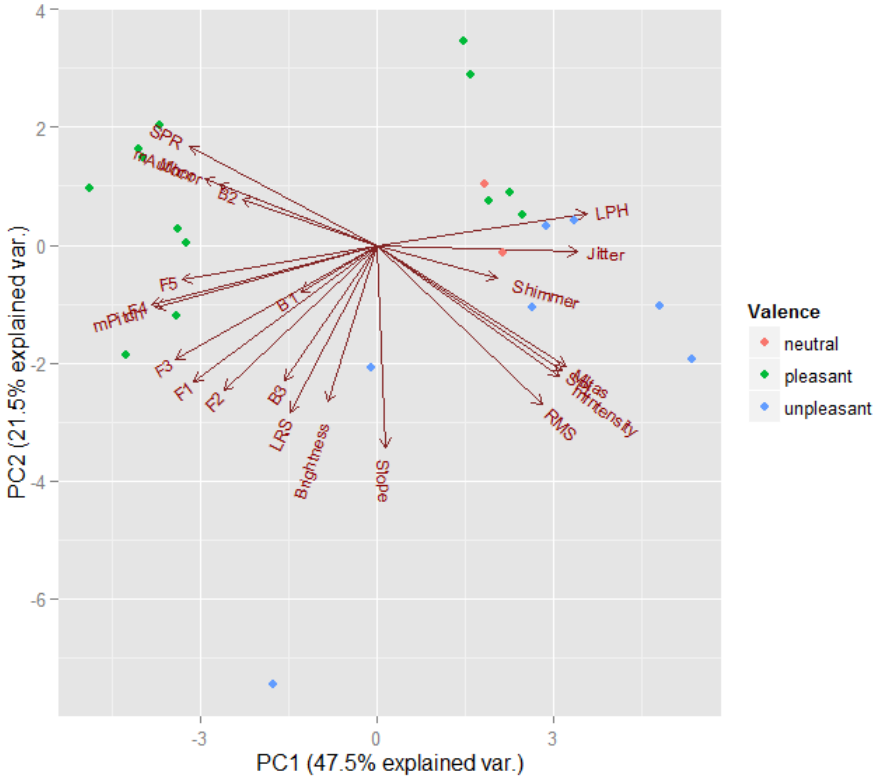

Figure 2.9: PC1 and PC2 of all the vocal features clustered by Valence

Second, an analysis is made on the remaining set of acoustic cues composed of F1, F5, Slope, B1, B2, B3, RMS, SPL, mean LTAS, LPH, LRS and B4. The first two components of vocal cues explained merely 42.7% and 25.5% of the original variance, and those of the glottal cues explained 32.6% and 25.1% of the original variance. See figures 2.13, 2.14 and 2.15 and 2.16.

A third and final test was made with a feature set consisting of selected prosodic and spectral features. The idea is to identify the largest possible set of features that is best accounted for by the components and that separates the sound files fairly well on the affective dimensions. The first two components of vocal features explain 86.3% of the original variance and those of the

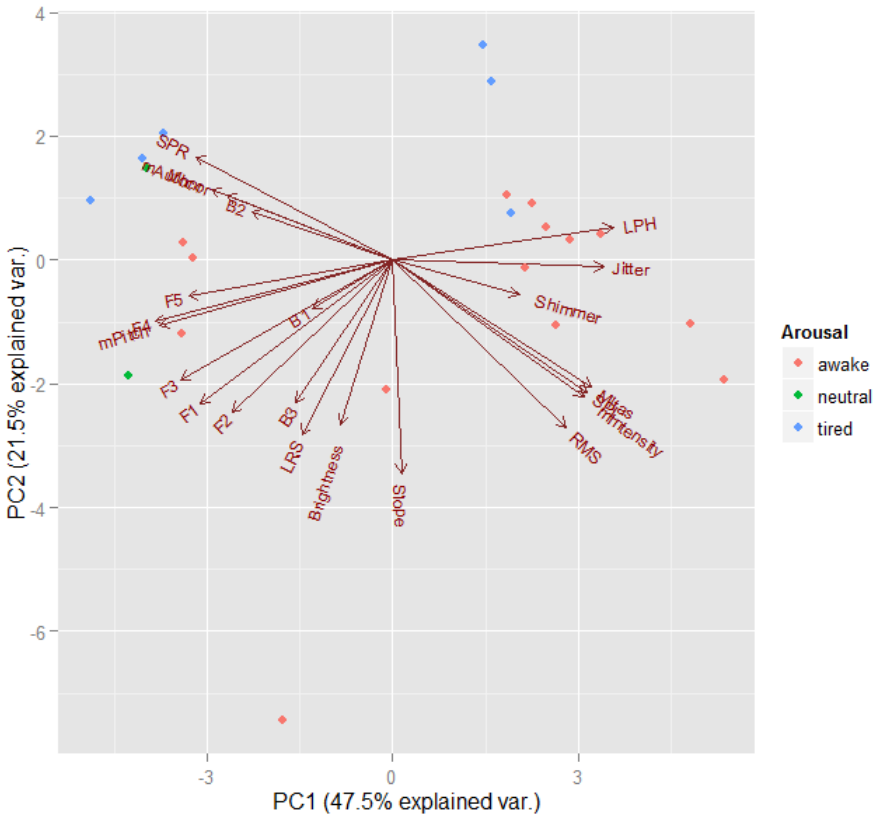

Figure 2.10: PC1 and PC2 of all the vocal features clustered by Arousal

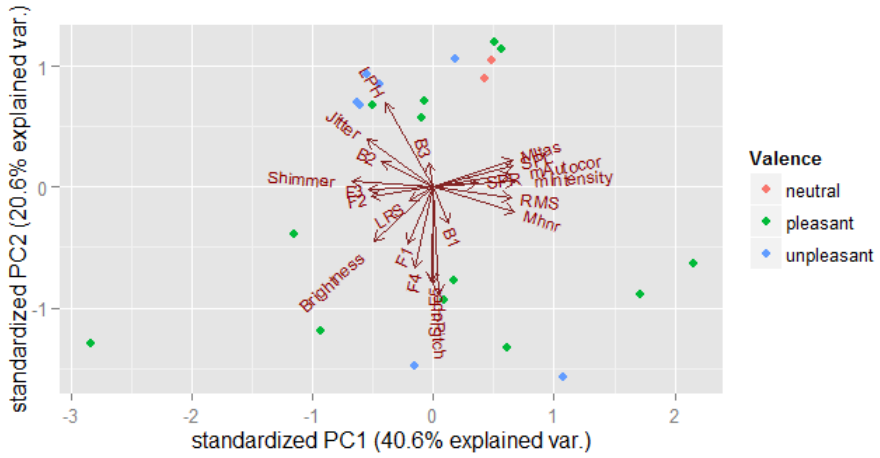

Figure 2.11: PC1 and PC2 of all the glottal features clustered by Valence

Figure 2.12: PC1 and PC2 of all the glottal features clustered by Arousal

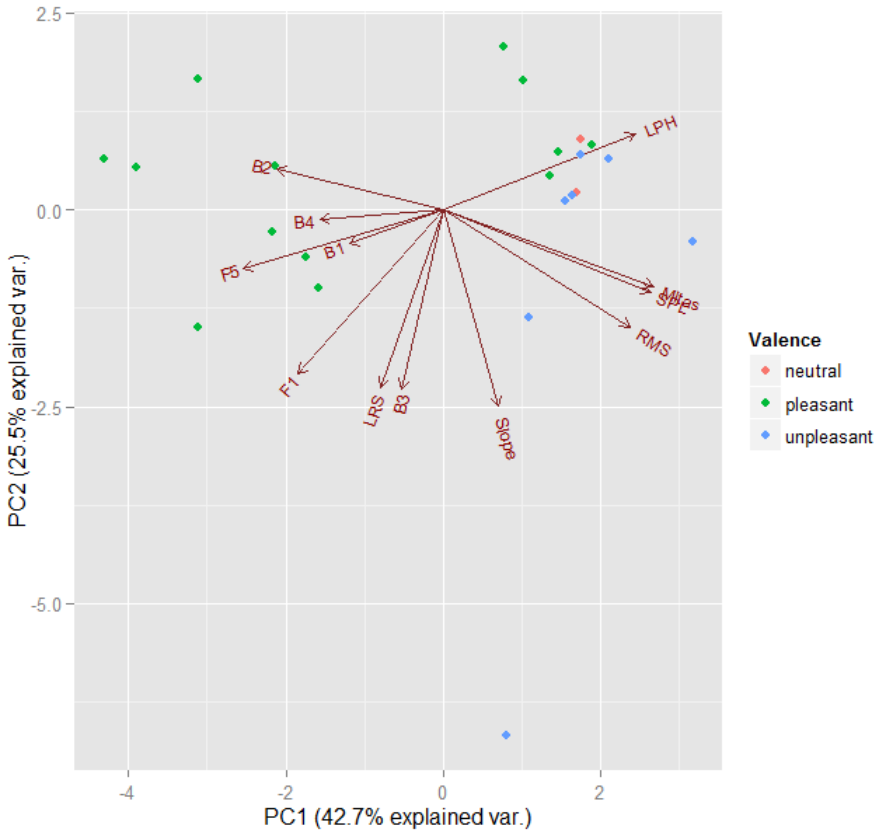

Figure 2.13: PC1 and PC2 of a subset of vocal features clustered by Valence

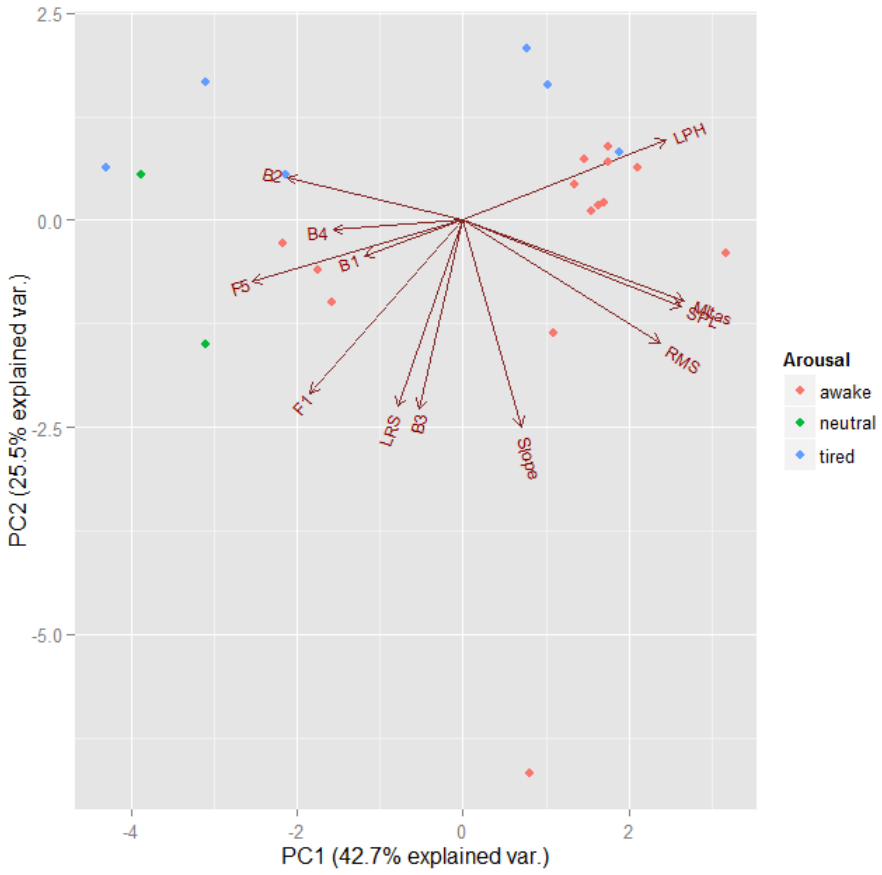

Figure 2.14: PC1 and PC2 of a subset of vocal features clustered by Arousal

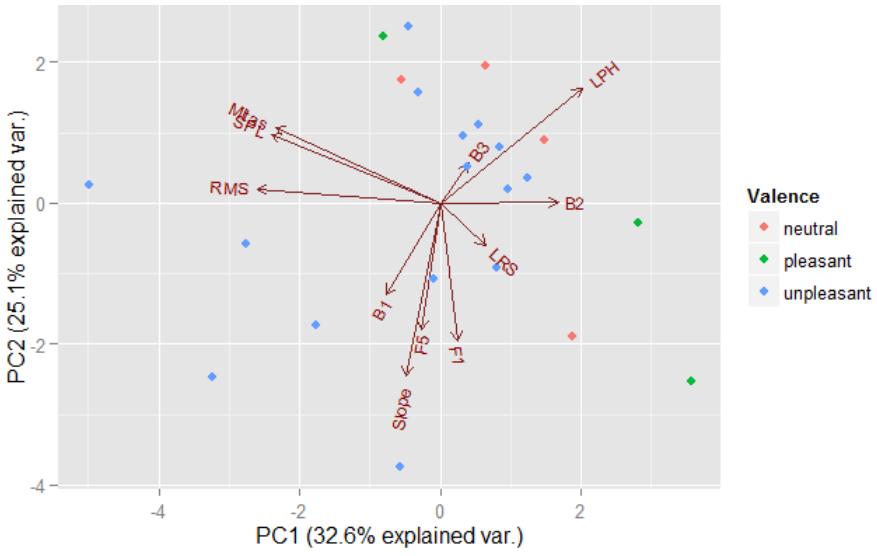

Figure 2.15: PC1 and PC2 of a subset of vocal features clustered by Valence

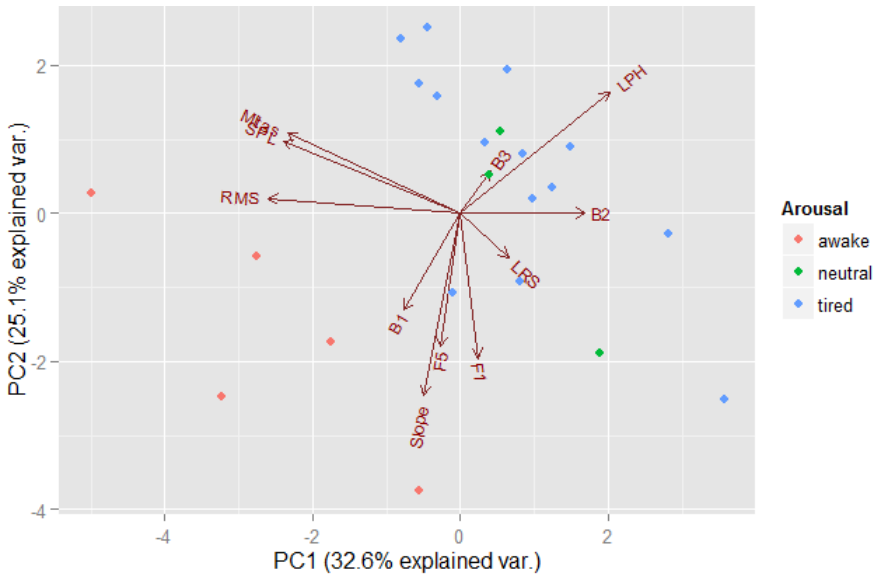

Figure 2.16: PC1 and PC2 of a subset of vocal features clustered by Arousal

glottal features explain 86.7% of the original variance (see figures 2.17 and 2.18). A summary of the PCA tests and their first two components are in table and table .

For vocal cues, the first component explains 61.1% of the original variance and accounts mainly for variations in SPR, F4, F5, and mean Pitch opposed to LPH and mean LTAS; the second component explains a further 25.2% of the original variance and accounts for variations in LRS and Slope. Figure 2.17 shows that the vocal files projected onto the PC1-PC2 planes cluster fairly well on valence, with a better clustering noted for the pleasant-unpleasant factors than in the previous result. For example, unpleasant files are those having high values for LPH, mean LTAS, SPL and slope and low values for F4, F5 and mean pitch. Pleasant files are those having high values for F4, F5, mean pitch and SPR, and low values for mean LTAS, SPL, LPH and mean intensity. Figure 2.18 shows an improved clustering of the vocal files on the awake-tired factors over the previous test result; sounds perceived as tired are those having high values for SPR and LPH and low values for RMS, slope, mean intensity and SPL. For glottal cues, the first component explains 62.9% of the original variance and accounts for variations in mean LTAS, SPL, mean autocorrelation, mean intensity, RMS and mean HNR opposed to Shimmer and Jitter; the second component explains a further 23.8% of the original variance and accounts for variations in Slope, mean pitch opposed to LPH. Figures 2.19 and 2.20 show that the glottal files projected on the PC1-PC2 plane cluster better for the pleasant-unpleasant factors than in the previous test result and rather quite better for the awake-tired factors. For example, unpleasant sounds are those having high values for slope, mean pitch, mean HNR, mean LTAS and/or low values for jitter and LPH. Sounds perceived as awake are those having high values for slope, mean pitch and mean HNR and low value for LPH and jitter. Sounds perceived as tired are those having high values for LPH, Jitter, mean LTAS and low values for slope and mean pitch.

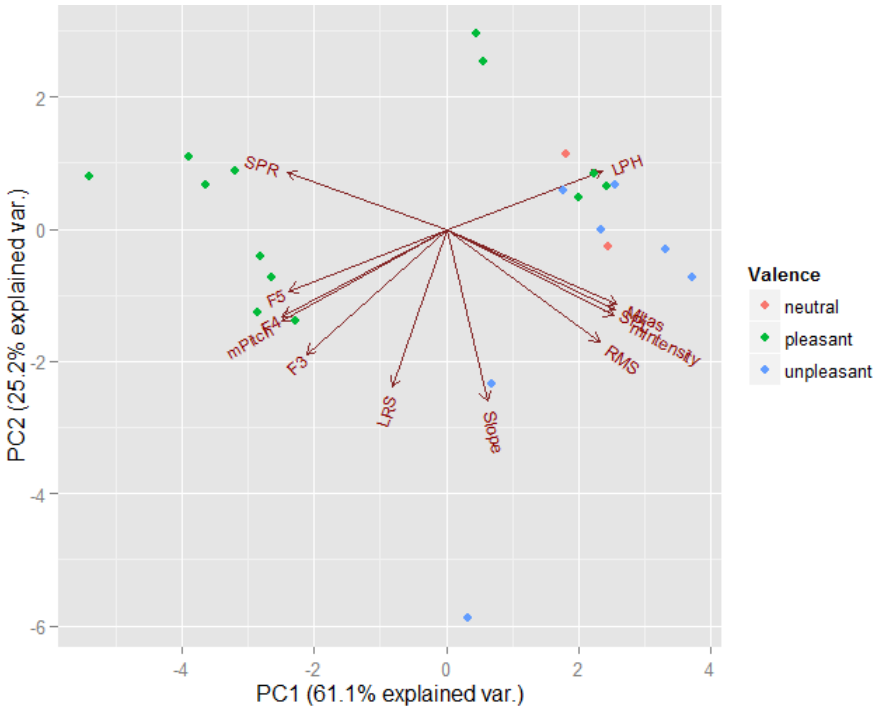

Figure 2.17: PC1 and PC2 of spectral and prosodic vocal features clustered by Valence

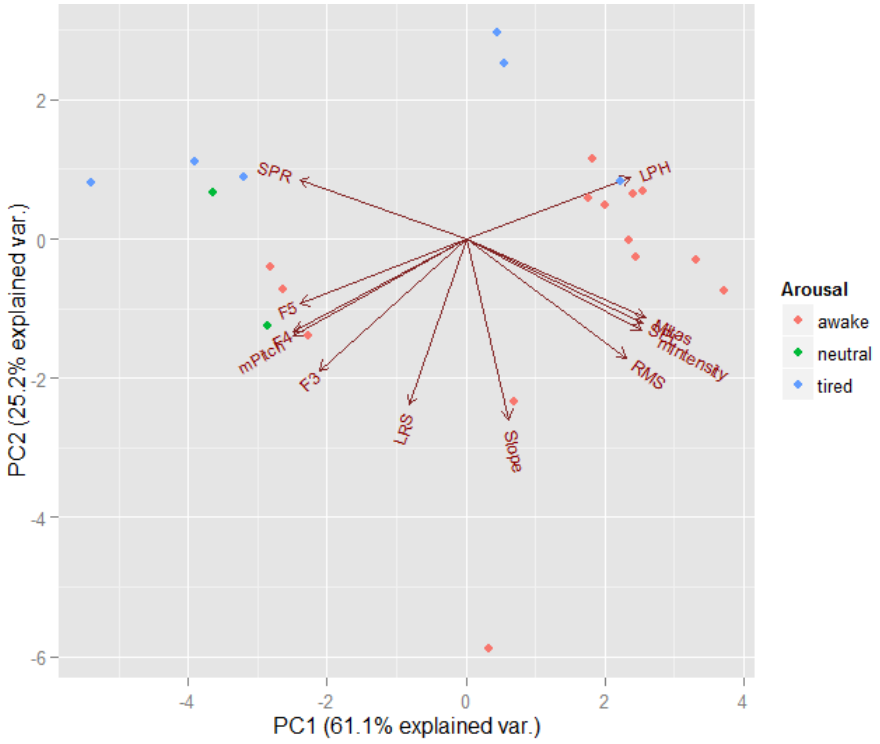

Figure 2.18: PC1 and PC2 of spectral and prosodic vocal features clustered by Arousal

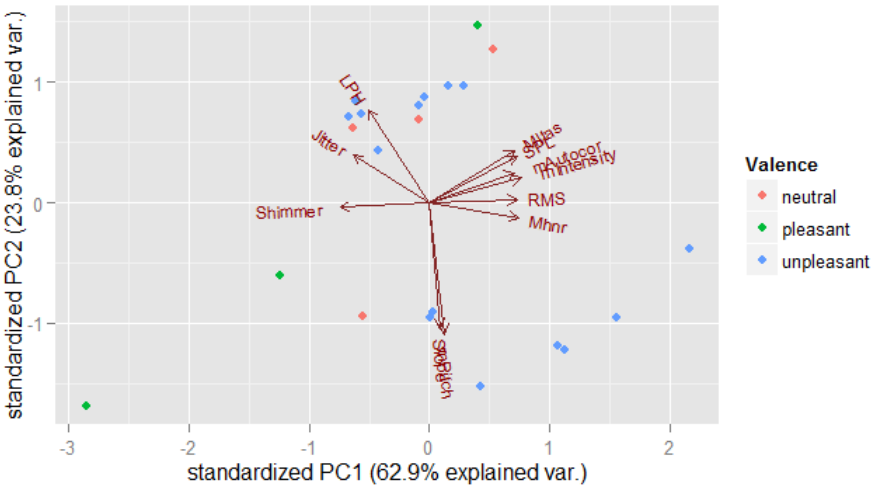

Figure 2.19: PC1 and PC2 of spectral and prosodic glottal features clustered by Valence

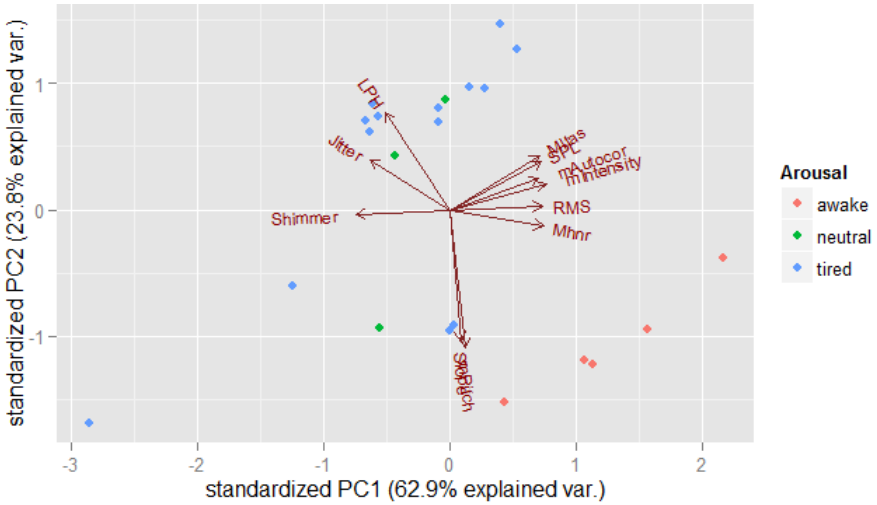

Figure 2.20: PC1 and PC2 of spectral and prosodic glottal features clustered by Arousal

2.5 Discussion

The experiment results obtained from the first test show that broad affect dimensions are perceived in the singing voice independently of the singer's emotional intent. The acoustic analysis revealed 8 features that explained the variance with respect to emotion; these are: SPR, mean intensity, brightness, jitter, shimmer, mean pitch, mean HNR and mean autocorrelation. PCA revealed 2 components that accounted for variations in 11 acoustic features. These results were encouraging to investigate the role of a larger set of acoustic measures in the perception of affect in vocal as well as glottal singing signals.

In the second experiment, a broader set of acoustic measures were studied and the statistical tests revealed 15 features that are statistically significant with respect to broad affect dimensions. Although the problem of recommending an optimal feature set for studies investigating the perception of affect in the singing voice is still an open problem, some recommendations can be made concerning the pertinent features to extract depending on whether vocal or glottal files are studied. A feature set that is appropriate for the discrimination of both affect dimensions consists of SPR, mean intensity, mean LTAS, RMS, LPH in vocal recordings, and mean intensity and shimmer for glottal sounds. A feature set consisting of cues that are significant for both file types includes mean intensity, shimmer, and mean autocorrelation for valence, and mean intensity and shimmer for arousal.

The PCA that is made on selected prosodic cues reveals 2 components that explain 78.1% and 73.5% of the original variance of vocal and glottal files respectively, and account for variations in a total of 11 acoustic cues, namely: SPR, mean pitch, jitter, mean intensity, brightness, shimmer, F2, F3, F4, mean autocorrelation and mean HNR.

The PCA made on a combination of prosodic as well as spectral cues reveals 2 components that explain 86.3% and 86.7% of the original variance of vocal and glottal files respectively, and accounted for variations in a total of 15 acoustic cues, namely: SPR, F4, F5, mean pitch, mean intensity, mean autocorrelation, mean HNR, shimmer, jitter, SPL, RMS, LPH, mean LTAS, LRS and Slope. A summary of the principal components and their corresponding features for each sound type, is found in tables 2.11 and 2.12.

