SINGING VOICE EXTRACTION FROM STEREOPHONIC RECORDINGS

Stratis Sofianos

A thesis submitted in partial fulfilment of the requirements of the University of Hertfordshire for the degree of Doctor of Philosophy

The programme of research was carried out in the Faculty of Science, Technology, and Creative Arts, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, United Kingdom

September 2012

«... Πάντων γάρ ὅσα πλείω μέρη ἔχει καὶ μὴ ἔστιν οἷον σωρὸς τὸ πᾶν ἀλλ' ἔστι τι τὸ ὅλον παρὰ τὰ μόρια...»

Αριστοτέλης, «Μετά τά Φυσικά», Βιβλίο Η΄ 1045α 9-10

'... The totality is not, as it were, a mere heap, but the whole is greater than the sum of its parts...'

Aristotle, "Metaphysics", Book H 1045a 9-10

ABSTRACT

Singing voice separation (SVS) can be defined as the process of extracting the vocal element from a given song recording. The impetus for research in this area is mainly that of facilitating certain important applications of music information retrieval (MIR) such as lyrics recognition, singer identification, and melody extraction.

To date, the research in the field of SVS has been relatively limited, and mainly focused on the extraction of vocals from monophonic sources. The general approach in this scenario has been one of considering SVS as a blind source separation (BSS) problem. Given the inherent diversity of music, such an approach is motivated by the quest for a generic solution. However, it does not allow the exploitation of prior information, regarding the way in which commercial music is produced.

To this end, investigations are conducted into effective methods for unsupervised separation of singing voice from stereophonic studio recordings. The work involves extensive literature review of existing methods that relate to SVS, as well as commercial approaches. Following the identification of shortcomings of the conventional methods, two novel approaches are developed for the purpose of SVS. These approaches, termed SEMANICS and SEMANTICS draw their motivation from statistical as well as spectral properties of the target signal and focus on the separation of voice in the frequency domain. In addition, a third method, named Hybrid SEMANTICS, is introduced that addresses time-, as well as frequency-domain separation.

As there is lack of a concrete standardised music database that includes a large number of songs, a dataset is created using conventional stereophonic mixing methods. Using this database, and based on widely adopted objective metrics, the effectiveness of the proposed methods has been evaluated through thorough experimental investigations.

To Dimitris Bamparis

ACKNOWLEDGEMENTS

They say that the vessel that carries us through a journey comprises the people we are connected with… Finally, this journey has now come to an end, and I feel that the least I can do is to mention the people who constituted my "vessel", for—without them—this project would not have been possible.

Firstly, I would like to express my deep and sincere gratitude to my principal supervisor, Prof. Aladdin Ariyaeeinia. His expertise and remarkable attention to detail have been invaluable to me. His understanding, encouraging, and personal guidance have transformed my project into not only an enjoyment, but also a life experience. Thanks are also due to my second supervisor, Dr. Richard Polfreman, for his useful comments and assistance.

Foremost, I am grateful to Mr. Dimitris Bamparis, to whom I dedicate this thesis, for his selfless and boundless trust he has shown me and without whom I would not have been able to pursuit the dreams of my adolescence.

I wholeheartedly thank my friend, and partner for the best part of this project, Miss Laura Callaghan. Her relentless support and dedication provided the horsepower to cut through challenging obstacles of this voyage.

I want to thank my parents for their support and unconditional love.

Thanks are also due to Dr. Maria Chrimatopoulou for kindly lending her elegant handwriting for the epigraph, for keeping my motivation in check throughout this endeavour, as well as for being a true friend and a valuable support throughout my life.

Last but not least, I want to thank my colleagues Dr. Miloš Milosavljević, Mr. Sat Juttla, Dr. Surosh Pillay, and Mrs. Zoe Jeffrey not only for the useful academic discussions, but also for their help that frequently went beyond the call of duty.