Table 2.11: PC1 and PC2 of vocal features

Set of Vocal Features	PC1	PC2
SPR, F1, F2, F3, F4, F5, mean Pitch, mean HNR, mean Autocorrelation, Jitter, mean Intensity, Shimmer, Brightness, B1, B2, B3, RMS, LRS, Slope, LPH, SPL, mean LTAS	47.5%	21.50%
SPR, F2, F3, F4, mean Pitch, mean HNR, mean Autocorrelation, Jitter, mean Intensity, Shimmer, Brightness	57.70%	20.4%
F1, F5, B1, B2, B3, B4, RMS, LRS, Slope, LPH, SPL, mean LTAS	42.70%	25.50%
SPR, F3, F4, F5, mean Pitch, mean Intensity, RMS, LRS, Slope, SPL, LPH, mean LTAS	61.10%	25.20%

Table 2.12: PC1 and PC2 of glottal features

Set of Glottal Features	PC1	PC2
SPR, F1, F2, F3, F4, F5, mean Pitch, mean HNR, mean Autocorrelation, Jitter, mean Intensity, Shimmer, Brightness, B1, B2, B3, RMS, LRS, Slope, LPH, SPL, mean LTAS	40.60%	20.60%
SPR, F2, F3, F4, mean Pitch, mean HNR, mean Autocorrelation, Jitter, mean Intensity, Shimmer, Brightness	53.40%	20.10%
F1, F5, B1, B2, B3, RMS, LRS, Slope, LPH, SPL, mean LTAS	32.60%	25.10%
Shimmer, Jitter, mean Pitch, mean Autocorrelation, mean Intensity, mean HNR, RMS, mean LTAS, Slope, LPH, SPL	62.90%	23.80%

2.6 Concluding Remarks

In this chapter, we conducted an experiment where we asked participants to annotate a database of singing voices with broad emotional descriptions. We investigated a set of acoustic features extracted from the vocal as well as the glottal waveform of the singing voice. A series of ANOVA statistical tests revealed the relevance of each feature with respect to the emotions perceived by the listeners. A dimension reduction algorithm with PCA was done to reduce the feature set size and keep only the features that explain the total variance of the set. As a result, we selected a set of spectral and prosodic features that are most relevant to the task of emotion recognition in nonverbal singing voice. Furthermore, the importance of the glottal waveform has been demonstrated as carrying important affective information that is captured by relevant acoustic parameters.

The findings reveal that the singing voice per se is intrinsically an instrument that conveys affective information clearly perceived by listeners, and that the absence of accompanying music and lyrics does not impact the emotional eloquence of this instrument. These findings are relevant for both the vocal and glottal sounds of singing.

2.7 Dissemination and Contribution

The work of this chapter appears in:

1. Pauline Mouawad, Myriam Desainte-Catherine, Anne Gegout-Petit and Catherine Semal, "The Role of the Singing Acoustic Cues in the Perception of Broad Affect Dimensions", in proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research, CMMR 2013.

2. Pauline Mouawad, Myriam Desainte-Catherine, Anne Gegout-Petit and Catherine Semal, "On the Perception of Affect in the Singing Voice: a Study of Acoustic Cues", in Sound, Music and Motion, pp. 105—121, Springer LNCS 2014.

3. Part of the Singing Voice Database (SVDB) has been annotated with affect dimensions and is now freely available for research purposes.

Chapter 3

Did Music and Voice Originate from a Common Affective Auditory Scenery? Insights from Acoustic Analysis on an Ongoing Debate

The previous chapter questioned the emotional expressiveness of the singing voice performing without an emotional intent, excluding lyrics and accompanying music. It made an acoustic-based analysis of the vocal as well as the glottal sounds and extracted a set of acoustic properties. Then using statistical analysis, it identified for each sound, the features that are highly correlated with the emotional dimensions of VA. This showed that the affective information is perceived in nonverbal singing voices regardless of the singer's emotional intent.

This chapter extends the analysis to nonverbal affective vocalizations. These are short interjections [182] that express spontaneous emotions such as laughters, screams or sighs. As a first aim, it investigates to what extent human affect is communicated through a shared acoustic code in the singing voice and the affective vocalizations. The second objective concerns a debate among scientists and philosophers, about whether musical sounds developed from primitive affective vocalizations [182]. Although recent works from psychology have addressed this question, there has been no systematic examination of the generalizability of acoustic features across voice and music. Particularly, to what extent spontaneous vocal and musical bursts share common acoustic properties of affect expression? Is it possible to build a holistic ER model for voice and music?

We hypothesize that if music evolved from primitive affective vocalizations, an exploration of the emotional expressions in the musical analogue of the vocal affective bursts would be expected to verify the claim. We address the following questions:

1. Is there a shared acoustic code of affect expression in singing voices as well as affective vocalizations?

2. Are acoustic features domain-specific, or do they generalize across vocal and musical sounds?

3. Is a holistic ER model capable of predicting emotions in voice as well as music?

3.1 Motivation and Related Work

There is an ongoing debate among scientists and psychologists about the common origins of musical and vocal expression of emotions. In his article "The Origin and Function of Music", Spencer 1857 stated that music is closely related to vocal expressions of emotion [94,202]. Both modalities are primary means of emotional communications, and both are nonverbal channels that rely on acoustic signals for emotional expression. Yet why should we care if the similarity in emotion communication exists in voice and music? As a possible answer, Juslin and Laukka 2003 suggest that this could shed the light on why we perceive music as a medium of emotion communication. What's more, it would lend support to the 'controversial hypothesis' that speech and music evolved from a common origin. In this respect, Bhatara et al. 2014 suggest that their common origin may be due to their ability to communicate emotions [13,94].

In our investigation, we don't seek to support or refute the similarity, but we aim to provide an insight through acoustic analysis and machine learning techniques on whether the acoustic code suggests a parallel between vocal and musical communication of emotion. The importance of this task is highlighted by Bhatara et al. 2014, who state that so far researchers in the domains of music and speech research have worked independently and rarely exchanged, which has bounded the research to one domain or the other [13]. In the following, we highlight the few researches that made a cross-domain investigation of the topic.

Some research suggests that "music expresses emotion by imitation and exaggeration of acoustical properties of emotional expression in the voice" [17], and that both speech and music may have evolved from primitive affective bursts [182]. In [182], the authors compare the acoustic patterns of emotional communication in speech and singing. Using statistical tests they find "many similarities in vocal emotion expression across speech and singing". Although they conclude that these similarities suggest "a parallel evolution of speech and music from primitive affective bursts", they don't include in their study musical sounds nor primitive affective vocalizations. Furthermore machine learning methods aren't implemented to assess the generalizability of the acoustic profiles across the singing voices and speech.

In [155] the roles of melodic and rhythmic contrasts in communicating emotion in speech and music, are investigated. The study found similarities in the use of "rhythmic contrastiveness between successive note durations", and differences "in the use of pitch contrasts as emotional information in the two domains". This suggested that these measures may "signal emotional intentions differently in speech and music". However no cross-corpora evaluation of features was made, and no machine learning methods were implemented for voice and music.

[186] investigates the expression of emotion in non-verbal vocalizations and finds a mean recognition rate of 81%. However the underlying acoustical cues are not studied. In a study of non-verbal vocal expressions of emotion, [174] investigates the relationship of affective perception in non-verbal vocal signals with their acoustical profiles. However, to what extent the results relate to non-verbal music is not addressed in the study. A review of research on non-verbal vocalizations can be found in [13,230].

Weninger et al. 2013 investigated the shared acoustic features in speech, music and general sound events [230]. They performed within-domain and cross-domain regression tasks. They find a "high degree of cross-domain consistency in encoding" valence and arousal and conclude that this may be due to the common origins of speech and music from affect bursts.

The present work takes this research a step further, and investigates the vocal encoding of emotion in singing voices and affect vocal bursts. Furthermore, it explores ER in the musical imitation of the affective vocalizations.

3.2 Our Approach

The framework is described in figure 3.1 and is described in the following subsections.

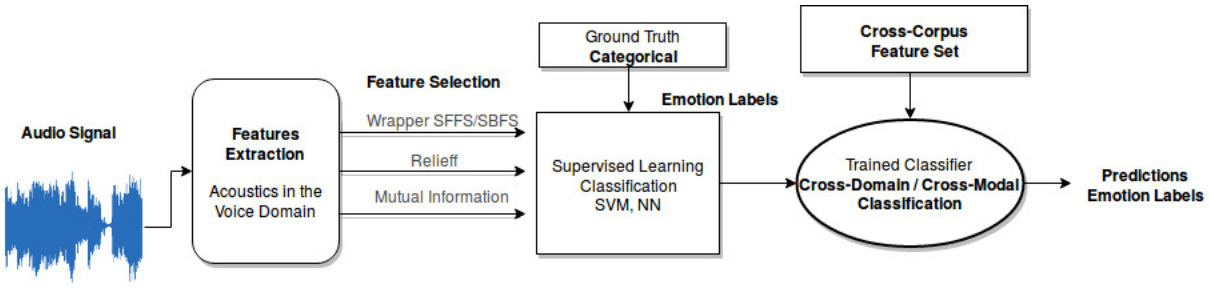

Figure 3.1: Hybrid Feature Selection with Cross-Corpora Classification Framework

3.2.1 Acoustic Features

In addition to the acoustic set detailed in Chapter 2, 28 new features are extracted resulting in a hybrid feature set of 48 content-based prosodic and spectral features. 6 features are discarded after feature transformation (see subsection 3.2.2).

The pertinence of the feature set for emotion recognition studies can be seen in table 1.1 of chapter 1. The additional 22 features investigated in this chapter are listed next. More elaborate details on the features as well as their respective definitions are included in the Appendix.

1. Spectral Centroid (SC): describes the timbre of a given sound, such as how sharp a musical or vocal sound is to the auditory perceptions. It is considered a measure of auditory brightness.

2. Spectral Spread (SS): mainly used to differentiate between noise-like and tone-like sounds, it describes the shape of a power spectrum and indicates whether the spectrum is concentrated in the vicinity of its centroid or is spread out over the spectrum [120].

3. Spectral Entropy (SE): describes the complexity of the signal, and is computed from the normalized power spectral density (PSD).

4. Spectral Flatness (SF): describes the flatness properties of the spectrum of an audio signal within a given number of frequency bands [120].

5. Mel Frequency Cepstral Coefficients (MFCC): MFCCs are widely used in speech processing such as in automatic speech recognition, speaker recognition or language recognition. Furthermore, they are also used for emotion recognition in speech. MFCCx are based on a signal's cepstrum, which has the advantage of allowing the separation of the respective contributions of the source and the vocal tract [**?**]. 13 MFFC coefficients are studied in this chapter.

6. Formant bandwidths (B_{Fn}): The formant bandwidth B_{Fn} of a formant frequency F_n, is defined as the frequency region in which the amplification differs less than 3 dB from the amplification at the centre frequency. The first formant F_1 has a corresponding bandwidth B_{F1}.

Data Preprocessing

The amount of data that is currently available on the internet has increased exponentially in terms of dimensionality as well as sample size and this large number of input features usually contains a high level of noise. Therefore an interesting challenge for researchers is to select from the high-dimensional data set a subset of features that contain the entire information while discarding noisy i.e. irrelevant features and redundant features.

A popular technique that removes noisy and redundant features is called dimensionality reduction, which involves feature transformation and feature selection.

3.2.2 Feature Transformation

After feature extraction from the signal, a feature transformation with PCA is made and the first two components are retained. As a result 6 descriptors are discarded resulting in a final set of 42 acoustic features. Then we employ filter as well as wrapper feature selection methods (see chapter 1): classifier-independent Relieff and MI, as well as classifier-dependent SFFS and SBFS. And we retained the set of features that achieves the highest classification performance on a given affective dimension.

3.2.3 Classification Schemes

Three different classification schemes are implemented: intra-domain, cross-modal and cross-domain. The intra-domain scheme refers to the task of training and testing a classifier on the same dataset, which is divided into training and testing sets. The cross-modal and cross-domain schemes are both cross-corpora learning tasks.

The cross-modal scheme refers to the task of training a classifier on a voice corpus such as the singing voice, and testing the classifier on another voice corpus such as the affective vocalizations, and conversely. A similar task is made for the musical stimuli where a classifier is trained on the clarinet corpus and tested on the violin corpus, and conversely. The objective

of this task is to probe the generalizability of the features as well as the classifier for different modalities of vocal expressions.

The cross-domain task refers to the task of training a classifier on a given corpus, and testing it on another. For example, training a classifier on a voice corpus, i.e. singing voice or affective vocalizations, and testing it on a musical corpus, i.e. clarinet or violin corpus.

In all the schemes, a multiclass one-versus-all classification is made per dataset and one classifier is retained for each of the six affective classes (see table 2).

3.3 Preliminary Experiments

.

In order to determine if our set of features from the voice domain have a good predictive quality for vocal and musical sounds, it is important to make sure that our results are not biased by the choice of the feature selection method nor the choice of the classifier. Therefore, we experiment with various feature selection methods, and evaluate the performance of multiple classifiers using confusion matrix performance measures, as well as the receiver operating curve (ROC). In total we build four models per corpus per affective dimension and evaluate their performances and then we propose the model with the highest performance.

For each classification scheme, four learning approaches are implemented: in a first approach, classifier-independent feature selection algorithm Relieff is applied and the feature subset is input to the following classifiers: K-nearest-neighbour (KNN) with the Euclidean distance metric, linear discriminant analysis (LDA), support vector machines (SVM) with the sequential minimal optimization (SMO) algorithm, ensemble learning (EL) using the Adaboost method and decision trees (DT). LDA was not suitable for the SVDB since the number of observations was inferior to the number of features, so it was discarded from the analysis.

In a second approach, classifier-dependent wrapper feature selection methods (SFS) are applied. The evaluation of the classifiers' performance was superior to the results in the first approach, therefore only where the SFS methods overfitted, we used Relieff for feature selection.

In a third approach, a hybrid feature selection method is applied. First the MI feature selection then SFS feature selection on the resulting subset. Then the resulting subset is input to a SVM.

In a fourth approach, the MI feature selection with the criteria of maximum-relevance minimum-redundancy (mRmR) is done as a preprocessing step, then the subset was input to a feedforward neural network (FNN) and the performance of the classifier was evaluated.

After evaluation of the performance of each of the four models, only model 2 and model 4 were retained. For comparison purposes, results from the two models are included later in the chapter and contrasted.

The details of each approach are described next.

Wrapper Feature Selection with Support Vector Machines

1. Given a dataset of observations – the affect labels – and predictors – the acoustic measurements –, standardize the predictor values and divide the set randomly into an 70%

training set and a 30% testing set.

2. Using the training sets of predictor values and response variables:
   (a) Fit a one-versus-all learning model for each of the 6 emotion categories.
   (b) Cross-validate the model with a 20 times repeated 10-fold cross-validation. For the SVDB a 20 times repeated 5-fold cross-validation was made due to the small size of the dataset.
   (c) Compute the mean and standard deviation (σ) of the generalization error (GE).
3. Tune the model by randomly modifying the parameter values of each classifier:
   (a) For KNN: k values ranged from 1 to 20.
   (b) For SVM, if the kernel type is:
      i. Linear: a grid search is made to find an optimal C value.
      ii. Gaussian: a grid search is made to find optimal C and γ values.
      iii. Polynomial: the degree varied from 1 to 3.
   (c) For Ensemble Learning: the tree, KNN and discriminant learners were tested with default values.
4. Apply the wrapper SFFS and SBFS. Where SFS overfits the selection, Relieff algorithm is used and the highest-ranking features are selected.
5. Refit the model with the selected features, repeat the cross-validation step and compute σ and mean of the GE.
6. Predict the emotions of the testing set and compute the GE. Get the prediction accuracy (PA) from the confusion matrix. The model is evaluated based on the GE as well as the PA.

MI Feature Selection with Neural Networks

1. Given a dataset of observations and predictors, standardize the predictor values.
2. Using the MI method with the mRmR criteria for each affective dimension, select a subset of salient features.
3. Divide the subset randomly into a 70% training set and a 30% testing set.
4. Train a simple feedforward neural network with one hidden layer and one output layer, with 10 hidden neurons.
   (a) Validate and test the network
   (b) Compute the mean squared error

(c) Compute the training, validation and test performance values.

5. Tune the model:

   (a) Reinitialize the weights and biases of the network

   (b) Test with different training functions namely, Levenberg-Marquardt,Bayesian Regularization, Scaled Conjugate Gradient and Resilient Backpropagation.

   (c) Increase the number of hidden neurons.

   (d) Increase the number of hidden layers.

6. Repeat step 4.

3.3.1 Intra-Domain Classification

The intra-domain learning task is implemented for each dataset. The means of the generalization error (GE) and of the prediction accuracy (PA) were computed and results for valence and arousal classification are reported in tables 3.1, 3.2, 3.3, and 3.4. The results are obtained using SVM as well as KNN classifiers. In every table, part (a) shows the results without feature selection, and part (b) shows the results after applying a sequential feature selection algorithm.

Table 3.1: MAV Classification performance

(a) SVM without feature selection

MAV	Mean GE	Mean σ GE	Mean PA
Valence	0.220	0.026	78.33%
Arousal	0.120	0.039	90.00%

(b) SVM with SFS feature selection

Mean GE	Mean σ GE	Mean PA
0.060	0.023	78.33%
0.077	0.007	91.67%

(c) NN with MI feature selection

MAV	Mean GE	Mean PA
Valence	0.17	95.26%
Arousal	0.05	97.63%

In tables 3.1, 3.2 and 3.3, the PA is considerably higher for arousal, which is consistent with previous findings that the "presence of vocals generally increases ratings of arousal but not of valence" and that this effect is "not limited to verbal lyrics but appears to generalize to non-verbal songs containing the human voice" [126]. Results in Table 3.2 show that the discrimination of affect using vocal features is high for the clarinet bursts, this may be due to the resemblance of the clarinet with the voice [158].

Results in Table 3.3 show that the communication of affect in violin bursts is captured by vocal cues, possibly due to the "voice-like" character of the violin that is perceived as a "super-expressive voice" [92]. However a larger dataset is needed to verify the generalizability of the results.

Table 3.2: Clarinet Classification performance

(a) SVM without feature selection

Clarinet	Mean GE	Mean σ GE	Mean PA
Valence	0.150	0.048	93.33%
Arousal	0.140	0.045	100.00%

(b) SVM with SFS feature selection

Mean GE	Mean σ GE	Mean PA
0.056	0.028	76.67%
0.067	0.033	96.67%

(c) NN with MI feature selection

MAV	Mean GE	Mean PA
Valence	0.06	100%
Arousal	0.2	100%

Table 3.3: Violin Classification performance

(a) SVM without feature selection

Violin	Mean GE	Mean σ GE	Mean PA
Valence	0.210	0.069	66.66%
Arousal	0.230	0.018	76.66%

(b) SVM with SFS feature selection

Mean GE	Mean σ GE	Mean PA
0.227	0.027	80.00%
0.163	0.034	93.33%

(c) NN with MI feature selection

MAV	Mean GE	Mean PA
Valence	0.01	96.67%
Arousal	0.03	100%

Table 3.4: SVDB Classification performance

(a) SVM without feature selection

SVDB	Mean GE	Mean σ GE	Mean PA
Valence	0.12	0.03	77.77%
Arousal	0.38	0.13	77.77%

(b) SVM with SFS feature selection

Mean GE	Mean σ GE	Mean PA
0.07	0.02	80.67%
0.16	0.04	82.33%

(c) NN with MI feature selection

MAV	Mean GE	Mean PA
Valence	0.03	100%
Arousal	0.08	100%

Results in table 3.4 show that the proposed vocal set captures the affective dimensions of the SVDB with a PA of 80.67% for valence and 82.33% for arousal, and a mean GE of 7% and 16% respectively.

3.3.2 Cross-Corpora Evaluation

In this section various cross-corpora learning tasks are implemented with the aim to identify the feature selection methods as well as the classifiers that achieve the highest emotion recognition accuracy across different datasets. The results from the different tasks are analysed and compared. The model achieving the highest performance accuracy is recommended for similar tasks. The learning tasks are described next.

Cross-modal Classification

Four cross-modal learning tasks are conducted: two tasks for the singing voices and the affective vocalizations, and two tasks for the musical bursts played with the clarinet and the violin. The purpose of these tasks is to determine to what extent a model of emotion recognition in the singing voice, recognizes with good accuracy the emotional expression in the affective vocalizations (and vice versa). Similarly, to what extent a model of emotion recognition in a musical sound played with a clarinet can recognize emotions in a musical sound played with a violin (and vice versa).

The details of the tasks follow.

1. Given a classifier trained on the singing voices (SVDB) denoted by C_{SV}:
   (a) Test the classifier on the affective vocalizations dataset, using the features (MAV_{fs}) that correspond to the optimal feature subset of the trained classifier C_{SV} SVDB dataset. We refer to this task as C_{SV} x MAV_{fs}
2. Compute the GE and determine the PA from the confusion matrix.
3. Repeat steps 1 and 3:
   (a) Given a classifier trained on the vocalizations (MAV) denoted by C_{MAV}:
      i. Test the classifier on the SVDB dataset using the features SV_{fs} that correspond to the optimal feature subset of C_{MAV}. The task is denoted as C_{MAV} X SV_{fs}. See Table 3.5a.

The results in table 3.5a show that the singing voice classifier fails to predict the valence of the affective vocalizations with good accuracy, however it achieves 80.83% PA on arousal. Conversely, the MAV classifier achieves a PA of 78.78% on valence and 82.33% on arousal. This shows that a learning model trained on the emotional vocalizations can predict with good accuracy the emotional expression in the singing voice. A similar task is made for the violin and the clarinet musical bursts. Table 3.5b shows that the SVM classifier trained on the violin

Table 3.5: Cross-modal prediction performance with SVM and NN

(a) SVDB and MAV

	SVM		NN	
Task	Valence	Arousal	Valence	Arousal
C_{SV} x MAV_{fs}	58.33%	80.83%	67.03%	74.80%
C_{MAV} x SV_{fs}	78.78%	82.33%	62.13%	74.23%

(b) Clarinet and Violin

	SVM		NN	
Task	Valence	Arousal	Valence	Arousal
C_V x CL_{fs}	68.33%	80.83%	66.67%	83.33%
C_{CL} x V_{fs}	80.00%	81.67%	60.00%	65.83%

dataset achieved 80.83% accuracy in classifying the clarinet dataset on the arousal dimension, while the NN classifier achieves 83.33% on the same dimension. On the valence dimension, the SVM classifier of the clarinet dataset performed better than the NN, with a performance of 80% in classifying the clarinet samples on the valence dimension.

Cross-Domain Classification

In order to estimate the degree of generalizability of the acoustic features between vocal and musical bursts, eight cross-domain tasks are realized. In [230] authors show that the "algorithms trained on emotional music are quite successful on emotional speech and vice versa".

The following approach is made for affective non-verbal voice and music:

1. Given a trained SVDB classifier C_{SV}:
   (a) Test the classifier on the feature subset of the violin musicals V_{fs} that corresponds to the optimal feature subset of C_{SV}. We refer to this task as C_{SV} x V_{fs}.
2. Compute the GE and determine the PA from the confusion matrix.
3. Repeat steps 1 to 3, testing:
   (a) C_{SV} is tested on the corresponding clarinet feature subset CL_{fs}: C_{SV} x CL_{fs}
   (b) A classifier trained on violin musicals C_V is tested on the corresponding feature subset SV_{fs}: C_V x SV_{fs}.
   (c) A classifier trained on clarinet musicals C_{CL} is tested on the corresponding SV_{fs}: C_{CL} x SV_{fs}
   (d) A classifier trained on the affective vocalizations C_{MAV} is tested on:
      i. The corresponding clarinet feature subset CL_{fs}: C_{MAV} x CL_{fs}

ii. The corresponding violin feature subset V_{fs}: C_{MAV} x V_{fs}

(e) A classifier trained on the violin musicals C_V is tested on the corresponding feature set in the vocalizations MAV_{fs}: C_V x MAV_{fs}.

(f) A classifier trained on the clarinet musicals C_{CL} is tested on the corresponding feature set in the vocalizations MAV_{fs}: C_{CL} x MAV_{fs}.

Table 3.6: Cross-domain prediction performance with SVM and NN

(a) SVDB versus musical bursts

	SVM		NN	
Task	Valence	Arousal	Valence	Arousal
C_{SV} x CL_{fs}	53.16%	64.16%	55.00%	57.50%
C_{SV} x V_{fs}	49.00%	67.50%	57.5%	56.67%
C_{CL} x SV_{fs}	59.00%	77.26%	45.43%	68.16%
C_V x SV_{fs}	48.48%	65.15%	69.70%	54.56%

(b) Vocalizations versus musical bursts

	SVM		NN	
Task	Valence	Arousal	Valence	Arousal
C_{MAV} x CL_{fs}	60.67%	71.67%	55.00%	69.16%
C_{MAV} x V_{fs}	63.33%	73.33%	55.00%	65.83%
C_{CL} x MAV_{fs}	64.43%	67.77%	51.50%	68.90%
C_V x MAV_{fs}	66.29%	76.66%	53.70%	68.53%

3.4 Results

3.4.1 Intra-domain

Table 3.1 shows that the classification performance with SVM achieves a higher performance when wrapper feature selection with SFS is implemented. The accuracy for (valence, arousal) is (78.33%, 91.67%) respectively, displayed in a blue colour in the table. The classification performance with NN and MI feature selection (thereafter referred to as the MINN model) achieves a remarkable accuracy of (95.26%, 97.63%), displayed in a red colour in the table. Furthermore, they highlight the strength of the MI feature selection method combined with NN classifier in achieving high prediction results. An important note to mention is that when a SVM classifier was implemented with a feature set obtained with MI feature selection, the results were lower than in table 3.1c. Additionally, when both feature selection methods were applied, MI first followed by SFS, and then the resulting feature set is input the a SVM, the results were also poorer than those reported in table 3.1c.

Table 3.2 surprisingly shows that the SFS feature selection reduces the classifier's performance. This indicates that consistently applying FS methods is not always an optimal practice for improving classification performance. This is reported in table 3.2a where the success rates are (93.33%, 100%) compared to (76.67%, 96.67%) with SFS feature selection. However, as can be seen in table 3.2c, an outstanding performance is achieved (100%, 100%), showing the power of the MINN model in recognising emotions in the clarinet corpus.

The highest results for the intra-domain task on the violin musical corpus are seen in tables 3.3b in blue as well as in table 3.3c in red. The MINN model outperforms the other models.

Similarly, the highest results on the singing voice corpus are obtained in tables 3.4b and 3.4c. These results clearly demonstrate the power of the MINN model over the other models in emotion recognition from singing voices.

Recommended model: In light of the above results, the MINN model using the mutual information feature selection with a feedforward neural network is recommended for intra-domain emotion recognition tasks, regardless of the type of the sound stimuli. Furthermore, the results show the strength of the proposed feature set in the recognition of emotion in vocal as well as musical expressions of affect.

3.4.2 Cross-modal

The results in table 3.5a show that the MINN model fails to outperform the SFSSVM model for the cross-modal task. The highest performance accuracy marked in a red colour, is of (78.78%, 82.33%) on (valence, arousal) when the classifier is trained on the vocalizations and tested on the singing voices. However when the classifier is trained on the singing voices and tested on the vocalizations, it fails to achieve a similar result on valence with a mere 58.33% accuracy, but achieves a fair accuracy of 80.83% on arousal. Several causes can be pointed out: first, the SVDB contains neutral productions of singing voices with a rather stable pitch, whereas the MAV consists of sounds that have a more complex dynamics and a richer information, and therefore a trained classifier on MAV can logically better predict simpler sounds such as the singing voices of SVDB. Second, the SVDB is very small in size, with only 22 sounds to train the classifier. Third, it is a known issue in the ER field, that arousal is generally better recognized in ER tasks applied to speech and music.

The results in table 3.5b show that a higher recognition rate is obtained with a SFSSVM model when the classifier is trained on the clarinet musical corpus and tested on the violin corpus. This confirms that for the current cross-modal tasks, the MINN model achieved a poorer performance than SFSSVM.

Recommended model: For cross-modal tasks, the SFSSVM model has a higher performance and is therefore recommended.

3.4.3 Cross-domain

The results of table 3.6a show a similar performance of the two models SFSSVM and MINN. Both models fail to achieve good accuracies in a cross-domain emotion recognition across voice and music. In general the results achieve less than 60% accuracy, which indicates a poor emotion recognition rate across corpora. Four results are superior to 60% but inferior to 70% and they are obtained for arousal. Also one result on arousal is of 77.26%.

The results of table 3.6b show low emotion recognition rates of both models in the cross-domain learning task across vocalizations and musical bursts.

Recommended model: No model is recommended for the cross-domain emotion recognition tasks. Both have failed to achieve high accuracy rate. The low accuracy rates compared with the high rates in intra-domain classification tasks, suggest that the models used are not holistic, because they fail to recognize emotion across corpora.

3.4.4 Feature subsets

The vocal features that achieve the highest PA are reported in this section. Table 3.7 lists features per database on each affect dimension. Table 3.8 reports the features that are common to the vocal and musical databases.

Table 3.7: Predictive features per dataset for Valence (v) and Arousal (a)

	Features	SVDB	MAV	CLARINET	VIOLIN
Formants	F1		v a	v	a
	F2		a		v a
	F3		a		a
	F4	v	a		a
	F5			a	a
Formant Bandwidth	B_{F1}		v		a
	B_{F2}		v	va	
	B_{F3}		va		va
	B_{F4}		v	va	a
	B_{F5}				v
Voice Perturbation Measures	Jit	a	va		va
	Shim		a	a	va
	MHNR	v	va	a	va
	MAC	v	a		va
Temporal Features	RMS	va	va		a
	ZCR	va		va	va
Voice Quality	SPR	a	v		va
	BR				a
	MI	va	va		va
	MP	v	va	v	a
	MR	v	va		a
LTAS Voice Quality	Mean	va	a	va	va
	Slope	a	v		va
	LPH	a		a	a
	LRS			va	v
	SPL	va	a		a
Spectral Features	SC		va	a	va
	SS			v	a
	SE	a	va	v	va
	SF		v		va
Mel Frequency Cepstral Coefficients	MFCC1		v	va	va
	MFCC2	v	v		va
	MFCC3	v	va		v
	MFCC4	a	va	v	a
	MFCC5	v	va	va	va
	MFCC6	v	va		va
	MFCC7	a	va	v	va
	MFCC8	va	v	v	a
	MFCC9	a	va	va	va
	MFCC10	va	va	va	va
	MFCC11	va	va		va
	MFCC12	va	a	va	
	MFCC13	va	va	va	va

Table 3.8: Common features among databases for Valence (υ) and Arousal (a)

	Features	MAV ∩ CLARINET	SVDB ∩ CLARINET	MAV ∩ VIOLIN	SVDB ∩ VIOLIN
Formants	F1	υ		a	
	F2			a	
	F3			a	
Bandwidths	B_{F2}	υ			
	B_{F3}			υa	
	B_{F4}	υ			
Voice Perturbation Measures	Jit			υa	a
	Shim	a		a	
	MHNR	a		υa	υ
	MAC			a	
Temporal	RMS			a	a
	ZCR		υa		a
Voice Quality	SPR			υ	a
	MI			υa	a
	MP	υ	υ	a	
	MR			a	
LTAS Voice Quality	Mean	a	υa	a	υa
	Slope			υ	a
	SPL			a	a
Spectral Features	SC	a		υa	
	SE	υ		υa	a
	SF			υ	
Mel Frequency Cepstral Coefficients	MFCC1	υ		υ	
	MFCC2			υ	υ
	MFCC3			υ	υ
	MFCC4	υ		a	a
	MFCC5	υa	υ	υa	υ
	MFCC6			υa	a
	MFCC7	a		υa	a
	MFCC8	υ	υ		a
	MFCC9	υa	a	υa	a
	MFCC10	υa	υa	υa	υa
	MFCC11	υ	a	υa	υa
	MFCC12	a	a	a	υa
	MFCC13	υa	υa	υa	υa

3.5 Discussion

In general, the MINN model is recommended for intra-domain emotion recognition tasks. The SFSSVM model is recommended for the cross-model recognition tasks. Cross-domain recognition results are noticeably lower than both intra-domain and cross-modal tasks, for both models.

These results have several implications:

1. First, the proposed feature set has proved highly efficient in recognizing emotions in the four corpora of sounds when the MINN model is used, and the classifier is trained and tested on the same corpus. In this sense their generalizability is established. However, when used for cross-corpora tasks, their generalizability.

2. Second, even the best performing ER model for intra-domain classification, is not holistic for ER across voice and music. In fact, both models are not holistic. This shows that the present results do not corroborate the notion of the common origins of music and voice. This finding is of significant importance, as it shows that although there is some kind of overlap in the acoustic features, music and voice seem to communicate emotions differently, which favours our intuition that music may not have entirely originated in voice. However, it is important to note that further work is needed to resolve this debate. For example, future tests would repeat the present experiments with acoustic features specific to the music domain, and would test how well these features capture emotions in the voice domain. Additionally, it would be interesting to develop vocal and musical stimuli that would be specifically designed to imitate each other on various scales of the sound's structure.

3. Third, there is a significant dependence of the results on the learning algorithm as well as on the feature selection method. This highlights the necessity to explore different approaches that don't rely on the acoustic measures but rather explore how the inherent dynamical nature of sounds communicates emotions over time.

3.6 Conclusion

In conclusion, using the approaches of this work, we found that it is not possible to build a holistic model of emotion recognition across voice and music. Some acoustic features are shared among voice and music, and they generalize across corpora only in intra-domain ER tasks. The findings confirm that although "Spencer's law should be part of any satisfactory account of music's expressiveness" [94], it "cannot be the whole story of music's expressiveness" and that some emotional cues in the music are contributed by other characteristics than what might be common with voice, such as the musician's performance. Additionally, further work is needed to identify features that enhance the recognition of valence. This is addressed in chapter 6. Since the SVDB dataset is small with only 22 samples, a future work will expand the dataset size to over 400 samples and the tests will be repeated in order to gain a better view of the

generalizability of the results. In conclusion, the debate that attributes the origins of musical emotions to primitive vocalizations must be nuanced as our results do not provide irrefutable evidence to this claim.

3.7 Dissemination

A major part of this work is pending submission, and the first part appears in:

1. Pauline Mouawad, Myriam Desainte-Catherine and Jean-Luc Rouas. Multilabel Classification of Nonverbal Communication of Emotions. In proceedings of the 21st International Symposium on Electronic Art, 8th International Workshop on Machine Learning and Music, MML2015.

Chapter 4

Chaos Theory and Nonlinear Dynamics State of the Art II

> You've never heard of Chaos theory? Non-linear equations? Strange attractors? Ms. Sattler, I refuse to believe you're not familiar with the concept of attraction.
>
> Michael Crichton

When in the early 1980's, concepts and theories of chaotic dynamics were applied to complex music signals, it became clear that a new perspective was needed to understand the mechanisms involved in sound production [145], a perspective that goes beyond linear analysis with low-level acoustic features. About a decade later, various studies started to show that nonlinear phenomena exist in sounds produced by musical instruments [67, 131], in voice [77], such as in infant cries [138], mammals vocalizations [232], bird songs [59], pathological voices [4, 74, 78] and emotional speech [75, 76, 167].

Cartwright et al. 2001 showed that dynamical attractors carry perceptual meaning in the case of pitch perception. Furthermore, a given sound reaches our brain through our auditory system which is a highly complex nonlinear dynamical system, and is processed in the nervous system, to which dynamical attractors are central [23].

In this part, we show that nonlinear phenomena in vocalizations as well as in music have an important role in carrying perceptible affective information to the listener, and that they are powerful measures for affect recognition tasks in various sounds. We review notions from chaos theory, which is intertwined with methods from nonlinear time series analysis. We discuss recurrence plots, recurrence quantification analysis, the concept of embedding and phase space reconstruction. Furthermore we review methods from symbolic time series analysis and discuss their relevance to the present work. Then we establish the grounds for our proposed approach that will be elaborated in chapter 5.

4.1 Chaos Theory

In Chaos Theory, the term 'chaos' describes the behaviour of nonlinear deterministic systems, that 'seems' random [88]. A core concept to such systems is known as an extreme sensitivity to initial conditions or the *butterfly effect*, a term coined by Edward Lorenz [124] to describe how small changes in the initial conditions of a chaotic dynamics system can cause significant changes in that system's outcome. Just prior to delving into the more technical aspects of the applications of the theory to our work, we indulge the interest of the reader with the anecdote behind the "butterfly effect".

The Butterfly effect One meteorologist commenting on Lorenz's 1963 seminal paper *Deterministic Nonperiodic Flow* [124], made the following statement:

"If the theory were correct, one flap of a sea gull's wings would be enough to alter the course of the weather forever".

Lorenz upheld:

"The controversy has not yet been settled, but the most recent evidence seems to favor the sea gulls" [125]

In later speeches, Lorenz would prefer to use the poetic word "butterfly". Until in 1972 during the 139^{th} meeting of the American Association for the Advancement of Science, he was about to present a talk to which he could not find a title, Philip Merilees designated his talk as[1]:

"*Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?*"

This title was subsequently widely referred to as the butterfly effect, where the flapping of the wings represents a small change in the initial conditions of a system, which alters the sequence of events that will significantly modify the system's outcome[2].

Deterministic and Nonlinear Systems In the popular usage, we often describe some state of things as chaotic when we do not distinguish a regularity or order. However chaos theory is concerned with dynamic systems that are deterministic in the sense that their evolution can theoretically be predicted if their initial states are precisely known. This means that the exact knowledge about the future state of a dynamic system sensitively depends on how accurate is our knowledge about its initial states, and changes dramatically if minor changes occur to the initial conditions. This property makes deterministic systems unpredictable, because the initial conditions are generally unknown, so that the prediction based on some knowledge of the initial condition and the real state of the system based on the real initial condition,

[1] http://www.telegraph.co.uk/news/obituaries/1895916/Professor-Edward-Lorenz.html

[2] It is remarkable to note that nineteen years earlier, in his fiction 'The Sound of Thunder', American author Ray Bradbury tells how the death of a butterfly 65 million years in the past, changes the presidential elections in year 2055 of our future.

diverge exponentially. Eventually the system's real trajectory will be entirely different from the predicted one, in the long run.

Additionally, when a system reveals dynamical chaos then it must be nonlinear, because nonlinearity coexists with chaotic dynamics. Nonlinearity means that measured values of the system's properties in a future state depend in a nonlinear way on the measured values in an earlier state [161].

A simple example of a nonlinear system is:

$$x_{n+1} = x_n^2 \tag{4.1}$$

The value of x in the $(n+1)^{th}$ state depends on the square of x in the n^{th} state. This is referred to as a nonlinear mapping of the n^{th} state to the $(n+1)^{th}$ state [161].

Another physical example is the temperature of water when it is being brought to a boiling point. When the boiling point is reached, "the temperature in the $(n+1)^{th}$ state is just equal to the temperature in the n^{th} state", however while the water is being heated, this is obviously not true. Hence the system evolves progressively in a fixed and known deterministic way, however this knowledge does not allow us to predict farther future states of the system [88]. Other examples of deterministic systems are found in fluid dynamics, chemistry as well as in the brain, the heart, in meteorology, and the solar system [14].

The study of such dynamic, deterministic, nonlinear systems is referred to as Nonlinear Dynamics Analysis (NLDA), and one of the most promising applications is within the field of nonlinear time series analysis (NLTSA). As pointed out by Kantz and Schreiber: "The most direct link between chaos theory and the real world is the analysis of time series from real systems in terms of nonlinear dynamics" [97]. Section 4.2 describes how nonlinear time series analysis can be efficient and crucial if we want to identify and predict dynamical systems.

Non-deterministic systems When the system under study is not clearly deterministic, the suitability of chaos theory is controversial [185]. However, serious efforts have been invested to apply methods from chaos theory to data observed from systems that are not evidently deterministic, but that cannot be described with traditional or linear methods. Such systems are typical in various areas such as medicine, astrophysics or social sciences.

4.2 Nonlinear Time Series Analysis

Nonlinear time series analysis (NLTSA) consists of a set of methods that characterize dynamical information from time-ordered values in a dataset. It is based on the fact that the real underlying dynamical state of a complex system is often unknown, and that all the information needed to determine the future behaviour of the system's state is independent of its past, and can be predicted based on knowledge of the present state, quantified by the time series.

A first sensational application of NLTSA was the prediction of the path of a ball on a roulette wheel by J. Doyne Farmer and Norman Packard at the University of California of Santa Cruz [148]. This experiment showed the power of NLTSA methods, and since then they

have been applied across all branches of science and engineering, as well as social sciences, the humanities and beyond [19].

A main assumption of time series is *stationarity*, which means that the statistical properties of the series like mean, variance and autocorrelation are constant and do not change over time. If the stationarity is strong, then the time series is *ergodic*, which means that a long sample of the series represents the entire process [198]. This ergodicity is an important characteristic of stationary time series, because it implies that statistical techniques can be directly applied to infer meaningful information from the series.

A time series consists of a few observed scalar measurements computed from a large sample of data, that represent a small observation of the underlying dynamical system. When analysing the time series, we have incomplete information about the original system that generated the data, therefore in order to recover the missing information and derive knowledge about the underlying dynamical system from the observed data, a framework is applied that extracts relevant physical information from time delayed copies of the available signal [185].

Phase Space Reconstruction The states of dynamical systems change in time, and their time evolution is defined geometrically in the shape of trajectories that belong to a *phase space* termed the *attractor*. The critical sensitivity of the dynamical system's states to the initial conditions endows them with a *strange* behaviour in time, which has granted their shape in phase space the term *strange attractor*.

One of the most important tasks in the study of dynamical systems is to predict how their states change in time. Some examples that highlight the importance of prediction include forecasting weather, earthquakes or epileptic seizures [134].
The phase space consists of points that represent the possible states of the system. Such that if we know the present state of a deterministic system, then we can determine the states at all future times as well. So the construction of a vector space, also called *state space* or *phase space* for the dynamical system is necessary, as it will allow to determine the state of the system by specifying a point in the space. In turn it becomes possible to study the dynamics of the system by studying the dynamics of those phase space points [97].
In practice, we do not have a full knowledge of the dynamical system in order to reconstruct its phase space. But we do have a time-discrete measurement of one observable, which results in a scalar and discrete time series, that is used to reconstruct the original system's dynamics, through the reconstruction of its phase space via embedding. The embedding theorems guarantee that for noise-free data, there is a dimension m such that the embedded vectors are equivalent to the original phase space vectors [97].
To reconstruct the phase space of a system from a time series, the Takens' embedding theorem is used [208] and the framework is the following [185]:

Let $x(t)$ be a trajectory of a dynamical system and $s(t) = s(x(t))$ the result of a scalar measurement on it. Then a delay reconstruction with time delay τ and embedding dimension m is given by:

$$s(t) = (s(t - (m-1)\tau), s(t - (m-2)\tau), ..., s(t)) \tag{4.2}$$

Embedding parameters One of the main challenges of the delay-coordinate embedding theorem, is choosing appropriate values of dimension m and time lag τ. In fact, the meaningful parameter is the product $m\tau$, rather than m or τ alone, because $m\tau$ represents the time interval described by an embedding vector [97].

Several methods exist but we are naming the most widely used ones in literature. First τ should be estimated: if τ is very small, consecutive elements of the delay vectors will highly correspond, and all the vectors will be clustered around the main diagonal, unless m is very large. if τ is very large, consecutive elements are independent, and the points will fill a large space in the phase space [97]. The first zero of the autocorrelation function of the time series returns the smallest τ that maximizes the linear independence of the coordinates of the embedding vector. The first minimum of the mutual information function of the time series, will be a suitable value for τ.

Once τ is chosen, the next step is to estimate the embedding dimension m. If m is too large, the embedded data will be redundant, which will confuse the performance of prediction algorithms.

Here also, two widely known methods can be used: the first one is the false-nearest neighbour algorithm (FNN) [99] and the 'asymptotic invariant approach'. The FNN method is used in this work since it is the most widely used one and the embedding dimension m is chosen where the number of false neighbours drops to zero. The attractor that results from the embedding is equivalent to the attractor in the original phase space if $m \leq 2d + 1$, were d is the dimension of the original phase space. In general we don't know the value of d, but using the FNN method the embedding dimension is guaranteed to fulfil that requirement.

Recently, NLTSA was used to analyse the nonlinear dynamics of audio signals, such as speech. Only few studies apply nonlinear time series analysis to music [66, 164]. Section 4.6 reviews some of the related works.

Limitations of embedding Generally NLTSA involves a delay-coordinate embedding of the time series, and if the embedding is accurately performed, then properties of the original dynamical system are preserved in the embedded space. However one important limitation is that accurate embeddings are not easy to construct [86], and for different values of the embedding parameters m and τ the quantification measures that describe the underlying dynamics of the system may vary as well.

It was suggested in [86] that with methods of recurrence plots analysis it may not be necessary to embed the data, and that NLTSA methods that circumvent the need for embedding are thus utterly required. Furthermore March et al. suggest that given an unembedded recurrence plot, it is possible to derive from it the statistical inferences of all the embedded recurrence plots [133].

Therefore it is interesting to construct recurrence plots with a NLTSA that does not require embedding, and subsequently derive the quantification measures that would help us understand and predict the states of the system. In the next section the method of recurrence plot analysis

is explained, and later in the chapter, we explain how we address those limitations with our method.

4.3 Recurrence Plots

Recurrence is a fundamental property of most dynamical systems, it is due to the systems' recurrence to former states, that we know how to make predictions about the future state of the system. Recurrence takes place in the system's phase space, and the tool that measures a recurrence of a trajectory in phase space is called a recurrence plot (RP) [134].

Given a trajectory $\vec{x}_i \in \mathbb{R}^d$ in a d-dimensional phase space of a dynamical system, the RP is a two-dimensional visualization of the square recurrence matrix of the embedded time series defined by:

$$R_{i,j}^{m,\varepsilon} = \Theta(\varepsilon - \left\|\vec{x}_i - \vec{x}_j\right\|), \quad i,j = 1,...,N \tag{4.3}$$

where $\vec{x}_i$ and $\vec{x}_j$ are phase space trajectories in an m-dimensional phase space, N is the number of measured points in a trajectory, ε is a threshold distance, $\Theta(.)$ the Heaviside function such that: $\Theta(x) = 0$, if $x < 0$ and $\Theta(x) = 1$ otherwise, and $\|.\|$ is some appropriate choice of a norm, such as the L_2-norm, otherwise known as the Euclidean distance. Both axes of the RP are time axes. The dots or pixels located at (i,j) and (j,i) on the RP are black if the distance between points x_i and x_j in the phase space fall inside a *ball* or *threshold corridor* of radius ε, the threshold distance [19, 168]. In this case, the black points refer to *recurring* states also termed ε-recurrent states since they occur in an ε-neighbourhood. The ε-recurrent states are represented by the relation [134]:

$$\vec{x}_i \approx \vec{x}_j \iff R_{i,j} \equiv 1. \tag{4.4}$$

The dots are white if $R_{i,j} \equiv 0$. The RP always displays a main black diagonal line called the line of identity (LOI), since $R_{i,i} \equiv 1$ by definition. For more in-depth description of the RP properties, the reader is referred to [134].

4.3.1 Factors to consider for RP construction

There are some factors to consider when constructing an RP by means of time-delay embedding. These include the choice of the threshold ε as well as the choice of the embedding parameters m and τ.

Choice of threshold ε So far it is clear that identifying recurrences in dynamical systems is a crucial step in NLDA , and that determining the recurrences rely on a key parameter, the threshold ε. Therefore selecting an appropriate value for ε is required before analysing a dynamical system. If ε is too small, there will be almost no recurrence points and little or no trajectories will fall into the same neighbourhood, and then we cannot learn much about the recurrence structure of the dynamical system. If ε is too large, then a large number

of trajectories will fall inside the same neighbourhood, and these may include points that are consecutive points on a given trajectory. This is known as the *tangential motion* and as a result the RP will display longer and thicker diagonal structures, that are not a realistic representation of the actual recurrences. An additional problem is added due to the influence of noise in the observed signal, that distorts the structure in the RP. Therefore in order to preserve that structure ε should be large. Ultimately the choice ε depends on the system under study. In our work we have circumvented the requirement of choosing ε for building the RP, so should the reader wish to gain a better understanding of the appropriate choices for ε, a thorough analysis is given in [134].

4.3.2 Qualitative Description of RP Structures

For the sake of consistency, the following description is taken from [134].

RPs display large-scale patterns as well as small-scale textures. Depending on the patterns and textures in the RP, a different information can be derived.

For example, RPs with large-scale patterns can be:

1. Homogeneous: stationary systems show homogeneous RPs "if all the characteristic times are short compared to the length of the time series", and "the overall pattern of the RP is uniformly grey although at small scale nontrivial texture may be visible" [45].

2. Periodic: periodic and quasi-periodic systems show periodic or quasi-periodic diagonal lines and checkerboard structures.

3. Drift: non-stationary systems show drift RPs with structures that move away from the LOI.

4. Disrupted: the RPs show white bands caused by abrupt changes in the systems' dynamics. Computing the frequency of the recurrences of these white areas helps in assessing extreme and rare events in the dynamics.

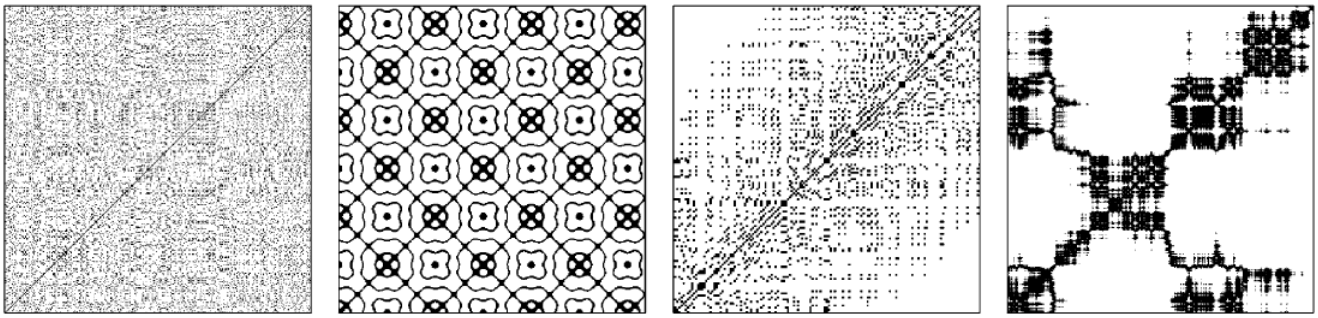

Figure 4.1: Characteristic Structures of recurrence plots: (A) Homogeneous, (B) Periodic, (C) Drift, (D) Disrupted [134]

Small-scale structures constitute the RP texture. They consist of single dots, diagonal, vertical and horizontal lines. Among these, diagonal and vertical lines are the most important ones, as they form the basis to derive quantitative measures that will be subsequently used in prediction tasks.

1. Single and isolate recurrence points occur if states are rare, persist for a very short time or vary considerably.

2. Diagonal lines exist when a segment of a trajectory runs in parallel to another segment within an ε-tube.

3. Vertical as well as horizontal lines indicate a lapse of time during which a state changes very slowly or does not vary at all.

4.3.3 Recurrence Quantification Analysis

In order to derive meaning from the structures of the RP, various complexity measures are computed from the RP, that quantify those structures. Such quantification is important since it will be employed to characterize various information and to perform predictions. These statistical measures are known as Recurrence Quantification Analysis (RQA), and are based on the density of recurrence points, the diagonal and the vertical line structures in the RP.

Measures based on the density of recurrence points

Given an RP thresholded at ε (Eq. 4.3), the Recurrence Rate (RR) measures the density of recurrence points in the RP:

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j} \tag{4.5}$$

The RR measure corresponds to the correlation sum (D2) measure, but D2 excludes the main diagonal line (LOI):

$$D2 = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ j \neq i}}^{N} R_{i,j} \tag{4.6}$$

Diagonal lines based measures

Given the histogram $P(l)$ of diagonal lines of length l, the following measures are computed:

Determinism (DET) is the percentage of points in diagonal line of at least length $l = l_{min}$, i.e. the ratio of recurrence points in the diagonals to all recurrence points, and is a measure of the predictability of the system. Processes with chaotic behaviour cause none or very short diagonals. Deterministic processes cause longer diagonals and less isolated recurrence points.