However, when one finally reaches his Ithaca, it is not only important to thank people who have had a positive contribution to his journey, but also the Laestrygonians and the Cyclops, the Scyllae and the Charybdes, that substantiated his expedition; alas, every exciting story requires a Nemesis, and for that I am thankful to Joy.

LIST OF ABBREVIATIONS

AD	**a**mplitude **d**iscrimination
ADRess	**a**zimuth **d**iscrimination and **re**-**s**ynthesi**s**
AIR	**a**udio **i**nformation **r**etrieval
ASA	**a**uditory **s**cene **a**nalysis
BSS	**b**lind **s**ource **s**eparation
BWV	**B**ach-**W**erke-**V**erzeichnis
CASA	**c**omputational **a**uditory **s**cene **a**nalysis
CLT	**c**entral **l**imit **t**heorem
CPP	**c**ocktail **p**arty **p**roblem
DCT	**d**iscrete **c**osine **t**ransform
DET	**d**etection **e**rror **t**rade-off
DNA	**d**irect **n**ote **a**ccess
EER	**e**qual **e**rror **r**ate
ENIC	**e**nergy and **n**on-vocal **i**ndependent **c**omponent correlation
ERB	**e**quivalent **r**ectangular **b**andwidth
EVD	**e**igen**v**alue **d**ecomposition
FFT	**f**ast **F**ourier **t**ransform
FICA	**f**ast **i**ndependent **c**omponent **a**nalysis
FIR	**f**inite **i**mpulse **r**esponse
H- SEMANTICS	**h**ybrid **s**inging **e**xtraction through **m**ultiband **a**mplitude e**n**hanced **t**hresholding & **i**ndependent **c**omponent **s**ubtraction
IBM	**i**deal **b**inary **m**ask
ICA	**i**ndependent **c**omponent **a**nalysis
IFFT	**i**nverse **f**ast **F**ourier **t**ransform
IID	**i**nterchannel **i**ntensity **d**ifference
IIR	**i**nfinite **i**mpulse **r**esponse
IPD	**i**nterchannel **p**hase **d**ifference
IR	**i**mpulse **r**esponse
ISA	**i**ndependent **s**ubspace **a**nalysis
ISM	**i**deal **s**oft **m**ask
ISMIR	**i**nternational **s**ymposium for **m**usic **i**nformation **r**etrieval
ISO	**i**nternational **s**tandardisation **o**rganisation

ISR	source **i**mage to **s**patial distortion **r**atio
ISTFT	**i**nverse **s**hort-**t**erm **F**ourier **t**ransform
ITD	**i**nterchannel **t**ime **d**ifference
KV	**K**öchel-**V**erzeichnis
LC	**l**ocal **c**riterion
MDL	**m**usic **d**igital **l**ibrary
MEG	**m**agneto**e**ncephalo**g**raphy
MFCC	**m**el-**f**requency **c**epstral **c**oefficient
MICED	**m**el-frequency cepstral coefficient of **i**ndependent **c**omponent **E**uclidean **d**istance
MIDI	**m**usical **i**nstrument **d**igital **i**nterface
MIR	**m**usic **i**nformation **r**etrieval
MP3	**MP**EG-2 Audio Layer **III**
NIC	**n**on-vocal **i**ndependent **c**omponent
PC	**p**rincipal **c**omponent
PCA	**p**rincipal **c**omponent **a**nalysis
PCM	**p**ulse **c**ode **m**odulation
PDF	**p**robability **d**ensity **f**unction
PMCC	**P**earson **m**oment **c**orrelation **c**oefficient
RMS	**r**oot **m**ean **s**quare
SAR	**s**ource to **a**rtefacts **r**atio
SDR	**s**ource to **d**istortion **r**atio
SEMANICS	**s**inging **e**xtraction through **m**odified **A**DRess and **n**on-vocal **i**ndependent **c**omponent **s**ubtraction
SEMANTICS	**s**inging **e**xtraction through **m**ultiband **a**mplitude e**n**hanced **t**hresholding & **i**ndependent **c**omponent **s**ubtraction
SIR	**s**ource to **i**nterference **r**atio
SNR	**s**ignal to **n**oise **r**atio
SPL	**s**ound **p**ressure **l**evel
STFT	**s**hort-**t**erm **F**ourier **t**ransform
SVD	**s**ingular **v**alue **d**ecomposition
SVS	**s**inging **v**oice **s**eparation
TF	**t**ime-**f**requency
TREC	**t**ext **re**trieval **c**onference
VIC	**v**ocal **i**ndependent **c**omponent