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)} \tag{4.7}$$

The average length of diagonal line length L is the average time during which two segments of a trajectory are close to each other, and it refers to the mean prediction time. The length l of diagonal lines refer to the number l of time steps during which a segment of the trajectory is close to another segment of the trajectory at a different time. Therefore the diagonal lines are related to the divergence of the trajectory segments.

$$L = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=l_{min}}^{N} P(l)} \tag{4.8}$$

Then the length L_{max} of the longest diagonal line in the RP excluding LOI is derived:

$$L_{max} = max(\{l_i\}_{i=1}^{N_l}) \tag{4.9}$$

And the inverse of L_{max} indicates the divergence (DIV) of the phase space trajectory. The faster the trajectory segments diverge, the diagonal lines will be shorter, and the value of DIV will be higher:

$$DIV = \frac{1}{L_{max}} \tag{4.10}$$

The next measure is the Shannon entropy of diagonal line length distribution in the RP (S_{RP}), which is the probability $p(l) = P(l)/N_l$ to find a diagonal line of exactly length l in the RP. It is a measure of complexity in the RP in terms of the diagonal lines, such that, for uncorrelated noise the value of S_{RP} will be small, which indicates a low complexity.

It is defined as:

$$S_{RP} = -\sum_{l=l_{min}}^{N} p(l)\ ln\ p(l) \tag{4.11}$$

The $RATIO$ is a measure that uncovers transitions in the system's dynamics:

$$RATIO = \frac{DET}{RR} \tag{4.12}$$

Vertical lines based measures

Measures based on vertical structures in the RP uncover chaos-chaos transitions [135] in a dynamical system that are not found using diagonal line based measures. These are *laminarity* and *trapping time*.

The laminarity (LAM) refers to the occurrence of laminar states in the system independently of their lengths. If the RP contains less vertical lines and more single recurrence points, then the value of LAM will be low. Its definition is analogous to the definition of DET for vertical lines of minimal length $v = v_{min}$.

$$LAM = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)} \tag{4.13}$$

The trapping time measure (TT) is the average length of vertical lines, and estimates the mean time that the system's state will be trapped:

$$TT = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)} \tag{4.14}$$

4.3.4 Limitations

A major concern with RQA estimates from RPs is that they are not invariants, since they heavily depend on the embedding performed to reconstruct the phase space. If the embedding parameters m and τ vary, we obtain different RQA estimates, which in turn impacts the understanding we derive about the system's dynamics and by the same token, the predictions we make. Therefore one viable approach is to derive RP as well as RQA without embedding, since as mentioned in paragraph 4.2 embedding might not be necessary with methods of recurrence plots.

Another issue to address is the case of discrete-valued time series. Most of the time series analysis methods have been developed for continuous-value time series only, and when they are applied to discrete-valued observables problems are encountered in interpreting the RQA results. Additional problems are also encountered if the variability of a system happens at very different time scales [37].

These issues can be addressed using symbolic time series analysis that encodes a time series into a sequence of discrete symbols, to which statistical analysis is subsequently applied in order to characterize the dynamics of the system. The advantage of the symbolic discretization of the time series is that it prunes irrelevant information by filtering only the interesting aspects of the system's dynamics. The next section describes this method.

4.4 Symbolic Time Series Analysis

Symbolic time series analysis (STSA) also known as *data symbolization*, are a collection of methods from symbolic dynamics, a branch of dynamical systems. The main objective of STSA is to recognize patterns in complex dynamical systems. In order to find such patterns, STSA methods transform time series measurements into a sequence of discretized symbols, i.e. a finite string, (an operation also known as *discretization*), that retain the essential temporal information of the system. Then the sequence of symbols is analysed by looking for regularities to describe the dynamics of the system. A main strength of STSA is that no assumptions are made with respect to the structure of the underlying dynamical system, therefore it applies to deterministic or stochastic, linear or non-linear systems [21, 36].

A main STSA method is the *static transformation* of the time series through generating partitions. Such method partitions the phase space into a finite number of regions, then a

symbol is assigned to each region. After symbolization, temporal patterns present in those symbol sequences can be identified, in addition to a global characterization of such sequences in terms of entropy and other recurrence statistics.

Although the application of generating partitions is effective for estimating dynamical invariants from the time series such as correlation functions, mutual information, permutation entropy or transfer entropy, the resulting invariants depend strongly on the distribution of the symbols [37]. Furthermore in the presence of noise in the signal, estimating the generating partitions becomes very difficult.

In the next section, we describe a recently developed method of time series symbolization [224] that finds the best symbolic representation of the time series in terms of the symbolized sequence recurrence properties, instead of the individual time series. It does not rely on partitioning the phase space, and circumvents the need for determining the phase space embedding. Another advantage of this method is that it allows the construction of symbolic recurrence plots that depict only the essential information retained from the data measurement.

4.5 Variable Markov Oracle

The Variable Markov Oracle (VMO) [224] is a suffix tree data structure that is derived from Factor Oracle (FO) [3, 6] as well as Audio Oracle (AO) [42].

FO is a suffix automaton that finds factors (*repeated substrings*) in a word (or sequence of symbols), as well as patterns (*repeated suffixes*) [3]. It has been employed mainly for optimal string matching algorithms, such as biosequence pattern matching. Assayag et al. 2004 showed how the FO can be adapted to learn symbolic musical sequences and generate symbolic musical improvisations in real-time [6].

AO is an extension of FO for audio signals, that is independent of the audio feature representation. AO extends the applications of FO to multivariate time series such as an audio signal sampled at discrete times. Based on a distance measure, the AO structure finds and links all the possible combinations of audio sub-clips that are similar. AO has been successfully applied to audio generation.

Since VMO is an extension of both FO and AO, a brief description of the original algorithms is given, and then the VMO construction algorithm is explained [222, 224].

Factor Oracle Given a sequence of symbols $\mathcal{S} = \sigma_1, \sigma_2, ..., \sigma_N$, an oracle structure is constructed with N states, where every symbol σ_i is identified with a state. Typically a FO has two types of links (figure 4.2).

1. A forward link denoted by $\delta(i-1, \sigma_i) = i$ starts at the beginning of $\mathcal{S}$ and constitutes a unique path to any of the substrings of $\mathcal{S}$, such that by following the paths, we are able to retrieve all the substrings of $\mathcal{S}$. Forward links can be internal or external:

   (a) An internal forward link is a pointer from state $i-1$ to state i, and is labeled by the symbol σ_i.

(b) An external forward link denoted by $\delta(i, \sigma_{i+k}) = i + k$ is a pointer from state i to state $i + k$, and is labeled by σ_{i+k} with $k > 1$. It is created when:

$$\sigma_{i+1} \neq \sigma_{i+k}$$
$$\sigma_i = \sigma_{i+k-1}$$
$$\delta(i, \sigma_{i+k}) = \emptyset$$

2. A suffix link is a backward pointer from state i to state k given that $i > k$. It locates repeated patterns in $\mathcal{S}$, due to an essential characteristic: a suffix link goes from i to k iff the longest repeated suffix of $\mathcal{S}$ is located at k, and is denoted by $sfx[i] = k$.

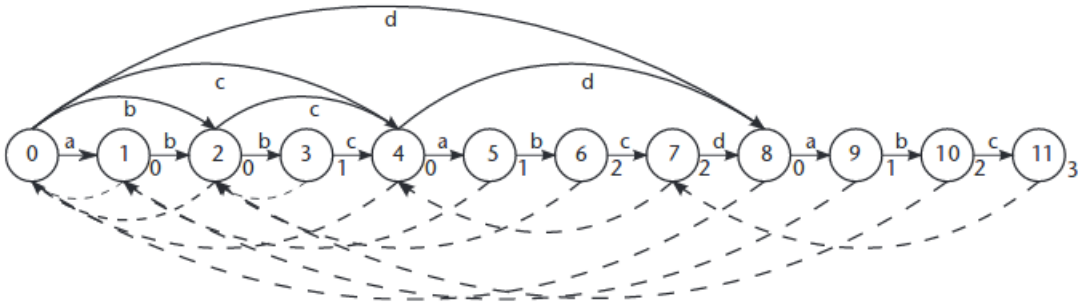

Figure 4.2: Factor Oracle for string $\mathcal{S}$ ="abbcabcdabc". A symbol is attributed to each forward link [222]

Audio Oracle AO extends FO for audio signals. Given an audio signal as input, AO reconstructs the signal into a sequence of feature vectors, and analyses them to determine similar subsequences according to a similarity threshold θ. The output is an oracle that contains pointers to similar subsequences that occur in different locations of the original sequence. The suffix structure of AO looks like FO however with no symbols assigned to the states. Given a continuous-valued time series $O[n]$, two samples $O[i]$ and $O[j]$ are similar if $|O[i] - O[j]| \leq \theta$. An example of an AO for a *tire skids* sound is in figure 4.3.

4.5.1 VMO Construction

VMO inherits the strengths of both FO and AO. The important improvement over its predecessors, is that VMO assigns symbols to the signal frames connected by suffix links during AO construction: it accepts a signal O as input, outputs the oracle structure, and keeps track of the sequence of assigned labels $Q = q_1, ..., q_N$ as well as a list of pointers to their corresponding observations $O = O[1], ..., O[N]$. As such VMO performs a symbolization of a signal's time

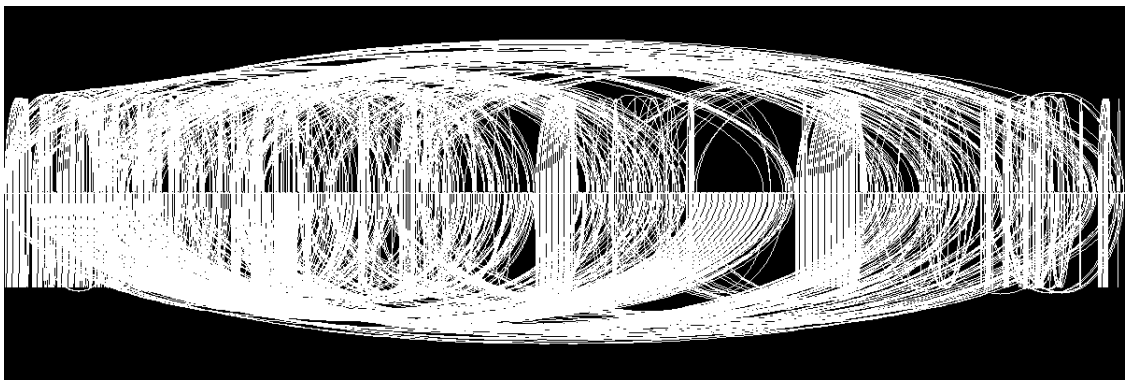

Figure 4.3: Audio Oracle of a tire skids sound [42]

series by storing the information regarding the repeated substrings via the suffix links created during AO construction and upgrades AO by assigning labels to the frames connected by suffix links.

The notations of the forward and suffix links remain the same as in FO construction. The detailed algorithm is found in [222, 224].

As mentioned earlier, a similarity threshold θ is introduced to determine if a signal sample $O[i]$ is similar to another sample $O[j]$.

In order to find the best symbolization of the signal, different VMO models can be created with different θ values. There is a tradeoff to consider when choosing θ values. If θ is very low, every frame will be different than every other frame, and VMO assigns a different symbol to each frame in O. If θ is very high, frames that are different are considered similar, and the same symbol is assigned to every frame in O. In both cases no structure in the time series can be captured by VMO.

Hence θ should be determined before VMO construction. Dubnov et al. have shown that the value of θ can be resolved by computing the Information Rate (IR) over candidate θ values [43]. The optimal θ value is the one that yields a highest IR value.

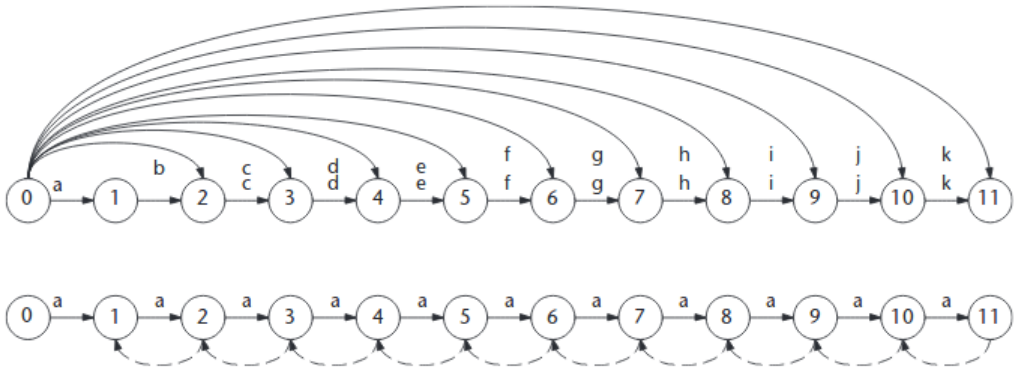

Figure 4.4: Two oracle structures. The top oracle has a very low θ value. The bottom oracle has a very high θ value [222]

Information Rate IR is an information theoretic metric that measures the information content of a time series.

Let $x_1^N = x_1, x_2, ..., x_N$ be a time series x with N observations, where $H(x) = -\sum P(x) log_2 P(x)$ is the entropy of x, then the definition of IR is [223]:

$$IR(x_1^{n-1}, x_n) = H(x_n) - H(x_n|x_1^{n-1}) \tag{4.15}$$

And it is approximated by replacing the entropy terms in equation 6.3 by a complexity measure C associated with a compression algorithm [223]. The complexity measure is the number of bits used to compress x_n independently using the past observations x_1^{n-1}:

$$IR(x_1^{n-1}, x_n) \approx C(x_n) - C(x_n|x_1^{n-1}) \tag{4.16}$$

Compror is a lossless compression algorithm based on FO and the length of the longest repeated suffix link (**lrs**). Details on *Compror* as well as on the method of combining Compror with AO and IR are found in [116]

IR is the mutual information between past and present observation in a signal $O[t]$ and is maximized when there is balance between variation and repetition in the symbolized signal, which means that a VMO with a higher IR value captures more of the repeating patterns than a VMO with a lower IR value [223]. Complete details about this can be found in [223].

4.5.2 Symbolic Recurrence Plots

From the generated VMO-symbolized time series, we obtain the symbolic RP, plotted from the binary self-similarity matrix.

In our work, we redefine the RP as a symbolic recurrence plot RP_S, obtained through the binary self-similarity matrix of the optimal VMO-symbolization of the signal's time series:

$$R_{i,j}^{\sigma_M,\theta} = \Theta(\theta - d(\sigma_{q_i}, \sigma_{q_j})) \quad i,j = 1, ..., N \tag{4.17}$$

Such that:

$$R_{i,j}^{\sigma_M,\theta} = \begin{cases} 1 & if\, d(\sigma_{q_i}, \sigma_{q_j})\, is \leq \theta \\ \\ 0 & otherwise \end{cases} \tag{4.18}$$

Where N is the number of states considered, σ_M refers to the M^{th} symbolized substring, Θ is the Heaviside step function (i.e. $\Theta(x) = 0$ if $x < 0$, and $\Theta(x) = 1$ otherwise). θ is a threshold distance, and $d(\sigma_{q_i}, \sigma_{q_j})$ is a distance metric between pairs of symbolized substrings q_i at $t = i$ and q_j at $t = j$.

4.5.3 Symbolic RQA

RQA characterizes the short-term dynamics of a system. So in order to study long-term features, some kind of filtering of the data is necessary. In such cases, there exists a relationship between the analysis of symbolic time series and RQA [37].

In our work we combine the advantages of RPs and RQAs with the symbolization using VMO, to derive symbolic complexity measures.

4.6 Applications of Nonlinear Dynamics

In [229], authors argue that RQA has become a powerful general purpose nonlinear methodology that is applicable to wide range of domains. Some of the areas of RQA applications are reported next.

4.6.1 Emotion Recognition in Voice

In their investigation about the potentials of NLD in discriminating emotions from speech, Lombardi et al. 2016 use recurrence plots as well as recurrence quantification analysis to explore the recurrence properties of vowel segments extracted from a set of spoken sentences [123]. They evaluate their performance for six categories of basic emotions: anger, boredom, fear, happiness, neutral and sadness. In the qualitative analysis of the RP results, they find that some RQA measures are most distinctive for some emotions, and that the density of points in the RPs, the length of the lines as well as the measures that differentiate the various time periodicities, can effectively discriminate between different emotional levels. They employ statistical tests in order to determine whether the six different groups of data of each RQA measure originate from the same distribution. Also, they compute the mean, standard deviation, median and interquartile range values of the nine RQA measures for all the groups of emotions. They find that some RQA measures are better associated with some basic emotions, and that all of them are statistically significant for discriminating the six groups of emotions, with the exception of $RATIO$, LAM, TT and V_{max} for neutral-boredom, and L for fear-happiness. It would have been interesting to see how their measures perform in an automatic classification task, to determine to what extent they are successful in the automatic recognition of emotions from the speech vowels, and to what extent they would generalize across different datasets of vowels. Furthermore, knowing that the RQA measures depend heavily on the embedding parameters and their method relies on the phase space embedding [208], it is unclear how their measures would perform if the embedding parameters were to change.

In [167], nonlinear dynamics measures are used for the recognition of three categorical emotions in speech, namely: neutral, fear and anger. The measures are: mutual information, correlation dimension, correlation entropy, Shannon's entropy, Lempel-Ziv complexity and Hurst exponent. The measures are extracted from the time series, after performing a phase space embedding using Takens' time-delay embedding theorem [208]. In addition, summary statistics of the measures are computed resulting in a set of 24 features. In order to select the best features among the extract measures, the authors apply a feature selection methodology that

results in four features: standard deviation of Hurst exponent, mean of CD, mean of Lempel-Ziv complexity and skewness of Takens estimator of CD. Then the performance of the measures in discriminating between neutral and two emotional states, fear and anger, in speech is evaluated using a neural network classifier. Authors conclude that fear and angry speech show more complexity than neutral speech. Their approach achieves a classification success rate of 93.78±3.18% in discriminating between neutral, anger and fear speech using only four complexity features, although it would be beneficial to see how those measures perform in discriminating neutral and positive emotions, and how their measures compare to emotion classification using acoustic features.

In searching for new measures by which to characterize emotional speech uttered by males and females, Shahzadi et al. 2015 studied the geometrical properties of the reconstructed phase space of speech signals [197]. They extracted four descriptor contours, in addition to statistical information of the contours. In order to compare the performance of their NLD measures with baseline methods, they extract prosodic and spectral features from the speech signals, and perform different classification tasks on 7 emotion categories for each type of features: NLD, prosodic, spectral and a combination of NLD, prosodic and spectral. They achieve an overall recognition rate of 82.72% and 85.90% for males and females respectively. Given that their NLD measures describe geometrical properties of the phase space, they highly depend on the embedding parameters needed for phase space reconstruction. Therefore it is uncertain to what extent these measures would generalize if the embedding parameters changed, and whether they can be considered as invariant estimates.

4.6.2 Pathology Detection in Voice

Recent works in nonlinear dynamics have investigated the detection of various pathologies from the voice signal. Henriquez et al. investigate the suitability of NLD features in characterizing healthy and pathological voice in speech signals. The measures are: the first- and second-order Rényi entropies, the correlation entropy, the correlation dimension, the first minimum of the mutual information function (FMMI) and Shannon entropy [74]. Authors found that the proposed measures are useful to discriminate between healthy and pathological speakers, with accuracies of 82.47% and 99.69% for two respective databases.

In their investigation of the detection of laryngeal pathologies, Alonso et al. compute the correlation dimension and the largest Lyapunov exponent (LLE) from a database of recorded voices of healthy and pathological speakers uttering five Spanish vowels (‘a’, ‘e’, ‘i’, ‘o’ and ‘u’) [4]. Global success rates of 91.77% was achieved with classic parameters, whereas a success rate of 92.76% was achieved with the complexity estimates.

Another area of research uses NLD in order to diagnose Parkinson’s disease from voice signals. One such study uses recurrence pitch entropy density (RPDE) and detrended fluctuation analysis (DFA) [98]. They achieve a classification performance rate of 93.82% using k-nearest neighbour (KNN) classifier.

4.6.3 Music Information Retrieval

Serrà et al. 2009 [195] investigated the use of cross recurrence plots (CRP) and cross recurrence quantification analysis (CRQA) measures for the identification of cover songs. A CRP is a bivariate extension to RP and is constructed similarly. It is used to analyse the dependencies between two different systems by comparing their states through the study of interrelations between time series [134]. Cover songs are alternative renditions of a previously recorded musical piece, that resemble their originals with respect to some features. Given two songs, Serrà et al. extract chroma descriptor time series and transpose one song to the main tonality of the other. The time series are the embedded using Takens' delay-coordinate embedding theorem [208], using arbitrary values of m and τ. The best combination of m and τ is selected that yield the highest cover song identification accuracy. Subsequently they construct a CRP and apply Q_{max} algorithm - derived from the RQA measure L_{max} - to extract features that are sensitive to cover song CRP characteristics. The dataset consists of 1953 commercial songs with an average song length of 3.5 min ranging from 0.5 to 7 min. They test their method at the 2008 MIREX [38] cover song identification contest. By using this approach, they are able to identify cover songs with a higher accuracy as compared to previously published methods.

In another work [194], Serrà et al. 2011 employ RPs and recurrence histograms to extract information from music audio frames, that is subsequently used for music genre classification. To perform delay-coordinate embedding, they do not refer to the known algorithms for the selection of m and τ but choose their values arbitrarily. They compare their method against a baseline obtained from standard spectrum-based descriptors namely: MFCC, chromas, brightness, roll-off, spectral centroid, spectral spread and spectral flatness in addition to mean, variance, skewness and kurtosis of histograms and spectra. They apply correlation feature selection as well as dimension reduction with PCA to feature set. The selected feature set is then evaluated using several classifiers, including nearest neighbours and SVM. They find that the classification accuracy using the spectral-based descriptors surpasses the one achieved by the histogram-based descriptors. However, when they combine both types of descriptors they notice an improvement by up to 5%. This finding was true regardless of the combinations of the embedding parameters. Their results encourage future work to create feature sets combining acoustic as well as NLD descriptors.

4.6.4 Scene Event Classification

Aiming to quantify the temporal dynamics of auditory scenes, Roma et al. 2013 [169, 170] propose to extract RQA measures from MFCC vectors extracted from the time series. By employing such an approach, they are able to analyse the temporal evolution of the features by computing RPs from overlapping windows of fix size for long time series. Using a feature set that consists of MFCC features as well as 11 RQA measures, their method is able to achieve a classification accuracy of 71%. When they use only RQA features in the classification task, their accuracy is of 55%. We emphasize on the fact that the RQA measures are estimated after embedding the times series. Therefore they are not invariant, and it would be interesting to see how the measures will perform when the embedding parameters are modified.

4.6.5 Emotion Recognition from Physiological Signals

Other methods that apply NLD estimates to emotion recognition tasks, extract NLD measures from physiological signals such as: ElectroCardioGram (ECG), ElectroDermal Response (EDR), ReSPiration activity (RSP). One such method [216] extract standard features from those signals as well as NLD features. The stimuli consist of images taken from the IAPS database [105]. Using a Quadratic Discriminant Classifer (QDC), the method evaluates the measures' performance in recognizing five levels of valence and five levels of arousal. The classification accuracy using standard features was acceptable for the neutral class only, while the remaining valence and arousal classes were misclassified. Interestingly, when NLD features were considered, each level of valence and arousal as well as the neutral class were successfully recognized. This highlights the great contribution of NLD estimates to the field of emotion recognition from physiological signals.

Other studies use NLD complexity features for emotion recognition in electroencephalograph data (EEG). The features include the Hurst exponent, approximate entropy and fractal dimension [226], or Lempel-Ziv complexity measurement [25]. Further works include the application of NLD measures for the analysis of epilepsy from EEG [2], the study of the differences between sleep stages and wake using cardio-respiratory signals [168]. As these works are outside the scope of this thesis, we omit elaborating further on the respective methods.

4.7 Our Contribution

In chapters 5 and 6, we address some of the limitations mentioned earlier by performing the following:

1. We combine nonlinear dynamics approach with STSA to estimate symbolic RQA_{S} measures from the RP_{S} of sounds.

2. The RQA_{S} estimates are also computed for the logistic map. We compare them with the same measures derived with embedding. The advantage of our method is that our RQA_{S} measures can be considered as dynamical invariants since they are obtained independently of the embedding parameters m and τ.

   (a) The performance of the RQA_{S} measures in the recognition of affect is evaluated on three datasets: vocal, musical and auditory scenes.

   (b) The generalization of the symbolic estimates is evaluated in a cross-domain classification task.

3. We compute symbolic dynamical invariants, namely the correlation dimension, correlation entropy, the Shannon entropy and the Lyapunov exponent. We show that they correlate in a similar way with the same measures obtained with embedding. We evaluate their performance in emotion recognition with and without embedding, using two datasets. We further compare our approach using nonlinear dynamics features with a baseline approach

using standard acoustic features. Finally we examine the performance of a hybrid set of acoustic as well as complexity features, for affect recognition.

Chapter 5

Emotions in Strange Attractors: Modeling Affect with Nonlinear Symbolic Dynamics

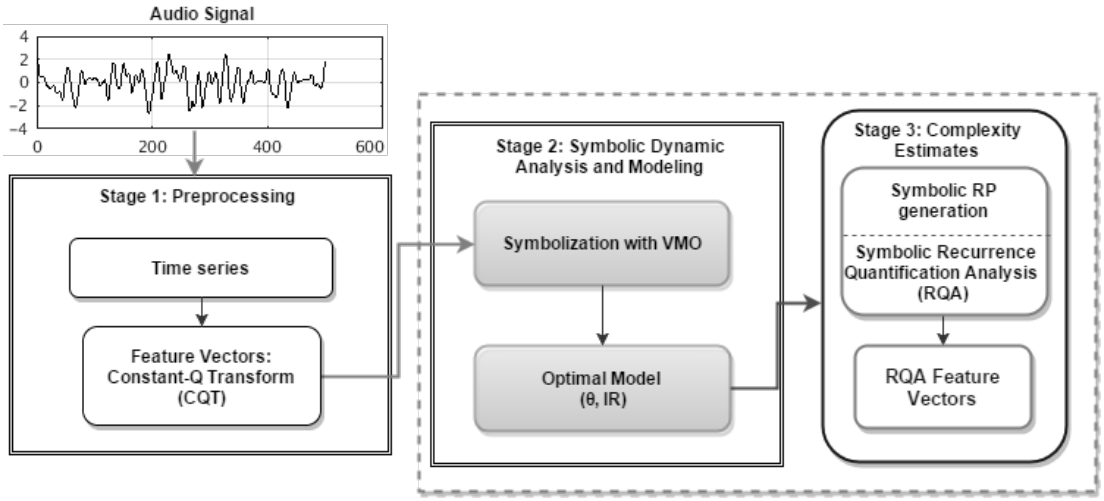

Figure 5.1: Nonlinear Symbolic Dynamical Framework

The previous part focused on the estimation and analysis of low-level acoustic measures from sounds, and mapping them to higher-level affective semantics. Although this approach is widely applied, it assumes implicitly that the affective knowledge communicated in the signal is captured exclusively by the signal's acoustic properties. A primary concern with this approach is that there is yet no consensus on a recommended set of acoustic features for ER studies, that would generalize for different sounds and emotions, despite the multitude of acoustic measures proposed and evaluated with various learning algorithms.