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Singing voice separation

The human auditory system is able to perform a significant number of complex tasks, given only two input streams (from the left and the right ears). These tasks include identifying the nature of a source that is present in the stream (e.g. if a source is speech, musical instrument, or noise), its position and pitch in relation to other sources and, in the case of a speech source, the words spoken. More formally, humans are able to derive a semantic understanding of audio and they are able to perform these tasks with streams that contain multiple time- and frequency-overlapping sources, even when the interfering energy is close to or exceeds the energy of the target source. The human ability to focus on a specific source from within a mixture is known as *auditory scene analysis* (ASA) [1].

Conversely, machines are not yet able to fully separate multi-sourced/polyphonic audio streams of which music is a particularly challenging example. Although today's music recording and production is largely carried out using computers, the processing of information in these recordings often

requires manual intervention. This is because, in contrast with humans, machines cannot yet provide the full capability required for recognising the genre of a music piece, its harmonic structure, the lyrics, or the identity of the singer.

The main obstacle in this respect appears to be the lack of an automated process to "focus" on individual sources of a polyphonic stream [2], in the way humans can. As a result, the research in this field has been largely concerned with the separation of individual sources in a given music mixture [3, 4]. A major facet of the effort in this respect is the extraction of human voice (singing), which is arguably the most information-rich content of music [5]. The specific field of research that addresses this area of separation/extraction is termed *singing voice separation* (SVS).

1.2 Motivation for singing voice separation

In the pre-digital computing era, musicians and enthusiasts relied on thematic catalogues. These catalogues allowed positive identification of music pieces while using a minimum of space and symbols [6]. They contained representative fragments, i.e. *incipits*, of scores that usually depicted the beginning of a score, but sometimes they represented the principal melody or *theme* of a musical work [7]. Thematic catalogues were usually produced by scholars that concentrated on the works of a particular composer. The *Köchel-Verzeichnis* (KV) [8] and the *Bach-Werke-Verzeichnis* (BWV) [9] are two popular examples of such catalogues that refer to the works of Wolfgang Amadeus Mozart and Johann Sebastian Bach respectively.

This method of organising music information has become unsustainable as well as inefficient, due to the inability of Western music notation to accurately represent the contemporary literature of music, the rapid change of musical forms (e.g. frequent absence of thematic structure), the emergence of

electronically synthesised innovative sounds, and the sheer growth of music track production. In addition, the exuberance of music that is available online demonstrates the need for an automated and robust system that can aid users to identify, verify, and locate music pieces.

The intuitive solution to the above is computer-assisted *music information retrieval* (MIR). Although this idea has been addressed as early as in 1966 [10], the truth remains that the field is still in its infancy [2]. One of the main reasons is that important applications of MIR such as song identification [11], singer identification [12], melody extraction, lyrics recognition [13], and lyrics alignment [14] require the vocal element alone and hence the effective separation of this from the accompanying music [15, 16]. This is supported by [17, 18] where the authors have found limited success when they tried to extract information from a music track without prior separation of sources. MIR can be broadly separated in two categories (Figure 1.1): the symbolic, which is based on retrieval based on symbolic representation (e.g. music notation and musical instrument digital interface, a.k.a. MIDI information [19]), and retrieval that draws its information from the audio signal of the piece, i.e. audio information retrieval (AIR) [20]. The symbolic route, however, exhibits the aforementioned shortcomings, while AIR is more appropriate to the current digital era, where music is frequently produced without the use of a score. As can be seen in the figure, particularly for the case of AIR, source separation can play a catalytic role in the workflow of MIR.

Evidently, tackling this "bottleneck" in MIR clears the pathway for content-based multimedia search. Currently, Internet search engines (Google, Yahoo, Ask, etc.), that have significantly facilitated Internet use, can only use text-based information that their algorithms/spiders can crawl through. When targeting multimedia, such engines rely solely on metadata [21], which are entered manually and are, thus, prone to human error and subjectivity.

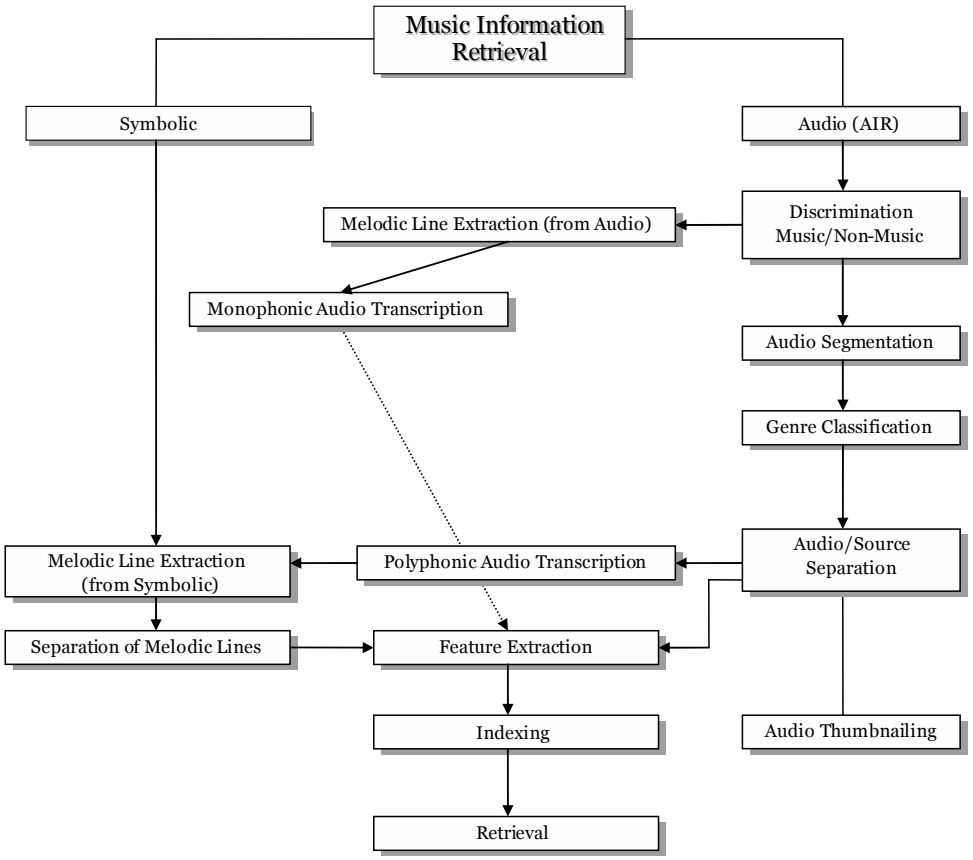

Figure 1.1: Schematic representation of the workflow of MIR

A successful approach to SVS would eliminate such restrictions and help applications that can expand and promote the multimedia domain of the Internet. In general, this can prove an enormous aid to musicians, ethnomusicologists, intellectual property solicitors, librarians, sound engineers, and, last but not least, music fans.

1.3 General approach

The isolation of a source from its surrounding environment is considered analogous to a person that is standing in an arbitrary position of a fairly crowded room and is faced with the challenge of focusing on a particular speaker. The audio content of such a room may include music, noise, as well as

people's conversations. This problem was initially described by Helmholtz in 1863 as the *Ball-room problem* [22] but is nowadays rather known by Cherry's description in 1953 as the *cocktail party problem* (CPP) [23]. This case is closely related to the research field of *blind source separation* (BSS) [24], where the term "blind" implies that only weak assumptions can be made about the nature of the sources.