Complex audio signals such as vocal, musical or textures of auditory scenes have intrinsic chaotic dynamics phenomena that carry complex perceptual meanings, that we as listeners perceive and appraise in our minds. Arguably, capturing the relevant emotional meaning from audio structure while disregarding trivial or irrelevant information is a complex process that cannot be inferred by linear approaches nor by using low-level acoustics. Such complex infor-

mation is shaped in the nonlinear dynamical structure of audio content that is brought together by repeating patterns evolving in a temporal order. Nonlinear dynamics analysis consists of a set of methods that unravel these fine-grained patterns, and study their role in conveying meaningful information.

In this part, the focus is on characterizing the perceptual salience of the dynamical attractors of the sounds. Since our hearing system captures the sounds first, the first step consists of transforming the audio signal into feature vectors that approximate human auditory analysis. Then, in order to quantify the meaningful patterns that capture the perceptual information, the second step applies a method of nonlinear symbolization that extracts salient patterns from the signal as well as their temporal order. The third step applies methods of dynamical statistical quantification analysis that quantifies the patterns. In the last step, we evaluate the performance of the quantification measures for ER using supervised learning tasks.

5.1 Motivation

Recent studies have shown that nonlinear phenomena exist in complex signals such as voiced, musical or audio textures from auditory scenes [22, 23, 75–77, 197]. This notion has oriented a good deal of research into characterizing complex signals in terms of their inherent nonlinear dynamics. More importantly, the same studies have shown that the dynamical attractors carry perceptual meaning to the listener:

1. Generally in the case of ER from vocal sounds, the acoustic features studied are based on the source-filter model of voice production (see chapter 1 figure 2.1). This model involves different mechanisms for the generation of voiced sounds, and acoustic properties reflect these mechanisms. However there are 2 main limitations to this approach:

   (a) There is still no theoretical basis for associating the acoustic properties of the vocal sounds to the speaker's emotional state [5, 123, 127].

   (b) The source-filter model does not explain the nonlinear dynamical phenomena that are inherent to the physiological processes that are involved in sound production, and that exist in vocal signals [22, 75–77, 197]

2. Similarly to voice, musical sounds have inherent nonlinear dynamical characteristics [66, 171, 194, 196]. Conventional low-level features cannot capture the properties of musical instruments that sway our affective perceptions.

3. Audio textures from auditory scenes exhibit chaotic behaviour that can only be depicted using nonlinear measures [169, 170].

As explained in chapter 4, since traditional statistical methods are insufficient for describing the dynamics of chaotic systems, nonlinear time series analysis (NLTSA) methods as well as symbolic time series analysis (STSA) are employed to model affect in sounds. Our approach is detailed next.

5.1.1 Contribution

The contribution of our proposed method is that our model combines the nonlinear dynamics analysis with a method of time series symbolization that circumvents the need for attractor reconstruction with embedding and that does not involve a partitioning of the phase space with generating partitions.
During phase space reconstruction, different values of the embedding parameters m and τ generate different phase space structures that are depicted differently in the RP (figure 5.2). Consequently the values of the derived RQA estimates will vary considerably, because RQA measures are not invariants but they depend on the embedding parameters [134].
Figure 5.2 shows three different phase spaces that correspond to a segment of a *tire skids* sound. Phase space (C) displays the phase space obtained given an optimal τ, (A) and (B) display the phase spaces for a small and large τ respectively.

We apply our symbolic measures of periodicity and complexity to the time series of the logistic map, and show that they are similar to the same measures computed directly from the time series as reported in literature [134]. Then the performance of our symbolic measures is evaluated for the recognition of emotion in voice, music as well as auditory scenes.

Our method is computationally efficient as it runs in linear time and space, and performs better than state-of-the-art methods that are based on low-level acoustic features, or that employ standard nonlinear dynamics approach.

Finally we advocate for a broader investigation of this feature set for ER tasks, and to integrate them in ER challenge tasks such as INTERSPEECH or MIREX.

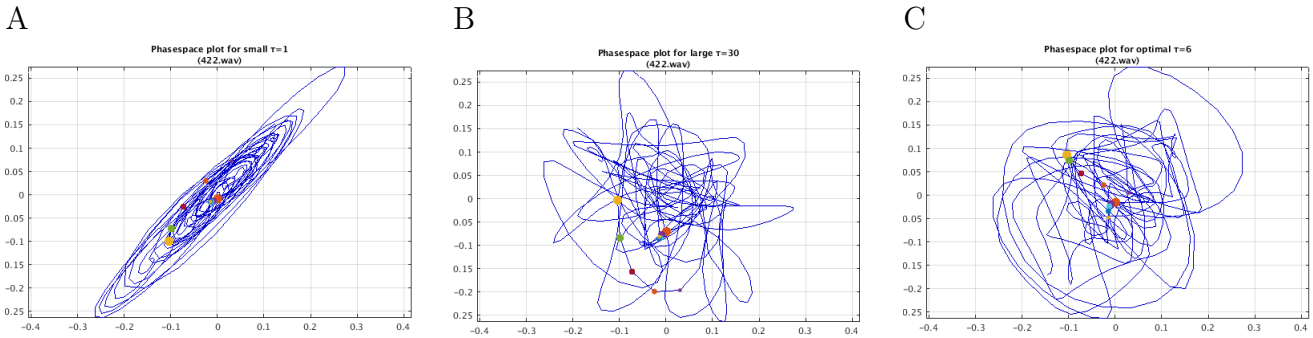

Figure 5.2: Phasespaces for different τ values. Sound: *Tire Skids*. A) Small τ. B) Large τ. C) Optimal τ

5.2 Our Approach

From an information theoretic approach, the signal and the listener are part of one model, since the signal's structure can be defined as the elements of audio material that the listener can predict [41]. In order to model the emotional information in sounds in a way that accounts for the listener's auditory perceptions, we propose a model that integrates nonlinear dynamics

with an information theoretic framework that quantifies the amount of affective information shared between the sound and the listener's auditory system.

Framework The framework is described at the beginning of the chapter in figure 5.1. In a preprocessing stage, the time series is transformed into a constant-Q feature vector (CQT). CQT is a logarithmic spacing of filter center frequencies versus bandwidths, that represents the audio signal in a form that approximates human auditory analysis. Next in stage 2, the CQT feature vector is passed as input to the VMO construction algorithm, that generates several symbolizations of the features in terms of their recurrence properties. Then by means of information rate (IR), the optimal threshold θ is evaluated to obtain the optimal VMO symbolization model $\mathcal{M}_{\mathcal{S}}$. In stage 3, the symbolic $RP_{\mathcal{S}}$ is generated from the self-similarity matrix created from the longest repeated substrings (LRS) of $\mathcal{M}_{\mathcal{S}}$. Then a recurrence analysis of the $RP_{\mathcal{S}}$ infers the $RQA_{\mathcal{S}}$ complexity estimates.

5.2.1 Symbolic Recurrence Plot

Given the optimal VMO symbolization $\mathcal{S}$ of the signal, the corresponding $\mathrm{RP}_{\mathcal{S}}$ is obtained. Here we redefine the RP as a symbolic recurrence plot $\mathrm{RP}_{\mathcal{S}}$, obtained through the binary self-similarity matrix of the optimal VMO-symbolization of the signal's time series:

$$R_{i,j}^{\sigma_M,\theta} = \Theta(\theta - d(\sigma_{q_i}, \sigma_{q_j})) \quad i,j = 1,...,N \tag{5.1}$$

Such that:

$$R_{i,j}^{\sigma_M,\theta} = \begin{cases} 1 & if\ d(\sigma_{q_i}, \sigma_{q_j})\ is \leq \theta \\ \\ 0 & otherwise \end{cases} \tag{5.2}$$

Where N is the number of states considered, σ_M refers to the M^{th} symbolized substring, Θ is the Heaviside step function (i.e. $\Theta(x) = 0$ if $x < 0$, and $\Theta(x) = 1$ otherwise). θ is a threshold distance, and $d(\sigma_{q_i}, \sigma_{q_j})$ is a distance metric between pairs of symbolized substrings q_i at $t = i$ and q_j at $t = j$.

Once the optimal VMO symbolization of the signal is obtained, then several $\mathrm{RQA}_{\mathcal{S}}$ features are extracted from the $\mathrm{RP}_{\mathcal{S}}$.

5.2.2 Symbolic Recurrence Quantification Analysis

Recurrence quantification analysis is a set of nonlinear statistical measures based on the recurrence matrix, and that quantify the topological structures seen in the recurrence plots.

Fifteen $RQA_{\mathcal{S}}$ features studied in this chapter are recurrence rate ($\mathrm{REC}_{\mathcal{S}}$), determinism ($\mathrm{DET}_{\mathcal{S}}$), average diagonal line length ($\mathrm{L}_{\mathcal{S}}$), length of the longest diagonal excluding the main diagonal $L_{S_{max}}$, divergence ($\mathrm{DIV}_{\mathcal{S}}$), entropy of the diagonal line lengths distribution $S_{RP_{\mathcal{S}}}$ and $RATIO$ which is the fraction $\frac{DET_S}{REC_S}$.

In order to account for our definition of the $\mathrm{RP}_{\mathcal{S}}$, two of the $RQA_{\mathcal{S}}$ measures are redefined next:

1. $\mathrm{REC}_{\mathcal{S}}$ is the percentage of points in the $\mathrm{RP}_{\mathcal{S}}$ excluding the points on the main diagonal line:

$$RR_{\mathcal{S}} = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j:i\neq j}^{\sigma_M,\theta} \tag{5.3}$$

2. $\mathrm{DET}_{\mathcal{S}}$ is the percentage of points in the diagonal lines of minimal length l_{min}:

$$DET_{\mathcal{S}} = \frac{\sum_{l=l_{min}}^{N} l \times P^{\theta}(l)}{\sum_{i,j}^{N} R_{i,j}^{\sigma_M,\theta}} \tag{5.4}$$

Application of $RQA_{\mathcal{S}}$ for the logistic map

We illustrate the application of the $RQA_{\mathcal{S}}$ for the logistic map (figure 5.3), and a qualitative comparison can be made with plots from [134] that are included in section 5.6 at the end of this chapter.

Mathematically, the equation of the logistic map is defined as:

$$x_{i+1} = ax_i(1 - x_i) \tag{5.5}$$

where x_i is a real number between zero and one and a is a positive constant. We generate multiple time series from the logistic map and define the control parameter $r \in [3.5, 4]$, with $\Delta r = 0.0005$, so that for each r we have a separate time series T of length 1000. The values of the parameters are set in order to compare the results with [134] and accordingly, we embed the time series with dimension $m = 3$ and time delay $\tau = 1$ [134].

Figure 5.3 shows plots of our VMO-derived $RQA_{\mathcal{S}}$ measures. A comparison of the plots below is made with similar plots derived directly from the time series after embedding in [134].

The measures $DET_{\mathcal{S}}$, $L_{\mathcal{S}}$, $Lmax_{\mathcal{S}}$, based on the diagonal lines, show similar maximas at the periodic-chaos/chaos-periodic transitions as in [134]. Also, $Lmax_{\mathcal{S}}$ detects all such transitions, but $DET_{\mathcal{S}}$ and $L_{\mathcal{S}}$ do not find them all.

Similarly, the chaos-chaos transitions to the laminar states, are depicted by the measures based on the vertical structures, namely: $LAM_{\mathcal{S}}$, $TT_{\mathcal{S}}$ and $Vmax_{\mathcal{S}}$. The difference between $LAM_{\mathcal{S}}$ and $Vmax_{\mathcal{S}}$ is that $LAM_{\mathcal{S}}$ only measures the amount of laminar states, while $Vmax_{\mathcal{S}}$ estimates the maximum duration of the laminar states [134]. The lines in $Vmax_{\mathcal{S}}$ plot fall to zero within the period windows, indicating that the chaos-order transitions are also identified. This is in agreement with [134] who states that RQA measures are able to identify bifurcation points. However the $LAM_{\mathcal{S}}$ plot shows a different structure, it displays minimas or drops that correspond to the chaos-chaos transitions, while in the referenced work the LAM plot displays maximas or peaks at the same locations. This may be due to the fact that our $LAM_{\mathcal{S}}$ is derived from a symbolic representation of the series rather than the data itself. However as in the

referenced paper, $LAM_{\S}$ is different from the other two vertical-based measures $Vmax_{\S}$ and $TT_{\S}$, in that it does not peak at inner crises, possibly because it is more robust against noise in the data. Finally similarly to the method in [134], with our symbolization method a 1000 data points are enough to derive the RP-based measures.

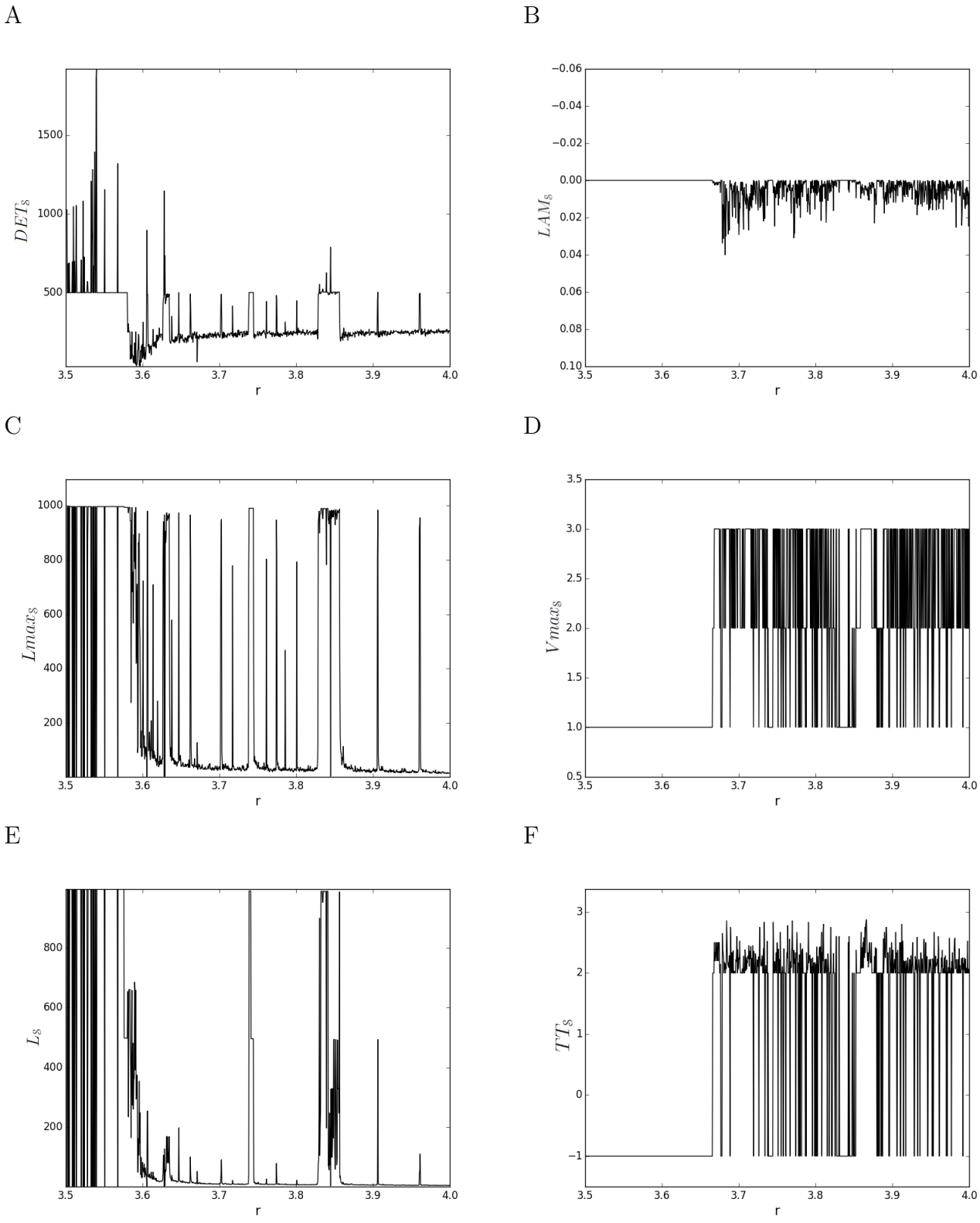

Figure 5.3: RQA_S of the logistic map: $r \in [3.5, 4.0]$, $\Delta r = 0.0005$ and $T = 1000$

5.3 Experiments

5.3.1 Stimuli

In addition to the databases investigated in chapter 3 namely, the Montreal Affective Voices (MAV) [12] and the Musical Emotional Bursts (MEB) [149], the International Affective Digitized Sounds (IADS-2) [20] is included in the study, and it consists of a set of standardized affective stimuli designed for the purpose of experimental investigations of emotion and attention. The stimuli consists of 167 auditory scenes of 6 seconds duration each at 44100 Hz sampling rate. Human affective labelling on the two-dimensional model of affect showed that the stimuli evoke reactions across the entire range of valence and arousal.

Dataset

For each sound belonging to the three databases, 15 RQA_S measures are computed, as mentioned in section 5.2.2.

5.3.2 Classification Scheme

This section details the classification scheme that was conducted to evaluate the dynamical features performance for each of the three datasets.

Feature Selection No feature selection algorithms were applied prior to the classification task.

Learning Model A feedforward neural network (NN) [39] in which the all the links are unidirectional, i.e. the information moves forward from the input nodes, through the hidden nodes and then to the output nodes. There are no cycles or feedback links involved, that is, the output information does not travel back to the network. The feedforward NN is chosen for its simplicity.

Learning with NN One the network's structure is defined, it is trained, tested and validated. The Levenberg-Marquardt and the scale conjugate gradient backpropagation learning algorithms are used and the validation is performed using the mean squared error (MSE).

Feature Extraction The CQT of each sound is obtained at 44100 Hz sampling rate, hop lenth of 512 and 84 bins. Then after symbolization, 7 RQA features are extracted and constitute the dataset used in the learning tasks.

Preprocessing The dataset is preprocessed before training such that column features are centered to have mean 0 and scaled to have standard deviation 1.

Classification The dataset is divided into 70% training, 15% validation and 15% testing. The classification tasks were conducted using the Neural Network Toolbox in MATLAB, in a multiclass one-versus-all fashion, such that each of the 6 affection subdimensions is in turn considered as positive and negative class. For example, the classification on the arousal dimension consists of 3 tasks: 1) classifying high arousal samples (positive class) versus non-high samples (low and neutral as negative class), 2) low arousal (positive class) versus non-low samples (high and neutral as negative class), 3) neutral (positive class) versus non-neutral (high and low as the negative class). Once the initial results are obtained, the number of neurons and hidden layers as well as the training algorithm are adjusted experimentally to improve the network's performance. The final results are subsequently averaged to get the classifier's performance for valence and arousal.

Performance Evaluation Overfitting can be a problem if the dataset is too small or biased. Therefore in addition to the MSE function, the Adaptive Synthetic Sampling (ADASYN) algorithm is applied in order to account for the imbalance in the dataset. ADASYN is an extension of the Synthetic Minority Oversampling Technique (SMOTE) and is a powerful method that deals with dataset imbalance by generating synthetic samples belonging to the minority class [73]. The classifier's performance is evaluated before and after dataset rebalancing, using a combination of performance metrics based on the confusion matrix. Since the results did not vary much between unbalanced and rebalanced dataset, we report below the original results on unbalanced dataset.

In order to ensure impartiality in the interpretation of the results and avoid the temptation of selecting the performance metrics that favour our model, we are reporting the classification's performance on seven metrics in addition to MSE. The metrics are: accuracy (ACC), precision or positive predictive value (PPV), recall or true positives rate (TPR), F_1-measure, F_2-measure, Cohen's Kappa (κ) as well as area under the receiver operating characteristic curve (ROC). The measures are explained in chapter 1.

Traditionally, the most frequently adopted metrics are ACC and MSE. However ACC does not provide accurate information on a classifier's functionality and is ineffective when there's data imbalance [73]. Therefore in our work, we use additional metrics that account for dataset imbalance, and that are not particularly sensitive to data distributions. For example precision is sensitive to changes in data, while recall is not. But recall doesn't give an insight about how many examples are incorrectly labelled as positive. And precision cannot assert how many positive examples are incorrectly labelled. The F_1-measure is the weighted harmonic mean of precision and recall that tends towards the lowest of the two, and by means of a single value, it provides a better insight into the functionality of a classifier. The F_2-measure sways recall more than precision, therefore emphasizing the false negative value which is the most critical element of the confusion matrix. Cohen's Kappa (κ) takes chance into account and overcomes the problem of overestimating the ACC, it measures the extent to which the agreement between observed and predicted values is higher than that expected by chance alone. According to [60], Cohen's kappa ranges from -1 (total disagreement), through 0 (random classification) to 1 (perfect agreement) [64]. Particularly, [104] considers 0-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect. Finally the AUC

provides a visual representation of the relative trade-offs between recall (TPR) and the false positives rate (FPR), and measures the ability of a classifier for various threshold settings [73].

5.4 Analysis

5.4.1 Results

In this section, the classification performance is reported for two datasets, the vocalizations (MAV) and the film music excerpts (FME). The tables depict the results obtained before as well as after rebalancing of the datasets with ADASYN.

Qualitative description of RP_S

The two figures below show symbolic recurrence plots generated from the VMO-symbolized series of environmental sounds. The RQA_S computations quantify those recurrences seen in the plots, that are the optimal representation of the signal. Figure 5.4 shows three RP_S of three different sounds. The sounds are labelled with a 'pleasant' valence. The patterns depicted on the plots provide a description of the temporal behaviour of the dynamic trajectories. It can be seen that all three plots show regularities in the patterns. Figure 5.5 shows three plots of three different sounds labelled with 'unpleasant' valence. All three plot show sudden changes in the system's dynamics, with no regularities. Comparing the two figures with each other, it is clear that the dynamics of a sound expressing a positive emotion are different that those of a sound expressing a negative emotion. Moreover, the VMO-symbolization captures the salient dynamics that are responsible for the different affective perceptions. By introducing a symbolic recurrence quantification analysis from the VMO-symbolization, we hope to quantify those dynamics and therefore discriminate affective information in sounds.

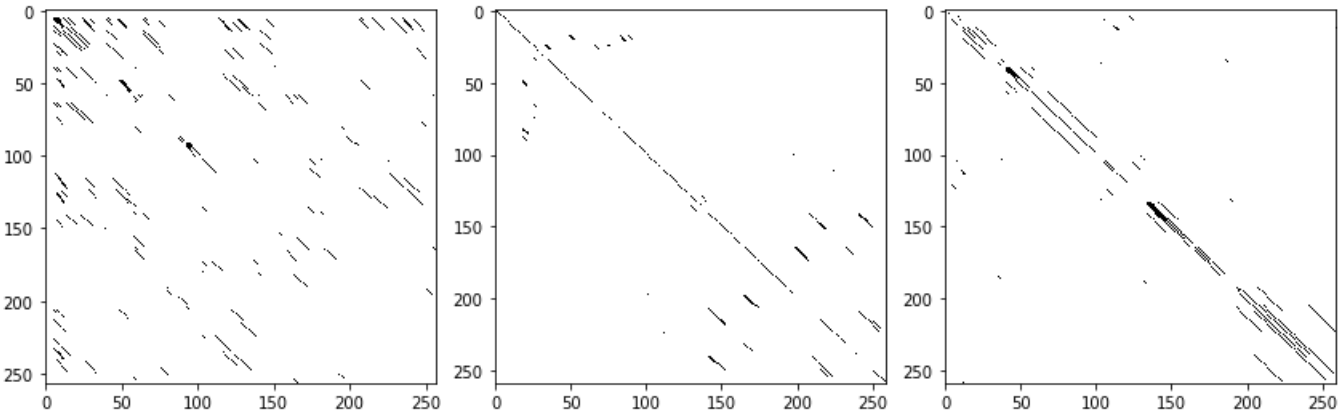

Figure 5.4: RP_S for *pleasant* scenes: countrynight (*left*), carousel (*middle*), tropical (*right*)

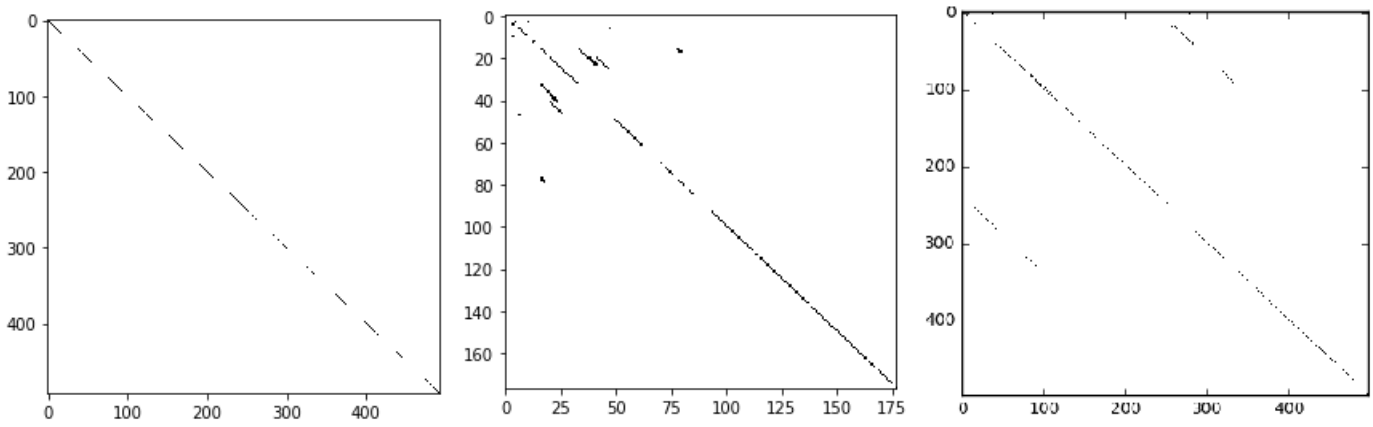

Figure 5.5: RP_S for *unpleasant* scenes: explosion (*left*), injury (*middle*), gun shot (*right*)

Performance of RQA_S

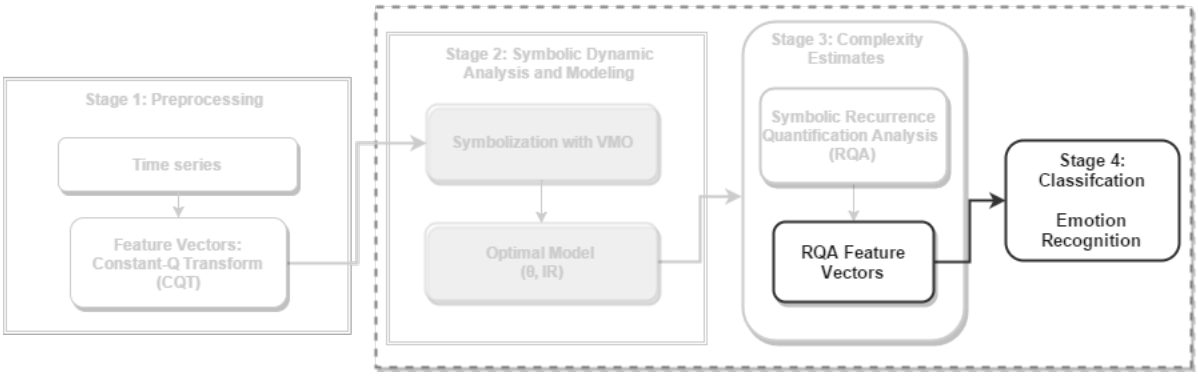

Figure 5.6: Dynamic Model of Affect Framework: stage 4

This part analyzes the performance results of the RQA_S in discriminating sounds by their emotional expressions. It corresponds to stage 4 in figure 5.6.

Tables 5.1 and 5.2 report the performance metrics for each of the four datasets, on valence and arousal respectively.

In table 5.1, the recognition of valence is high for musical clips, with 97% and 81% accuracy for the violin and the clarinet music respectively. The corresponding metrics values are high, especially for the violin with F_1, F_2 and κ values close to 1. This demonstrates that the recurrence measures capture with an impressive accuracy the dynamics of emotions in the violin musical clips. The ACC for MAV is 79% with similarly high values for the remaining metrics. κ is >0.40 for MAV, Violin and Clarinet, which shows a fair to good agreement between observed and predicted values for MAV and the clarinet, and excellent agreement for the violin music.

The lowest ACC is for the auditory scenes (IADS) with a 74% ACC and a κ=0.17, showing

poor agreement between the human labels and the classifier's performance using the RQA_s measures. This may be due to the nature of the sounds, that comprise human voices as well as various machinery sounds such as a lawn mower, plane, car, tires and various other sounds, that may arguably have different inherent dynamics that are perceived differently on the affective dimensions.

Table 5.1: Performance measures on valence

Datasets	ACC	PPV	TPR	F_1	F_2	κ	AUC
IADS	0.74	0.93	0.73	0.82	0.76	**0.17**	0.66
MAV	0.79	0.78	0.77	0.74	0.75	0.47	0.79
Violin	**0.97**	0.95	0.98	0.96	0.97	**0.93**	**0.98**
Clarinet	**0.81**	0.75	0.71	0.73	0.72	**0.56**	**0.85**

In table 5.2 the ACC values are >=90% for all datasets. Except for the IADS sounds, the κ values are >70% which shows an excellent agreement between human affective annotations and the classifier's performance, and the AUC values are > 0.90 which indicates a high discrimination ability of the classifier using the RQA_s measures. The ACC=90% for IADS, but κ =0.06, which indicates a poor agreement between the human labels and the classifier's results. In this work, the kappa score is used as an unbiased metric for comparing the reliability of the classifier's performance obtained with RQA_s across four different datasets. In other words, we want to determine the role of the sounds' nonlinear dynamics in conveying affect information. It may be that a larger dataset and further experiments are needed to better interpret the high recognition rate along high values for PPV, TPR, F_1, F_2 and AUC, and on the other hand, the low κ value. Another observation is that the recognition rates and performance metrics are higher for arousal than for valence, with the exception of the violin music.

Table 5.2: Performance measures on arousal

Datasets	ACC	PPV	TPR	F_1	F_2	κ	AUC
IADS	**0.90**	1.00	0.90	0.94	0.92	**0.06**	0.72
MAV	**0.93**	0.78	0.81	0.79	0.80	**0.75**	**0.96**
Violin	**0.90**	0.73	1.00	0.83	0.92	**0.76**	**0.92**
Clarinet	**0.90**	0.83	0.89	0.86	0.88	**0.78**	**0.95**

Cross-Corpora Evaluation In chapter 3 we referred to literatures that suggest the notion that music evolved from primitive affective bursts, and we conducted cross-domain classification task using a set of acoustic features to verify the generalizability of acoustic features across datasets, and to determine whether a holistic learning model could predict emotional information across voice and music. Our classification performance results showed that a modest overlap exists between affective vocalizations and music bursts with a learning performance

around 60%, however the overall results weighed more favourably towards the opposite notion, that music and voice do not entirely share a common acoustic code, and that although some overlap may exist between voice and music, music may not have entirely emerged from primitive vocal interjections.

In this section, we consider this theme from the angle of chaotic dynamics and we ask the question:

What do nonlinear dynamics reveal about the underlying dynamics of affect expression in voice and music? Are there inherent similarities in the nonlinear dynamics, that corroborate the notion that music evolved from primitive vocalizations?

In order to address this question, the RQA_S measures are evaluated in a cross-corpora classification task: a classifier trained on a given dataset, is tested on another dataset, and conversely.

Additionally, in order to make a comparative analysis of the performance of the RQA_S in a cross-domain task, i.e. between voice and music, we implemented an inter-domain classification task between the clarinet and the violin musical clips. We anticipated that the performance for this task will be higher, as it involves the dynamics of two musical sounds. Therefore a classifier trained on the clarinet dataset was tested on the violin dataset, and conversely.

Finally, the same task was implemented for the auditory scenes dataset (IADS-2) with the vocal and musical datasets. Here we expected the performance results to be rather low, because of the nature of the auditory sounds, consisting of various environmental activity sounds including machinery, human voice and other random sounds, compared to voice and music. Therefore a classifier trained on the IADS dataset was tested on the remaining datasets and conversely. The results are reported in tables 5.3 and 5.4 for valence and arousal respectively.

In table 5.3 we observe the very low kappa scores for the classification between voices and music clips: κ =[0, 0.07, 0.03, -0.0002]. The ACC is very low as well <60%. This shows that the RQA_S with the NN classifier fail to capture affective information, when the classifier is trained on one dataset (i.e. voice) and tested on another (i.e. music), and conversely. These results are important as they show that the dynamics of valence in voices may be very different from music clips. Our findings could not corroborate the notion that music evolved from primitive affect bursts, as our results do not support such claim. A similar analysis can be made for the cross-domain classification between IADS and the other datasets. This clearly highlights different dynamics on the expression of valence between environmental sounds compared to voices and music. An accuracy of 72% is achieved when a classifier is trained on a clarinet dataset and tested on the violin dataset, with a κ =0.30, indicating a fair agreement between human labels and the classification performance, and a AUC =0.95. However the remaining metrics are rather low, so perhaps a larger dataset is needed in order to draw conclusive remarks about this task. However it is important to note that performance metrics for the inter-domain task involving only the violin and the clarinet, are considerably higher than the remaining tasks. This may be due to the fact they are both music clips and may have similarities in their inherent dynamics. This is something that cannot be established for the cross-domain classification tasks.

In table 5.4, a 73% ACC and κ =0.21 is seen for the cross-domain learning task between voices and the musical clips. However the remaining metrics are considerably low. Although

Table 5.3: Cross-Corpora classification results on valence

Task	ACC	PPV	TPR	F_1	F_2	κ	AUC
$MAV_{classifier}$ x $Violin_{features}$	0.50	0.33	0.27	0.29	0.27	0	0.65
$MAV_{classifier}$ x $Clarinet_{features}$	0.55	0.48	0.37	0.38	0.37	0.07	0.79
$Violin_{classifier}$ x $MAV_{features}$	0.55	0.39	0.44	0.38	0.40	0.03	0.98
$Clarinet_{classifier}$ x $MAV_{features}$	0.52	0.30	0.40	0.30	0.34	-0.0002	0.95
$Violin_{classifier}$ x $Clarinet_{features}$	0.64	0.40	0.43	0.41	0.42	0.13	0.98
$Clarinet_{classifier}$ x $Violin_{features}$	0.72	0.51	0.51	0.50	0.50	0.30	0.95
$IADS_{classifier}$ x $Clarinet_{features}$	0.40	0.78	0.35	0.45	0.38	0.04	0.64
$IADS_{classifier}$ x $Violin_{features}$	0.44	0.83	0.36	0.48	0.40	0.07	0.64
$IADS_{classifier}$ x $MAV_{features}$	0.43	0.84	0.43	0.51	0.46	0.08	0.64

the values indicate a better recognition rate for arousal than for valence, they rather emphasize that the dynamics of arousal are also different for voice and music. The performance metrics for the cross-domain task involving the IADS sounds are considerably low for arousal as well, which suggest that the dynamics of affect communication differ considerably between auditory scenes compared with voices and music. There are slightly higher ACC and κ scores for the inter-domain task between the music clips, with κ =[0.12, 0.18], which shows a slight or poor agreement. Although the values are not high, however they are higher than the remaining tasks, which suggest that the characterization of affect is better depicted among music clips than across sounds belonging to different domains.

Table 5.4: Cross-Corpora classification results on arousal

Task	ACC	PPV	TPR	F_1	F_2	κ	AUC
$MAV_{classifier}$ x $Violin_{features}$	0.73	0.36	0.68	0.56	0.61	0.21	0.90
$MAV_{classifier}$ x $Clarinet_{features}$	0.64	0.25	0.34	0.35	0.34	-0.003	0.90
$Violin_{classifier}$ x $MAV_{features}$	0.66	0.35	0.18	0.23	0.20	0.04	0.89
$Clarinet_{classifier}$ x $MAV_{features}$	0.57	0.35	0.12	0.18	0.14	-0.03	0.95
$Violin_{classifier}$ x $Clarinet_{features}$	0.62	0.46	0.41	0.44	0.42	0.12	0.89
$Clarinet_{classifier}$ x $Violin_{features}$	0.66	0.45	0.45	0.44	0.45	0.18	0.95
$IADS_{classifier}$ x $Clarinet_{features}$	0.33	0.66	0.25	0.40	0.29	0	0.62
$IADS_{classifier}$ x $Violin_{features}$	0.32	0.66	0.16	0.40	0.29	-0.01	0.62
$IADS_{classifier}$ x $MAV_{features}$	0.33	0.66	0.11	0.20	0.13	0	0.62

5.4.2 Comparison to Previous Work

In chapter 3, we had made similar classification tasks using a set of acoustic measures. We had also employed several feature selection techniques in order to select the most informative

features and achieve a high emotion recognition rate. In this section we compare the results we obtained in chapter 3 with the present results.

In table 3.3a, the emotion recognition performance for the violin dataset on (valence, arousal) using 42 acoustic features, with SVM and without feature selection was of (66.66%, 76.66%), and (90%, 93.33%) with SFS greedy feature selection. With our present method using nonlinear dynamics quantification measures, we achieve (97%, 90%) using NN, using 14 RQA features only, and without any feature selection. The best result in chapter 3 was achieved using the nonlinear mutual information (MI) feature selection method, with accuracies of (96.67%, 100%) and a subset of 20 acoustic measures.

The classification performance on the clarinet stimuli was higher using the acoustic measures, achieving (93.33%, 100%) without feature selection using 42 features and (100%, 100%) using MI feature selection (see table 3.2a). In the present work we achieve (81%, 90%) using 14 RQA features.

The classification performance on the MAV stimuli (table 3.1a) was (78.33%, 90%) without feature selection and 42 features. After applying MI feature selection we achieved (95.26%, 97.63%). Here we achieve (79%, 93%) with the 14 RQA measures.

In [170], Roma et. al perform a scene classification task using a feature set that consists of MFCC features as well as 11 RQA measures. Using their method, they are able to achieve a classification accuracy of 71%. When they use only RQA features in the classification task, their accuracy is of 55% only. With our method and using RQA_8 only, we are able to achieve a classification accuracy of 90% in the recognition of affect in auditory scenes. Although our task classifies emotions rather than detects scenes, the higher accuracy we obtain reveal the strength of our method.

It is important to note that a known and common challenge in the ER field, is that it is very difficult to compare different methods, since researchers employ different stimuli, as well as different acoustic features and classification schemes that involve different dimension reduction techniques and feature selection methods. So it is hard to conclude what method performs best. Our previous method using bag-of features and feature selection methods reported in chapter 3 suffers from this drawback as well. However, our current method is proven successful in the characterization of affect in auditory scenes, affective vocalizations, as well as musical emotional bursts using two different instruments: the clarinet and the violin. This shows the effectiveness of our approach for ER studies. Furthermore, the measures employed are standard nonlinear dynamics quantifications, that can be used for any type of stimuli having nonlinear dynamics states.

5.4.3 Perspectives

We aim to test our approach and symbolic features on an international evaluation task such as MIREX[1]. Furthermore, initial experimentations proved our method efficient in discriminating Parkinson's disease from vocal recordings, as well as in recognizing instruments in music. Future work will develop this further.

[1] http://www.music-ir.org/mirex/wiki/2007:AMC

5.5 Dissemination

This work appears in:

1. Pauline Mouawad and Shlomo Dubnov. Novel Method of Nonlinear Symbolic Dynamics for Semantic Analysis of Auditory Scenes. In proceedings of the First International Workshop on Semantic Computing for Entertainment, 11th International Conference on Semantic Computing, 2017.

2. Pauline Mouawad and Shlomo Dubnov. On Symbolic Dynamics and Recurrence Quantification Analysis for Affect Recognition in Voice and Music. In proceedings of the International Conference on Perspectives in Nonlinear Dynamics, 2016.

5.6 Reference Work

5.6.1 Logistic map bifurcation diagram

Multiple time series are generated from the logistic map. The values of the control parameter r are: $r \in [3.5, 4]$, with $\Delta r = 0.0005$, so that for each r we have a separate time series T of length 1000.

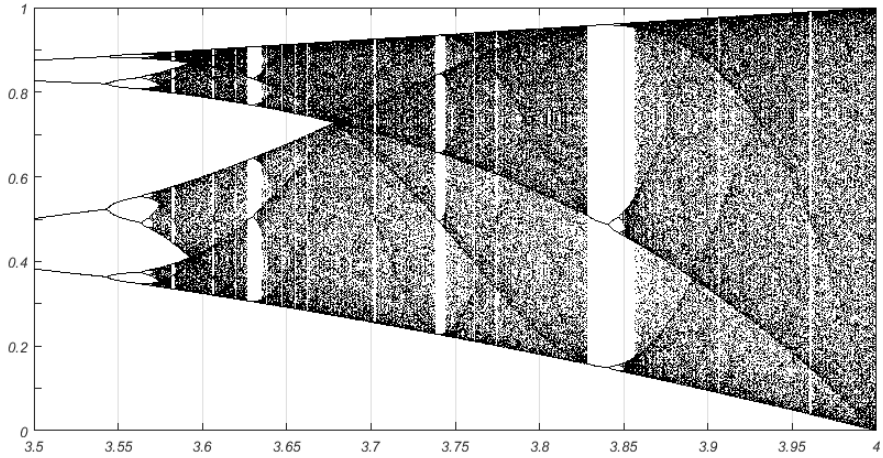

Figure 5.7: Bifurcation diagram of the logistic map. Control parameter $r \in [3.5, 4.0]$

5.6.2 *RQA* measures from the logistic map

The following plots display six *RQA* measures are computed from the logistic map time series: $DET, L_{max}, L, LAM, V_{max}, TT$ taken from [134]. The dotted vertical lines indicate points at which laminar states and band merging behaviour occur.

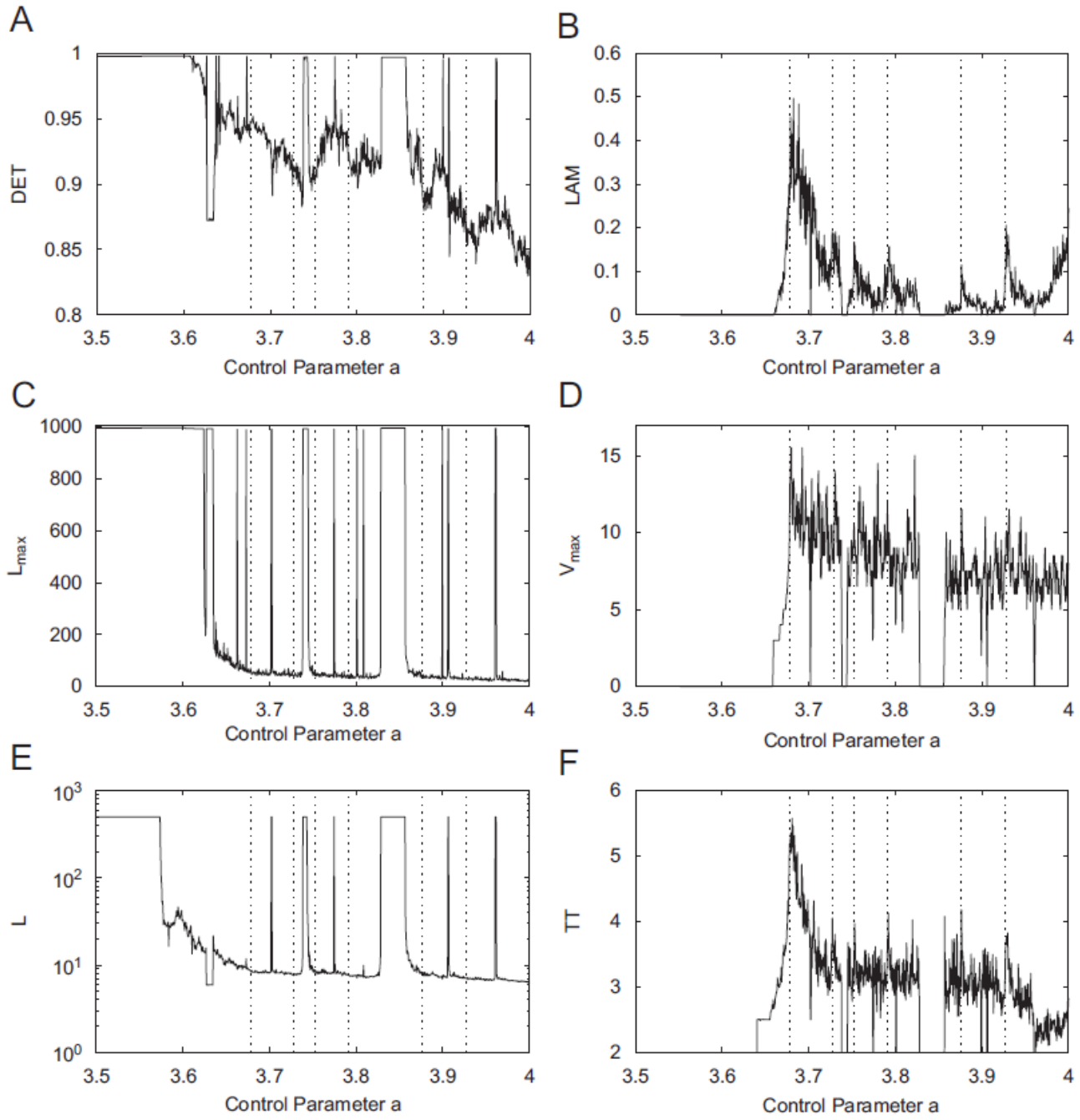

Figure 5.8: RQA measures for the logistic map series. Control parameter $r \in [3.5, 4.0]$ [134]

Chapter 6

A Novel Feature Extraction Method of Dynamical Invariants for Emotion Recognition

In the previous chapter, we proposed nonlinear symbolic measures of recurrence quantification analysis from symbolic RPs obtained with an adaptive symbolization of the time series, and evaluated their performance in the recognition of human emotions in audio signals. In this chapter, we further develop our approach and derive dynamical invariants from the symbolized model of the time series, namely: the correlation dimension (D2), the correlation entropy (K2) and the Lyapunov exponents (LE).

Our contribution lies in the method we employ to extract the invariants, and particularly, we show that the nonlinear estimates obtained with our method are in agreement with known methods from literature. Furthermore, we show that the symbolic invariants are well suited for the analysis of affect in sounds and that they achieve a higher performance rate for valence, compared to the rate achieved with RQA_S alone. Furthermore, we show that when the dynamical invariants are combined with acoustic features, they improve the overall ER rate.

6.1 Motivation

Chaos theory has been applied to different tasks of speech signal processing in the past two decades [76]. The complexity estimates that have been employed include the correlation dimension (D2), correlation entropy (K2) as well as the Lyapunov Exponent (LE), and these dynamical invariants can be accurately estimated from RPs [134]. The correlation dimension and Lyapunov Exponent have been successful in discriminating voice quality as well as in characterizing pathologies in voice [4, 74]. Furthermore, D2, K2 as well as the Shannon's entropy have been effective in the detection of emotion in speech [75, 76]. In the music domain, LE and D2 were used to characterize the clarinet tone [233], however rare are the literatures that investigate the potential of dynamical invariants in characterizing emotion in music.

In this chapter, we compute these dynamical invariants from a symbolized model of the

time series. We first probe to what extent our estimates are in agreement with known methods by illustrating their application to a simple example, the logistic map. Then we compare them with similar work from literature. Then we question the role of these symbolic dynamical invariants in discriminating emotion in voice as well as in film music. Finally we propose a dynamical model of affect suitable for voice and music.

6.1.1 Contribution

Our contribution in this work lies in the method we employ to compute dynamical invariants from the time series. Normally, in order to derive the LE using Rosenstein's algorithm or Eckmann's algorithm, the algorithm operates directly on the time series after embedding, and then computes the LE. In our approach we embed the time series and then symbolize it with VMO. Then we obtain from the VMO model the longest repeated substrings or motifs (LRS) in the signal as well as their lengths. Then we compute the LE, D2 and K2 from the LRS. This is a novel aspect where the dynamical invariants describe the chaotic behaviour of only the most important repetitions found in the series.

6.2 Framework

Given a one-dimensional time series obtained from an audio signal, first we embed it using Takens' time-delay embedding method. The dimension m is determined by the false-nearest neighbour algorithm, and the value is chosen where the false nearest neighbours are zero. The value of the time delay τ is defined by the first minimum of the mutual information function (FMMI). Next, we symbolize it with VMO, and select the optimal VMO model by means of IR. Then from the selected VMO model we obtain a representation of the lengths of LRS found in the series. From this representation, we proceed to extract three dynamical invariants: the correlation dimension (D2), the correlation entropy (K2) and the Lyapunov Exponent (LE). The framework is depicted in figure 6.1.

6.2.1 Complexity Features

In addition to the dynamical invariants, the IR as well as the optimal threshold (θ) of the VMO model construction are included in the nonlinear feature set. Their respective definitions included in this section.

Correlation Dimension The correlation dimension (D2) is a geometric measure that estimates the complexity of the system's dynamics: a higher D2 indicates a more complex dynamics. $D2_{\mathcal{S}}$ is computed from the symbolic $\mathrm{RP}_{\mathcal{S}}$ by the correlation sum [134]:

$$D2_{\mathcal{S}} = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ j \neq i}}^{N} \Theta(\theta - d(\sigma_{q_i}, \sigma_{q_j})) \tag{6.1}$$

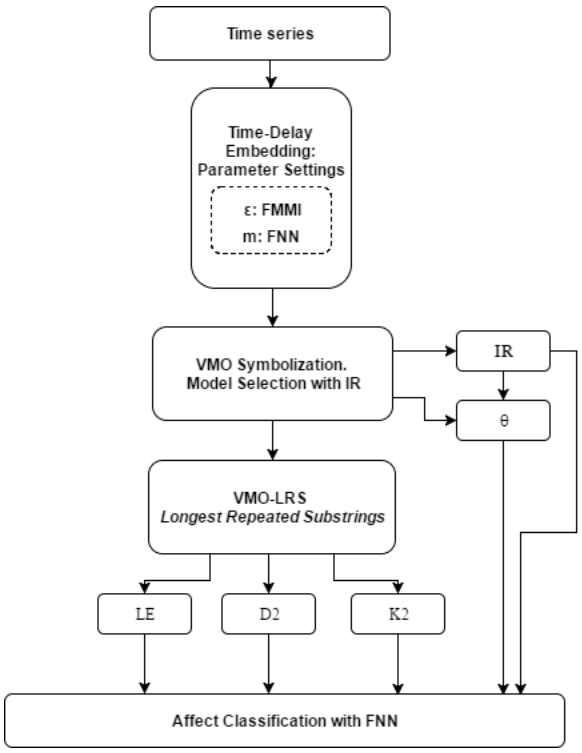

Figure 6.1: Framework: Extraction of Complexity Features

Correlation Entropy The correlation entropy (K2) also known as 2^{nd} order Rényi entropy quantifies the loss of information in time, in a dynamic system. It is estimated from the symbolic RP_S as:

$$K2_S = -ln(D2_S) \tag{6.2}$$

Lyapunov Exponent Lyapunov exponents (LE) estimate the amount of chaos in a dynamical system by quantifying the exponential divergence of initially close phase-space trajectories. A system with one or more positive LEs is defined to be chaotic.

Information Rate IR is the measure used in the VMO construction algorithm that selects the most informative symbolized sequence among different structures generated by different θ values. Let $x_1^N = x_1, x_2, ..., x_N$ be a time series x with N observations, where $H(x) = -\sum P(x) log_2 P(x)$ is the entropy of x, then the definition of IR is [223]:

$$IR(x_1^{n-1}, x_n) = H(x_n) - H(x_n|x_1^{n-1}) \tag{6.3}$$

And it is approximated by replacing the entropy terms in equation 6.3 by a complexity measure C associated with a compression algorithm [223]:

$$IR(x_1^{n-1}, x_n) \approx C(x_n) - C(x_n|x_1^{n-1}) \tag{6.4}$$

IR is maximized when there is balance between variation and repetition in the symbolized signal, which means that a VMO with a higher IR value captures more of the repeating patterns than a VMO with a lower IR value [223]. For more details about the IR in VMO construction, the reader is referred to chapter 4.

Next we illustrate the application of the complexity measures for the logistic map, and show that they correspond to measures computed directly from the time series in the case of LE, and to measures computed from RPs in the case of D2 and K2.

6.3 Application to the logistic map

The definition of the logistic map is provided in Eq.(5.5). To test the suitability of our symbolic complexity measures, we compute them for the logistic map. Similar to what is done in chapter 5, for the generation of multiple time series from the logistic map, we define the control parameter $r \in [3.5, 4]$, with $\Delta r = 0.0005$, so that for each r we have a separate time series T of length 1000. The values of the parameters are set in order to compare the results with [134] so accordingly we embed the time series, with dimension $m = 3$ and time delay $\tau = 1$, although an embedding is not required for maps (i.e. $m = 1$) [134]. For each of LE, D2, K2, IR as well as θ, there is one value per time series.

Figure 6.2 portrays plots of the LE computed directly from the time series after embedding, the $LE_{\mathbb{S}}$ obtained from the symbolized series, the Shannon entropy ($S_{RP_{\mathbb{S}}}$) estimated from the $RP_{\mathbb{S}}$, and $K2_{\mathbb{S}}$. The formal relationship between the correlation entropy K2 and the Lyapunov Exponents LE is [134]:

$$K_2 \leq \sum_{\lambda_i > 0} \lambda_i \tag{6.5}$$

where λ_i denote the Lyapunov exponents. From Eq.(6.5) one sees that K2 is a lower bound for the sum of the positive Lyapunov exponents.

6.3.1 Qualitative Comparison of Dynamical Invariants

There are different dynamical regimes and transitions that occur between the values in the range of r of the logistic map. They appear in the form of accumulation points, periodic and chaotic states, band merging points, period doublings and various order-chaos, chaos-order as well as chaos-chaos transitions [29, 134].

In [214], authors applied various measures to derive statistical meaning from the graphical structures of recurrence plots. One such measure is the Shannon entropy of diagonal line length distribution in the RP ($S_{RP_{\mathbb{S}}}$), which is the probability to find a diagonal line of exactly length l in the RP. The definition of $S_{RP_{\mathbb{S}}}$ is included in chapter 4.

The Shannon entropy is a measure of the complexity of the RP, for example, for uncorrelated noise it takes a small value, which indicates a low complexity [134]. However it is not a dynamical invariant [156].

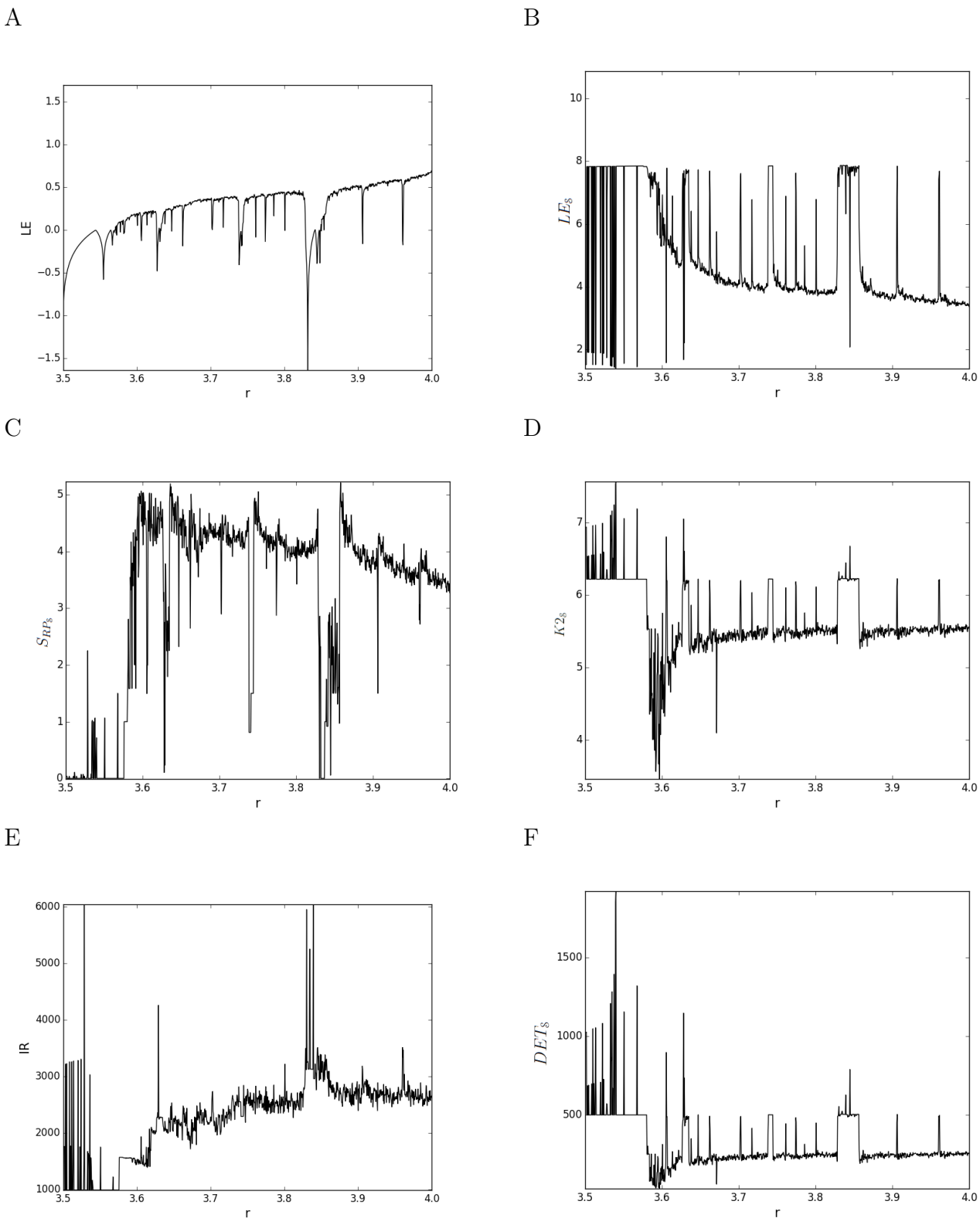

Figure 6.2: Dynamical Invariants of the logistic map: $r \in [3.5, 4.0]$, $\Delta r = 0.0005$.(A) LE from the time series. (B) LE_S from LRS. (C) Shannon entropy from RP_S. (D) $K2_S$ from RP_S.

According to Eckmann et al. 1987, the diagonal line lengths on RPs are related to the inverse of the largest LE [45]. This is true for some cases despite the fact that empirical studies have shown that S_{RP_S} is capable of identifying dynamical transitions, and therefore should grow as the system's complexity grows. [53].

Since the Shannon entropy quantifies the complexity of the dynamical system being studied, it is expected that its values increase when the system develops, that is when it varies from nonchaotic to chaotic regime [53]. Hence it is expected to be positively correlated with the LLE, rather than negatively correlated. And within periodic windows, the entropy should considerably decrease.

Therefore in a recent work, [117] has proposed another estimation of the Shannon entropy from RPs from the relative frequency of the occurrence of the diagonal segments of nonrecurrent points formed by white dots, that are a signature of complexity within the data. In this case a one-to-one correspondence was seen between the new Shannon entropy estimate and the positive LLE. That is, the Shannon entropy increased as the bifurcation parameter of the logistic equation increased, as illustrated in the plots found in the referenced paper. However, [53] claim that the definition of the entropy from the white nonrecurrent dots does not solve the problem of the negative correlation between the entropy and the LLE.

Here we compute the LLE from the logistic map time series, after embedding and symbolization with VMO. The LLE is computed from the VMO-derived LRS, and the Shannon entropy is estimated from the VMO-derived symbolic RP_S.

Figure 6.2 (A) shows the plot of the Lyapunov exponent obtained from the time series of the logistic map. Plot (C) displays the Shannon entropy estimated from the black diagonal lines of the symbolic recurrence plot. The S_{RP_S} plot detects the chaos-chaos transitions as well as the periodic-chaos and chaos-periodic transitions. The plot shows that the S_{RP_S} correlates positively with the LE rather than with its inverse. We note though that for $r \in [3.9, 4.0]$ the values of the entropy seems to slightly decrease with the chaotic behaviour of LE. Further investigation is needed to understand this behaviour, that seems to correlate with the inverse of LE for that particular region of r.

It is clear that with our method, while maintaining the computation of the entropy from the black recurrence dots of the symbolic RP_S generated from the LRS of the symbolized time series, we obtain an estimate of the diagonal lines entropy that correlates positively with the Lyapunov exponents plot, with the exception of the region of $r \in [3.9, 4.0]$.

Plot (B) illustrates the LE_S obtained from the VMO-derived LRS. Plot (D) is the K2 plot derived from the RP_S and shows that the K2 plot is a lower bound of the LE_S plot. This verifies the relation expressed in Eq. 6.5, where K2 is defined as being a lower bound on the sum of the positive LEs.

Plot (E) displays the IR values that correspond to the optimal VMO models selected for each time series we generate from the logistic map. Plot (F) portrays the value of determinism DET_S from chapter 5. By contrasting the two plots, we notice that the IR plot corresponds positively to the determinism measure. Both depict the chaos-order transitions. Additionally, IR captures the chaos-chaos transitions as well.

6.4 Analysis

Classification tasks using the complexity measures were conducted for the MAV and the FME datasets. The affective dimensions of the FME are: high valence, low valence, high energy, low energy. The classification framework is described in section 3.3 of chapter 3.

As reported in table 6.1, the classification success rates of the affective vocalizations using the complexity measures are (84%, 91%) on VA, with a precision of (72%, 61%) and a recall of (80%, 90%). Compared to the results of table 5.1 of chapter 5, the classification performance with the symbolic RQA_S features was (79%, 93%) on VA. We note that the success rate for the arousal dimension has decreased by 2%, however the prediction rate for valence increased by 5%. Considering that generally valence is recognized with less accuracy than arousal in literatures, this increase in valence prediction is important. The values of the remaining metrics are high, $\kappa > 0.40$ for both VA which shows a good agreement between observed and predicted values. AUC value is 0.91 indicates that the test accuracy is excellent.

Comparing these results with those from table 3.1c in chapter 3, we had achieved a success rate of (95.26%, 97.63%) on (valence, arousal) using acoustic features, after performing feature selection with mutual information. Although this shows the effectiveness of the acoustic measures in capturing affect, we note two advantages with the nonlinear dynamics features, be it RQA_S or the complexity features. First the decision about the number of features and their types is straightforward: since these are metrics that describe the underlying system's dynamics, no prior work is needed in order to motivate the choice of the measures for a given type of sound, as is the case with acoustics. In our case the number of features are five in total in this chapter, and a maximum of 15 RQA measures in chapter 5. Second, with the present approach, there is no need to perform feature selection which considerably alleviates the burden of selecting the appropriate algorithm, as well as the computation resources required when large databases are involved.

Table 6.1: Dynamical Invariants performance measures for MAV on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.84	0.72	0.80	0.75	0.78	0.59	0.91
Arousal	0.91	0.61	0.90	0.72	0.82	0.45	0.91

A classification was made for the IADS database as well, in order to compare the performance of the dynamical invariants with that of the symbolic RQA_S measures. In table 6.2 the accuracies are (72%, 82%) on VA, with rather high values achieved on precision, F_1 and F_2 for valence. However κ values are rather low, which indicates that the agreement is not above chance level. This should be further investigated, with a much larger dataset, to be able to conclude on the viability of the dynamical invariants with respect to auditory scenes.

Four learning tasks are conducted on the FME dataset. A first one using the complexity measures. A second baseline task using a set of acoustic features. A third task using a set of RQA_S measures as developed in chapter 5. And a final task using a hybrid set combining acoustics as well as complexity measures.

Table 6.2: Dynamical Invariants performance measures for IADS on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.72	0.91	0.71	0.79	0.74	0.14	0.64
Arousal	0.82	0.78	0.76	0.75	0.75	0.09	0.61

The classification success rates on the FME dataset using the complexity measures are described in table 6.3. We note that the accuracy is rather low, attaining (69%, 73%) for VA with a precision of (69%, 72%) and a recall of (70%, 46%). This shows that our complexity measures do not capture well the affective dimensions in the film music excerpts.

In order to apprehend why the prediction performance was rather low using the complexity measures, we tested our symbolic RQA_S measures on the FME dataset. These measures achieved high recognition rates of (90%, 90%) on VA for music clips in chapter 5. The classification results on the FME dataset are reported in table 6.4. We note the performance rates of (65%, 77%) on VA, which is also rather low, in contrast to (90%, 90%) on the music clips. This shows that the complexity measures as well as RQA_S do not capture well the affective dimensions in the film music excerpts. Further investigations are needed in this respect, to determine what is impacting the differences in the recognition rates, between the short music clips and the short film music excerpts.

Table 6.3: Dynamical invariants performance measures for FME on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.69	0.69	0.70	0.69	0.70	0.39	0.75
Arousal	0.73	0.72	0.46	0.56	0.50	0.38	0.79

Table 6.4: RQA_S performance measures for FME on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.65	0.60	0.87	0.71	0.80	0.30	0.71
Arousal	0.77	0.87	0.46	0.60	0.51	0.47	0.83

Next we compare the performance of the complexity metrics with a baseline using acoustic measures. A total of 391 acoustic measures were extracted from the film music excerpts using MIRToolbox [106]. They consist of statistics of the following main features: dynamics, fluctuation, rhythm, spectral, timbre and tonal (see Appendix). The features' statistics are referenced in the MIRToolbox manual. Statistics with missing values were excluded from the final feature set. First the classification task was conducted on the entire feature set. Then the mutual information (MI) feature selection was made. The latter method achieved higher success rates with only 25 features instead of the full set, and the results are shown in table 6.5.

Table 6.5: FME Baseline performance measures using acoustics and MI

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.90	0.90	0.90	0.90	0.90	0.80	0.94
Arousal	0.92	0.91	0.87	0.88	0.87	0.82	0.96

A last and final classification task was made using a hybrid set combining acoustic measures with the complexity features. Feature selection with MI was made, and 50 features were retained that included D2 and K2 for valence, and D2, IR and the threshold θ for arousal. Table 6.6 shows that the success rates are higher than those reported in table 6.1. Furthermore, they are also higher than the results we obtained using 50 acoustic features only, without complexity measures. Therefore fusing the acoustic feature set with complexity measures improved the overall classification performance rates. An accuracy of (92%, 93%) on VA is attained, with an almost perfect AUC of (0.95, 0.95) respectively, indicating that the test separates almost perfectly the classes. κ values are (0.84, 0.86) indicating that the agreement is considerably above chance level.

Table 6.6: FME performance measures using acoustic and complexity features

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.92	0.89	0.95	0.92	0.94	0.84	0.95
Arousal	0.93	0.93	0.90	0.91	0.90	0.86	0.95

6.5 Conclusive Remarks

In this chapter we proposed novel estimates of dynamical invariants from the symbolic $RP_{\mathbb{S}}$ of the VMO symbolized time series of audio signals. Additionally we computed these measures for the symbolized time series of the logistic map, and showed that our measures are in agreement with the same measures obtained using a different method in literature. The key aspect of our dynamical invariants is that they describe only the meaningful recurrences of a symbolized form of the signal.

When they are assessed in emotion recognition tasks, they achieved high recognition rates of (84%, 91%) on VA for affective vocalizations, but rather low rates (69%, 73%) on VA for film music excerpts. Further work is needed to explore why both sets of complexity measures, the $RQA_{\mathbb{S}}$ as well as the dynamical invariants obtained in this chapter, achieve low recognition rates on film music excerpts, compared to the results obtained with the $RQA_{\mathbb{S}}$ on the music clips in chapter 5.

Finally, when the dynamical invariants are combined with acoustic features, the ER success rates of (92%, 93%) on VA are higher than those achieved by acoustic features alone (90%, 92%),

which shows that when combined with acoustics, they contribute positively by improving the overall emotion recognition rate.

As directions for future works, ER tasks could merge nonlinear dynamics features such as RQA as well as complexity invariants with other acoustic features in order to improve affect prediction rates in audio stimuli.

6.6 Dissemination

1. Pauline Mouawad and Shlomo Dubnov. Novel Feature Extraction Method of Dynamical Invariants for Emotion Recognition. Submitted to: Special Issue on Recent Advances in Engineering Systems, Advances in Science, Technology and Engineering Systems Journal, ASTESJ, 2017.

Conclusions and Future Research

The research presented in this dissertation investigated the recognition of affective information in nonverbal vocal and musical sounds, through the implementation of a variety of tasks spanning psychology, signal analysis and chaotic dynamics.

The main contributions of this work are:

- Provided a ground truth of emotionally annotated nonverbal singing voices, using categorical descriptors of the VA model of affect.
- Demonstrated that the nonverbal singing voice conveys affective information even when there is no emotional intent in the performance. Additionally, we showed that the glottal sound of the singing voice distinctly communicates emotions as well. This revealed that the production of singing is inherently emotional, independently of lyrics, music and emotional intent.
- Proposed a framework of affect recognition across the domains of voice and music:
  - We revealed the role of vocal acoustics as strong affective predictors in capturing musical emotions.
  - Showed that musical expressiveness did not entirely originate in voice. Although there is some overlap between vocal and musical acoustics, however our tests did not substantiate the claim about the common origins of voice and music.
  - Recommended learning models and frameworks for ER in affective vocalizations and musical emotional bursts.
- Proposed a new model of ER based on nonlinear dynamics and time series symbolization. Notably, we showed that our proposed model based on the Variable Markov Oracle (VMO) circumvents the necessity to resort to the traditional nonlinear dynamics methods that rely strongly on phase space reconstruction with embedding.
- Proposed new symbolic complexity features. A key advantage of these features is that they describe the dynamics of those parts of sounds that are most expressive. We showed their effectiveness in capturing affective information in voice as well as music.
- Provided additional insight to the debate on the common origins of voice and music: cross-domain classification tasks using dynamical properties of sounds failed to create a holistic learning model for both channels.

- Proposed new symbolic dynamical invariants and showed their suitability for ER studies in sounds.

- Developed a hybrid model for ER that combines dynamical invariants as well as acoustics.

Conclusive Remarks and Future Directions

This work investigated affective recognition in voice and music from two different perspectives: acoustic-based analysis and nonlinear dynamics analysis. In general we found that both approaches are suitable for the ER tasks. Yet there are key advantages in our proposed nonlinear symbolic dynamics framework: first, the number of features is well defined and no feature selection techniques are needed; second, the dynamical properties we employed describe the meaningful repetitions that are most important in communicating information.

The fusion of acoustics with nonlinear dynamical invariants improved the overall emotion prediction rate in film music excerpts, achieving a higher performance than the one obtained with the baseline using acoustic features.

However, it was found to be difficult to obtain a holistic learning model of affect that learns emotions across the vocal and musical channels, and this was true for both perspectives. Furthermore, although the symbolic dynamical invariants we estimated in chapter 6 achieved a high recognition rate for the affective vocalizations, their recognition performance was rather low for the film music excerpts. This was also true for the symbolic RQA_S although both types of complexity measures achieved high recognition rates for the musical emotional bursts.

In neuroscience and cognitive psychology, various aspects of emotion and cognition are viewed as dissimilar processes, and are addressed independently of each other. However, emotions are first perceived in various stimuli and then appraised. In this sense, emotion is closely coupled with perception and cognition, and the three are interrelated processes. Nonlinear dynamical theory describes how emotion perception and cognition interact and proposes a dynamical model of emotion-cognition interaction, based on the competition principle among brain centers and modes. In seeking to develop a dynamical model of affect, future works could investigate to what extent the proposed dynamical features relate to cognitive processes. Furthermore, studies of emotions show that significant interaction exist between valence and well-being [80]. Future works could investigate the nonlinear dynamics of time series and their role in positive affective states such as well-being.

Future work could complete the ground truth of nonverbal singing voices, potentially annotating it with dimensional as well as categorical emotions. The stimuli can be made available for future research, as a benchmark database of nonverbal singing voices.

In endeavouring towards a consensus on a recommended feature set for ER in sounds, future works could apply our proposed nonlinear symbolic dynamical invariants in a variety of ER tasks in order to test their predictive performance in a broad variety of stimuli, such as music, speech, verbal singing, various environmental sounds, as well as animal sounds. Furthermore, it would be interesting to see how they impact the recognition rates when combined with other

acoustic measures. One recommendation is to test them in challenges such as INTERSPEECH[1] or MIREX[2].

Another avenue for future studies is to examine why the RQA_S achieved high recognition rates on the short music clips, and rather low rates on the film music excerpts. This was also true for the symbolic dynamical invariants. More work is needed to understand the differences in the dynamics of the two types of musical sounds, and how they communicate affect.

The model that served as a basis for our nonlinear dynamics features was previously applied for music segmentation [225] and synthesis [222]. Future works could employ our features to guide music improvisation and segmentation by emotion.

Similarly, future works could investigate audio event detection tasks guided by the different emotions associated with the events.

The application of chaotic dynamics to the ER field is very recent, and the fusion of nonlinear methods with acoustics is still a young research approach. Naturally, as is the case with most research it could be that more questions are raised in the present dissertation than answered. I hope this work has demonstrated the potential contributions that chaotic dynamics can bring to the field. Hopefully findings of this thesis have helped to advance our understanding of emotion perception in nonverbal vocal and musical sounds.

The unlimited resources for vocal and instrumental expression lie in artistic deviation from the pure, the true, the exact, the perfect, the rigid, the even, and the precise. This deviation from the exact is, on the whole, the medium for the creation of the beautiful - for the conveying of emotion.

—Carl Seashore

[1] http://www.interspeech2017.org/
[2] http://www.music-ir.org/mirex/wiki/MIREX_HOME

Bibliography

[1] *Music and emotion: Theory and research*, chapter Psychological perspectives on music and emotion, pages 71–104. Oxford University Press, 2001.

[2] U Rajendra Acharya, S Vinitha Sree, G Swapna, Roshan Joy Martis, and Jasjit S Suri. Automated eeg analysis of epilepsy: a review. *Knowledge-Based Systems*, 45:147–165, 2013.

[3] Cyril Allauzen, Maxime Crochemore, and Mathieu Raffinot. Factor oracle: A new structure for pattern matching. In *International Conference on Current Trends in Theory and Practice of Computer Science*, pages 295–310. Springer, 1999.

[4] Jesús B Alonso, Fernando Díaz-de María, Carlos M Travieso, and Miguel Angel Ferrer. Using nonlinear features for voice disorder detection. In *ISCA tutorial and research workshop (ITRW) on non-linear speech processing*, 2005.

[5] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[6] Gérard Assayag and Shlomo Dubnov. Using factor oracles for machine improvisation. *Soft Computing*, 8(9):604–610, 2004.

[7] Philip Ball. *The music instinct: how music works and why we can't do without it*. Random House, 2010.

[8] Rainer Banse and Klaus R Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.

[9] Mathieu Barthet, György Fazekas, and Mark Sandler. Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based model. In *Proc. CMMR*, pages 492–507, 2012.

[10] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

[11] Russell Beale and Christian Peter. The role of affect and emotion in hci. In *Affect and emotion in human-computer interaction*, pages 1–11. Springer, 2008.

[12] Pascal Belin, Sarah Fillion-Bilodeau, and Frédéric Gosselin. The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods*, 40(2):531–539, 2008.

[13] Anjali Bhatara, Petri Laukka, and Daniel J Levitin. Expression of emotion in music and vocal communication: Introduction to the research topic. *Frontiers in psychology*, 5, 2014.

[14] Stefanos Boccaletti, Celso Grebogi, Y-C Lai, H Mancini, and Diego Maza. The control of chaos: theory and applications. *Physics reports*, 329(3):103–197, 2000.

[15] Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2002.

[16] Kanu Boku, Taro Asada, Yasunari Yoshitomi, and Masayoshi Tabuse. Speech synthesis of emotions using vowel features of a speaker. *Artificial Life and Robotics*, 19(1):27–32, 2014.

[17] DL Bowling. A vocal basis for the affective character of musical mode in melody. *Frontiers in psychology*, 4:464–464, 2012.

[18] Elif Bozkurt, Engin Erzin, Cigdem Eroglu Erdem, and A Tanju Erdem. Improving automatic emotion recognition from speech signals. In *INTERSPEECH*, pages 324–327, 2009.

[19] Elizabeth Bradley and Holger Kantz. Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097610, 2015.

[20] Margaret M Bradley and Peter J Lang. The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3*, 2007.

[21] Juan Gabriel Brida et al. *Symbolic time series analysis in economics*. Universidad de la República, Facultad de Ciencias Sociales, 2000.

[22] Caitlin J Butte, Yu Zhang, Huangqiang Song, and Jack J Jiang. Perturbation and nonlinear dynamic analysis of different singing styles. *Journal of Voice*, 23(6):647–652, 2009.

[23] Julyan HE Cartwright, Diego L González, and Oreste Piro. Pitch perception: A dynamical-systems perspective. *Proceedings of the National Academy of Sciences*, 98(9):4855–4859, 2001.

[24] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 2014.

[25] Dong-Wei Chen, Na Han, Jun-Jie Chen, and Hao Guo. Novel algorithm for measuring the complexity of electroencephalographic signals in emotion recognition. *Journal of Medical Imaging and Health Informatics*, 7(1):203–210, 2017.

[26] John Ellery Clark and Colin Yallop. An introduction to phonetics and phonology. *Scientific American*, 2000.

[27] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[28] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[29] Pierre Collet and J-P Eckmann. *Iterated maps on the interval as dynamical systems*. Springer Science & Business Media, 2009.

[30] Roddy Cowie and Ellen Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1989–1992. IEEE, 1996.

[31] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[32] Trevor J Cox. Tutorial: public engagement through audio internet experiments. University of Salford, 2011.

[33] Kathleen E Cummings and Mark A Clements. Improvements to and applications of analysis of stressed speech using glottal waveforms. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 2, pages 25–28. IEEE, 1992.

[34] Kathleen E Cummings and Mark A Clements. Application of the analysis of glottal excitation of stressed speech to speaking style modification. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 207–210. IEEE, 1993.

[35] Kathleen E Cummings and Mark A Clements. Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America*, 98(1):88–98, 1995.

[36] C Stuart Daw, Charles Edward Andrew Finney, and Eugene R Tracy. A review of symbolic analysis of experimental data. *Review of Scientific Instruments*, 74(2):915–930, 2003.

[37] Reik Donner, Uwe Hinrichs, and Bernd Scholz-Reiter. Symbolic recurrence plots: A new quantitative framework for performance analysis of manufacturing networks. *The European Physical Journal Special Topics*, 164(1):85–104, 2008.

[38] J Stephen Downie. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[39] Philip J Drew and John RT Monson. Artificial neural networks. *Surgery*, 127(1):3–11, 2000.

[40] Peter D Drummond and Saw Han Quah. The effect of expressing anger on cardiovascular reactivity and facial blood flow in chinese and caucasians. *Psychophysiology*, 38(2):190–196, 2001.

[41] Shlomo Dubnov. Spectral anticipations. *Computer Music Journal*, 30(2):63–83, 2006.

[42] Shlomo Dubnov, Gerard Assayag, and Arshia Cont. Audio oracle: A new algorithm for fast learning of audio structures. In *Proceedings of International Computer Music Conference (ICMC)*. ICMA, 2007.

[43] Shlomo Dubnov, Gérard Assayag, and Arshia Cont. Audio oracle analysis of musical information rate. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 567–571. IEEE, 2011.

[44] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[45] J-P Eckmann, S Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. *EPL (Europhysics Letters)*, 4(9):973, 1987.

[46] Tuomas Eerola, Olivier Lartillot, and Petri Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *ISMIR*, pages 621–626, 2009.

[47] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[48] Paul Ekman. Are there basic emotions? 1992.

[49] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[50] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[51] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[52] Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595, 2003.

[53] Deniz Eroglu, Thomas K DM Peron, Nobert Marwan, Francisco A Rodrigues, Luciano da F Costa, Michael Sebek, István Z Kiss, and Jürgen Kurths. Entropy of weighted recurrence plots. *Physical Review E*, 90(4):042919, 2014.

[54] Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009.

[55] Florian Eyben, Gláucia L Salomão, Johan Sundberg, Klaus R Scherer, and Björn W Schuller. Emotion in the singing voice—a deeperlook at acoustic features in the light of automatic classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–9, 2015.

[56] Paul R Farnsworth. A study of the hevner adjective list. *The Journal of Aesthetics and Art Criticism*, 13(1):97–103, 1954.

[57] Mehrdad Fatourechi, Rabab K Ward, Steven G Mason, Jane Huggins, Alois Schlögl, and Gary E Birch. Comparison of evaluation metrics in classification applications with imbalanced datasets. In *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications-Volume 00*, pages 777–782. IEEE Computer Society, 2008.

[58] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[59] Michale S Fee, Boris Shraiman, Bijan Pesaran, and Partha P Mitra. The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. *Nature*, 395(6697):67–71, 1998.

[60] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[61] Johnny JR Fontaine, Klaus R Scherer, and Cristina Soriano. *Components of emotional meaning: A sourcebook*. OUP Oxford, 2013.

[62] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[63] Jose Fornari and Tuomas Eerola. Predicting emotional prosody of music with high-level acoustic features.

[64] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959, 2009.

[65] Marylou Pausewang Gelfer and Dawn M Fendel. Comparisons of jitter, shimmer, and signal-to-noise ratio from directly digitized versus taped voice samples. *Journal of Voice*, 9(4):378–382, 1995.

[66] David Gerhard et al. Audio visualization in phase space. In *Bridges: Mathematical Connections in Art, Music and Science*, pages 137–144, 1999.

[67] Vincent Gibiat and Michèle Castellengo. Period doubling occurrences in wind instruments musical performance. *Acta Acustica United with Acustica*, 86(4):746–754, 2000.

[68] Janine Giese-Davis and David Spiegel. Emotional expression and cancer progression. 2003.

[69] Ioulia Grichkovtsova, Michel Morel, and Anne Lacheret. The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3):414–429, 2012.

[70] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE, 2011.

[71] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE, 2011.

[72] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[73] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[74] Patricia Henríquez, Jesús B Alonso, Miguel A Ferrer, Carlos M Travieso, Juan I Godino-Llorente, and Fernando Díaz-de María. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. *IEEE transactions on audio, speech, and language processing*, 17(6):1186–1195, 2009.

[75] Patricia Henríquez, Jesús B Alonso, Miguel A Ferrer, Carlos M Travieso, and Juan R Orozco-Arroyave. Application of nonlinear dynamics characterization to emotional speech. In *International Conference on Nonlinear Speech Processing*, pages 127–136. Springer, 2011.

[76] Patricia Henríquez, Jesús B Alonso, Miguel A Ferrer, Carlos M Travieso, and Juan R Orozco-Arroyave. Nonlinear dynamics characterization of emotional speech. *Neurocomputing*, 132:126–135, 2014.

[77] Hanspeter Herzel. Bifurcations and chaos in voice signals. *Appl. Mech. Rev*, 46(7):399–413, 1993.

[78] Hanspeter Herzel, David Berry, Ingo R Titze, and Marwa Saleh. Analysis of vocal disorders with methods from nonlinear dynamics. *Journal of Speech, Language, and Hearing Research*, 37(5):1008–1019, 1994.

[79] Kate Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268, 1936.

[80] Marlies Houben, Wim Van Den Noortgate, and Peter Kuppens. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis., 2015.

[81] Hao Hu, Ming-Xing Xu, and Wei Wu. Fusion of global statistical and segmental spectral features for speech emotion recognition. In *INTERSPEECH*, pages 2269–2272, 2007.

[82] Xiao Hu and J Stephen Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *ISMIR*, pages 67–72, 2007.

[83] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[84] Arefin Huq, Juan Pablo Bello, Andy Sarroff, Jeff Berger, and Robert Rowe. Sourcetone: An automated music emotion recognition system. In *Proceedings of the International Conference on Music Information Retrieval*, 2009.

[85] Carlos Toshinori Ishi and Nick Campbell. Analysis of acoustic-prosodic features of spontaneous expressive speech. *Revista de Estudos da Linguagem*, 12(2):37–49, 2004.

[86] Joseph S Iwanski and Elizabeth Bradleya. Recurrence plots of experimental data: To embed or not to embed? *Chaos*, 8(4):861, 1998.

[87] Susan Jansens, Gerrit Bloothooft, and Guus de Krom. Perception and acoustics of emotions in singing. In *Fifth European Conference on Speech Communication and Technology*, volume 4, pages 2155–2158, 1997.

[88] Jack J Jiang, Yu Zhang, and Clancy McGilligan. Chaos in voice, from modeling to measurement. *Journal of Voice*, 20(1):2–17, 2006.

[89] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[90] Aastha Joshi and Rajneet Kaur. A study of speech emotion recognition methods. *Int. J. Comput. Sci. Mob. Comput.(IJCSMC)*, 2(4):28–31, 2013.

[91] Dipti D Joshi and MB Zalte. Speech emotion recognition: A review. *Journal of Electronics and Communication Engineering (IOSR-JECE)*, 4(4), 2013.

[92] Patrik N Juslin. Music and emotion: Seven questions, seven answers. *Music and the mind: Essays in honour of John Sloboda*, pages 113–135, 2011.

[93] Patrik N Juslin and Petri Laukka. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, 1(4):381, 2001.

[94] Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770, 2003.

[95] Patrik N Juslin and Petri Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004.

[96] Patrik N Juslin and Daniel Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(05):559–575, 2008.

[97] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

[98] Hamid Karimi Rouzbahani and Mohammad Reza Daliri. Diagnosis of parkinson's disease in human using voice signals. *Basic and Clinical Neuroscience*, 2(3):12–20, 2011.

[99] Matthew B Kennel, Reggie Brown, and Henry DI Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403, 1992.

[100] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[101] Vladimir J Konečni. Does music induce emotion? a theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts*, 2(2):115, 2008.

[102] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[103] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. Citeseer.

[104] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[105] P Lang and Margaret M Bradley. The international affective picture system (iaps) in the study of emotion and attention. *Handbook of emotion elicitation and assessment*, 29, 2007.

[106] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[107] Petri Laukka. Categorical perception of emotion in vocal expression. *Annals of the New York Academy of Sciences*, 1000(1):283–287, 2003.

[108] Anne-Maria Laukkanen, Erkki Vilkman, Paavo Alku, and Hanna Oksanen. On the perception of emotions in speech: the role of voice quality. *Logopedics Phoniatrics Vocology*, 22(4):157–168, 1997.

[109] Cyril Laurier, Owen Meyers, Joan Serrà, Martin Blech, Perfecto Herrera, and Xavier Serra. Indexing music by mood: design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, 48(1):161–184, 2010.

[110] Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. Music mood representations from social tags. In *ISMIR*, pages 381–386, 2009.

[111] Richard S Lazarus. Emotions and adaptation: Conceptual and empirical relations. In *Nebraska symposium on motivation*. University of Nebraska Press, 1968.

[112] Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303, 2005.

[113] Chul Min Lee, Shrikanth S Narayanan, and Roberto Pieraccini. Classifying emotions in human-machine spoken dialogs. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 737–740. IEEE, 2002.

[114] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition based on phoneme classes. In *Interspeech*, pages 205–211, 2004.

[115] Sungyoung Lee, Young-Tack Park, Brian J d'Auriol, et al. A novel feature selection method based on normalized mutual information. *Applied Intelligence*, 37(1):100–120, 2012.

[116] Arnaud Lefebvre and Thierry Lecroq. Compror: on-line lossless data compression with a factor oracle. *Information Processing Letters*, 83(1):1–6, 2002.

[117] Christophe Letellier. Estimating the shannon entropy: recurrence plots versus symbolic dynamics. *Physical review letters*, 96(25):254102, 2006.

[118] Mark Levy and Mark Sandler. A semantic space for music derived from social tags. *Austrian Compuer Society*, 1:12, 2007.

[119] Xi Li, Jidong Tao, Michael T Johnson, Joseph Soltis, Anne Savage, Kirsten M Leong, and John D Newman. Stress and emotion classification using jitter and shimmer features. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1081. IEEE, 2007.

[120] Chang-Hung Lin, Yuan-Shan Lee, Ming-Yen Chen, and Jia-Ching Wang. Automatic singing evaluating system based on acoustic features and rhythm. In *Orange Technologies (ICOT), 2014 IEEE International Conference on*, pages 165–168. IEEE, 2014.

[121] C Liu, M White, and G Newell. Measuring the accuracy of species distribution models: a review. In *Proceedings 18th World IMACs/MODSIM Congress. Cairns, Australia*, pages 4241–4247, 2009.

[122] Huawen Liu, Jigui Sun, Lei Liu, and Huijie Zhang. Feature selection with dynamic mutual information. *Pattern Recognition*, 42(7):1330–1339, 2009.

[123] Angela Lombardi, Pietro Guccione, and Cataldo Guaragnella. Exploring recurrence properties of vowels for analysis of emotions in speech. *Sensors & Transducers*, 204(9):45, 2016.

[124] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.

[125] Edward N Lorenz. Section of planetary sciences: The predictability of hydrodynamic flow*. *Transactions of the New York Academy of Sciences*, 25(4 Series II):409–432, 1963.

[126] Psyche Loui, Justin Bachorik, Hui C Li, and Gottfried Schlaug. Effects of voice on emotional arousal. *Frontiers in Psychology*, 4:675, 2013.

[127] Iker Luengo, Eva Navas, and Inmaculada Hernáez. Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 12(6):490–501, 2010.

[128] Marko Lugger and Bin Yang. The relevance of voice quality features in speaker independent emotion recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–17. IEEE, 2007.

[129] Marko Lugger and Bin Yang. *Psychological motivated multi-stage emotion classification exploiting voice quality features*. INTECH Open Access Publisher, 2008.

[130] Donna S Lundy, Soham Roy, Roy R Casiano, Jun W Xue, and Joseph Evans. Acoustic analysis of the singing and speaking voice in singing students. *Journal of Voice*, 14(4):490–493, 2000.

[131] Christian Maganza, René Caussé, and Franck Laloë. Bifurcations, period doublings and chaos in clarinetlike systems. *EPL (Europhysics Letters)*, 1(6):295, 1986.

[132] Antonio Maratea, Alfredo Petrosino, and Mario Manzo. Adjusted f-measure and kernel scaling for imbalanced data learning. *Information Sciences*, 257:331–341, 2014.

[133] TK March, SC Chapman, and RO Dendy. Recurrence plot statistics and the effect of embedding. *Physica D: Nonlinear Phenomena*, 200(1):171–184, 2005.

[134] Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5):237–329, 2007.

[135] Norbert Marwan, Niels Wessel, Udo Meyerfeldt, Alexander Schirdewan, and Jürgen Kurths. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. *PHYSICAL REVIEW E Phys Rev E*, 66:026702, 2002.

[136] Brian Massumi. *Parables for the virtual: Movement, affect, sensation.* Duke University Press, 2002.

[137] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[138] Werner Mende, Hanspeter Herzel, and Kathleen Wermke. Bifurcations and chaos in newborn infant cries. *Physics Letters A*, 145(8-9):418–424, 1990.

[139] David Meyer, Friedrich Leisch, and Kurt Hornik. Benchmarking support vector machines. 2002.

[140] Leonard B Meyer. *Emotion and meaning in music.* University of chicago Press, 2008.

[141] Thomas J Millhouse and Frantz Clermont. Perceptual characterisation of the singer's formant region: a preliminary study. In *Proceedings of the Eleventh Australian International Conference on Speech Science and Technology*, pages 253–258, 2006.

[142] Luca Mion and Giovanni De Poli. Score-independent audio features for description of music expression. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):458–466, 2008.

[143] Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111, 2014.

[144] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[145] Jürgen Neubauer, Michael Edgerton, and Hanspeter Herzel. Nonlinear phenomena in contemporary vocal music. *Journal of Voice*, 18(1):1–12, 2004.

[146] Maria Nordenberg and Johan Sundberg. Effect on ltas of vocal loudness variation. *Logopedics Phoniatrics Vocology*, 29(4):183–191, 2004.

[147] Koichi Omori, Ashutosh Kacker, Linda M Carroll, William D Riley, and Stanley M Blaugrund. Singing power ratio: quantitative evaluation of singing voice quality. *Journal of Voice*, 10(3):228–235, 1996.

[148] Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. Geometry from a time series. *Physical review letters*, 45(9):712, 1980.

[149] S Paquette, I Peretz, and P Belin. The "musical emotional bursts": a validated set of musical affect bursts to investigate auditory affective processing. *Frontiers in Psychology*, 4(509):1–7, 2013.

[150] Sona Patel, Klaus R Scherer, Eva Björkner, and Johan Sundberg. Mapping emotions into acoustic space: The role of voice production. *Biological psychology*, 87(1):93–98, 2011.

[151] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[152] Paolo Petta, Catherine Pelachaud, and Roddy Cowie. Emotion-oriented systems. *The Humaine Handbook, ISBN*, pages 978–3, 2011.

[153] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[154] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[155] Lena Rachel Quinto, William Forde Thompson, and Felicity Louise Keating. Emotional communication in speech and music: The role of melodic and rhythmic contrasts. *Frontiers in Psychology*, 4:184, 2013.

[156] H Rabarimanantsoa, L Achour, C Letellier, A Cuvelier, and J-F Muir. Recurrence plots and shannon entropy for a dynamical analysis of asynchronisms in noninvasive mechanical ventilation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(1):013115, 2007.

[157] Lawrence R Rabiner and Ronald W Schafer. *Digital processing of speech signals*. Prentice Hall, 1978.

[158] Lucy Rainey. *The clarinet as extension of the voice and expressive conduit of musical styles in diverse ensembles: a thesis submitted for [ie to] the Victoria University of Wellington in fulfilment of the requirements for the degree of Master of Musicology, 2011: New Zealand School of Music, Wellington, New Zealand*. PhD thesis, Massey University, 2011.

[159] Srinivasan Ramakrishnan. Recognition of emotion from speech: A review. In *Speech Enhancement, Modeling and Recognition-Algorithms and Applications*. InTech, 2012.

[160] K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada. Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2):143–160, 2013.

[161] S Neil Rasband. *Chaotic dynamics of nonlinear systems*. Courier Dover Publications, 2015.

[162] AF Rasmussen. Emotions and immunity. *Annals of the New York Academy of Sciences*, 164(1):458–461, 1969.

[163] Byron Reeves and Clifford Nass. How people treat computers, television, and new media like real people and places. *CSLI Publications and Cambridge*, 1996.

[164] JD Reiss and MB Sandler. Nonlinear time series analysis of musical signals. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, pages 1–5, 2003.

[165] Fabien Ringeval and Mohamed Chetouani. Exploiting a vowel based approach for acted emotion recognition. In *Verbal and nonverbal features of human-human and human-machine interaction*, pages 243–254. Springer, 2008.

[166] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[167] Patricia Henríquez Rodríguez, Jesús B Alonso Hernández, Miguel A Ferrer Ballester, Carlos M Travieso González, and Juan R Orozco-Arroyave. Global selection of features for nonlinear dynamics characterization of emotional speech. *Cognitive Computation*, 5(4):517–525, 2013.

[168] Jerome Rolink, Martin Kutz, Pedro Fonseca, Xi Long, Berno Misgeld, and Steffen Leonhardt. Recurrence quantification analysis across sleep stages. *Biomedical Signal Processing and Control*, 20:107–116, 2015.

[169] Gerard Roma, Waldo Nogueira, and Perfecto Herrera. Recurrence quantification analysis features for environmental sound recognition. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

[170] Gerard Roma, Waldo Nogueira, Perfecto Herrera, and Roc de Boronat. Recurrence quantification analysis features for auditory scene classification. *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events, Tech. Rep*, 2013.

[171] Rui Rui and Changchun Bao. Musical instrument classification based on nonlinear recurrence analysis and supervised learning. *Radioengineering*, 22(1):61, 2013.

[172] JA Russel. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[173] Gláucia Laís Salomão, Johan Sundberg, and KR Scherer. Emotional coloring of the singing voice. In *PAN EUROPEAN VOICE CONFERENCE ABSTRACT BOOK*, page 80, 2015.

[174] Disa A Sauter, Frank Eisner, Andrew J Calder, and Sophie K Scott. Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology*, 63(11):2251–2272, 2010.

[175] Disa A Sauter, Nicole M McDonald, Devon N Gangi, Daniel S Messinger, et al. Nonverbal expressions of positive emotions. *Handbook of positive emotions*, pages 179–198, 2014.

[176] Klaus R Scherer. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[177] Klaus R Scherer. Affect bursts. *Emotions: Essays on emotion theory*, 161:196, 1994.

[178] Klaus R Scherer. Expression of emotion in voice and music. *Journal of Voice*, 9(3):235–248, 1995.

[179] Klaus R Scherer. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351, 2009.

[180] Klaus R Scherer, Rainer Banse, and Harald G Wallbott. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural psychology*, 32(1):76–92, 2001.

[181] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[182] Klaus R Scherer, Johan Sundberg, Lucas Tamarit, and Gláucia L Salomão. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*, 29(1):218–235, 2015.

[183] Ulrich Schimmack and Alexander Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345, 2000.

[184] Erik M Schmidt, Matthew Prockup, Jeffery Scott, Brian Dolhansky, Brandon G Morton, and Youngmoo E Kim. Relating perceptual and feature space invariances in music emotion recognition. *CMMR, London, UK*, 2012.

[185] Thomas Schreiber. Interdisciplinary application of nonlinear time series methods. *Physics reports*, 308(1):1–64, 1999.

[186] Marc Schröder. Experimental study of affect bursts. *Speech communication*, 40(1):99–116, 2003.

[187] Emery Schubert. Update of the hevner adjective checklist. *Perceptual and motor skills*, 96(3_suppl):1117–1122, 2003.

[188] Bjorn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Emotion recognition from speech: putting asr in the loop. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4585–4588. IEEE, 2009.

[189] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.

[190] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

[191] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010.

[192] Björn W Schuller, Ronald Müller, Manfred K Lang, and Gerhard Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *INTERSPEECH*, pages 805–808, 2005.

[193] Ashish Sen and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012.

[194] Joan Serrà, A Carlos, and Ralph G Andrzejak. Nonlinear audio recurrence analysis with application to genre classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. Citeseer, 2011.

[195] Xavier Serra, Ralph G Andrzejak, et al. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, 2009.

[196] Yong-tao Sha, Chang-chun Bao, Mao-shen Jia, and Xin Liu. High frequency reconstruction of audio signal based on chaotic prediction theory. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 381–384. IEEE, 2010.

[197] Ali Shahzadi, Alireza Ahmadyfard, Ali Harimi, and Khashayar Yaghmaie. Speech emotion recognition using nonlinear dynamics features. *Turkish Journal of Electrical Engineering & Computer Sciences*, 23(Sup. 1):2056–2073, 2015.

[198] Cosma Rohilla Shalizi. Methods and techniques of complex systems science: An overview. In *Complex systems science in biomedicine*, pages 33–114. Springer, 2006.

[199] Mohammad T Shami and Mohamed S Kamel. Segment-based approach to the recognition of emotions in speech. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 4–pp. IEEE, 2005.

[200] Eric Shouse. *Feeling, emotion, affect*. publisher not identified, 2005.

[201] Emiliana R Simon-Thomas, Dacher J Keltner, Disa Sauter, Lara Sinicropi-Yao, and Anna Abramson. The voice conveys specific emotions: evidence from vocal burst displays. *Emotion*, 9(6):838, 2009.

[202] Herbert Spencer. The origin and function of music. *Essays, Scientific, Political, and Speculative*, 2, 1857.

[203] Johan Sundberg. The acoustics of the singing voice. *Scientific American*, 236(3):82–86, 1977.

[204] Johan Sundberg, Jenny Iwarsson, and Håkan Hagegård. A singer's expression of emotions in sung performance. In *Vocal fold physiology*. Singular Publishing Group, Inc., 1995.

[205] Johan Sundberg, Sona Patel, Eva Bjorkner, and Klaus R Scherer. Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing*, 2(3):162–174, 2011.

[206] Johan Sundberg and Thomas D Rossing. The science of singing voice. *the Journal of the Acoustical Society of America*, 87(1):462–463, 1990.

[207] Ana Tajadura-Jiménez and Daniel Västfjäll. Auditory-induced emotion: A neglected channel for communication in human-computer interaction. In *Affect and Emotion in Human-Computer Interaction*, pages 63–74. Springer, 2008.

[208] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[209] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014.

[210] R Core Team. ISBN 3-900051-07-0, 2014.

[211] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[212] Ingo R Titze. *Principles of voice production*. National Center for Voice and Speech, 2000.

[213] Laurel J Trainor. The origins of music in auditory scene analysis and the roles of evolution and culture in musical creation. *Phil. Trans. R. Soc. B*, 370(1664):20140089, 2015.

[214] LL Trulla, A Giuliani, JP Zbilut, and CL Webber. Recurrence quantification analysis of the logistic equation with transients. *Physics Letters A*, 223(4):255–260, 1996.

[215] Athanasios Tsanas, Max A Little, Patrick E McSharry, Jennifer Spielman, and Lorraine O Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271, 2012.

[216] Gaetano Valenza, Antonio Lanata, and Enzo Pasquale Scilingo. The role of nonlinear dynamics in affective valence and arousal recognition. *IEEE transactions on affective computing*, 3(2):237–249, 2012.

[217] Naresh N Vempala and Frank A Russo. Exploring cognitivist and emotivist positions of musical emotion using neural network models.

[218] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1500–1503. IEEE, 2005.

[219] Dimitrios Ververidis, Constantine Kotropoulos, and Ioannis Pitas. Automatic emotional speech classification. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–593. IEEE, 2004.

[220] Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In *International Conference on Affective Computing and Intelligent Interaction*, pages 139–147. Springer, 2007.

[221] Thurid Vogt, Elisabeth André, and Johannes Wagner. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In *Affect and emotion in human-computer interaction*, pages 75–91. Springer, 2008.

[222] Cheng-i Wang and Shlomo Dubnov. Guided music synthesis with variable markov oracle. In *The 3rd International Workshop on Musical Metacreation, 10th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

[223] Cheng-i Wang and Shlomo Dubnov. Pattern discovery from audio recordings by variable markov oracle: A music information dynamics approach. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 683–687. IEEE, 2015.

[224] Cheng-i Wang and Shlomo Dubnov. The variable markov oracle: Algorithms for human gesture applications. *IEEE MultiMedia*, 22(4):52–67, 2015.

[225] Cheng-i Wang and Gautham J Mysore. Structural segmentation with the variable markov oracle and boundary adjustment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 291–295. IEEE, 2016.

[226] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from eeg data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.

[227] Ying Wang, Shoufu Du, and Yongzhao Zhan. Adaptive and optimal classification of speech emotion recognition. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 5, pages 407–411. IEEE, 2008.

[228] Christopher Watts, Kathryn Barnes-Burroughs, Julie Estis, and Debra Blanton. The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. *Journal of voice*, 20(1):82–88, 2006.

[229] Charles L Webber, Norbert Marwan, Angelo Facchini, and Alessandro Giuliani. Simpler methods do it better: success of recurrence quantification analysis as a general purpose data analysis tool. *Physics Letters A*, 373(41):3753–3756, 2009.

[230] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer. On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology*, 4:292, 2013.

[231] SP Whiteside. Note on voice and perturbation measures in simulated vocal emotions. *Perceptual and motor skills*, 88(3_suppl):1219–1222, 1999.

[232] Inka Wilden, Hanspeter Herzel, Gustav Peters, and Günter Tembrock. Subharmonics, biphonation, and deterministic chaos in mammal vocalization. *Bioacoustics*, 9(3):171–196, 1998.

[233] Teresa D Wilson and Douglas H Keefe. Characterizing the clarinet tone: Measurements of lyapunov exponents, correlation dimension, and unsteadiness. *The Journal of the Acoustical Society of America*, 104(1):550–561, 1998.

[234] Yi-Hsuan Yang and Homer H Chen. *Music emotion recognition*. CRC Press, 2011.

[235] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso. An acoustic study of emotions expressed in speech. In *INTERSPEECH*, 2004.

[236] Mingyu You, Chun Chen, Jiajun Bu, Jia Liu, and Jianhua Tao. Emotion recognition from noisy speech. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1653–1656. IEEE, 2006.

[237] Mingyu You, Chun Chen, Jiajun Bu, Jia Liu, and Jianhua Tao. A hierarchical framework for speech emotion recognition. In *Industrial Electronics, 2006 IEEE International Symposium on*, volume 1, pages 515–519. IEEE, 2006.

[238] Edumund Kim Youngmoo. *Singing voice analysis/synthesis*. PhD thesis, Program in Media Arts and Sciences, School of Architecture and Planning, MIT, Boston, 2003.

[239] Feng Yu, Eric Chang, Ying-Qing Xu, and Heung-Yeung Shum. Emotion detection from speech to enrich multimedia content. *Advances in multimedia information processing—PCM 2001*, pages 550–557, 2001.

[240] Noor Aina Zaidan and Md Sah Hj Salam. A review on speech emotion features. *Jurnal Teknologi*, 75(2), 2015.

[241] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.