In the case of SVS, the "room ambience" [25] of the Cocktail Party is represented by an observed mixture of musical sources (i.e. the music accompaniment of a song), and the challenge is the extraction of the singing voice. Generally, the observed mixture can be produced in a physical or simulated space (i.e. live or studio recording), and can comprise one, two, or multiple channels (mono, stereo, etc.). There are two essential phases in the process of SVS: the segregation of vocal from non-vocal segments in the time domain [26], and the separation of vocal from music accompaniment in the frequency domain.

Overall, the vocal vs. non-vocal segregation (also named singer's voice detection [27]), resembles the common approach in speaker change detection [28]: initially, the signal is divided into segments, and compact representations of these segments are obtained (i.e. features). The segments (i.e. the features of the segments) are tested against each other, and are classified into two classes (vocal and non-vocal).

In contrast, frequency-domain approach poses a far more complex challenge, and there are many pathways that have been explored over the recent years. The usual method in this scenario is one of considering SVS as a BSS problem [29, 30]. Given the inherent diversity of music, such an approach is motivated by the quest for a generic solution. For this purpose, the methods proposed by the community draw their inspiration from disciplines across a plethora of fields, such as neuroscience [31], cognitive psychology [32], and psychoacoustics [33].

However, the common denominator in all of these endeavours is that researchers try to simplify the challenge by assuming some sort of prior knowledge about the nature of the observed mixtures. In other words, the initial step is to make the case less "blind". As expected, different assumptions dictate different approaches. Nevertheless, it should be noted that one of the contributing factors to the complexity of this challenge is that these assumptions can only be minor due to the heterogeneity of music. Notwithstanding the above, common hypotheses that have been made throughout the literature derive from the production process of recordings in commercial music.

1.4 Challenges

Intuitively, the solution to the challenge of SVS is to imitate the cognitive functions of the brain. Indeed, this seems to be one of the most intriguing barriers in the path to a generic solution, as there is little understanding of how exactly the human brain perceives and processes audio signals in general, and in particular music.

The vocal vs. non-vocal discrimination in the time domain (mentioned in Section 1.3) has been the subject of extensive research and is producing promising results [26, 27]. On the other hand, the separation in frequency domain has not been found as effective. This is because it does not entail the use of such effective analysis means as those deployed in the time-domain methods (feature extraction – modelling – classification). As mentioned before, the diversity in music songs is vast: the singer can be any human being and the number and type of instruments can vary from a symphonic orchestra, to just a single synthesised sound accompaniment. Assumptions cannot even be made for a set of possible music instruments. On the contrary, originality of sound is a highly sought-after concept, which encourages the producers /composers to try "inventing" new sounds—synthesised or not [34]. Thus, a

generic "music accompaniment" model cannot be created. Even the distinction between music, noise, and silence has nowadays become a blurry one [35-37]. In addition, the separation in the frequency domain poses a particularly complex challenge, as the singing voice and the music not only overlap, but they are also strongly correlated. This is because, more often than not, music accompanies the singing voice harmonically. As a result, the two exhibit simultaneous presence in the same fundamental frequencies and partials.

A less theoretical and more practical issue of this research field is the lack of a concrete standardised music database that includes a large number of songs, so that researchers can have a common reference point for benchmarking the efficacy of their respective systems. This is a significant challenge for this area, as there is no reliable basis for pinpointing the state-of-the-art technology for SVS. The main hindrance in creating such a database is the copyright legislation that exists in most countries and the considerable financial resources that would be needed to obtain clearance for a large and comprehensive database.

In recent years there have been a few attempts to create such a database. The studies in [38, 39] involved the hiring of musicians in order to produce a specifically research-orientated music information retrieval (MIR) database (i.e. the *RWC Database).* Although this database tackles the copyright problem, the number of music tracks is very small (200 in total, with only 20 of them songs). In [40], the authors describe the construction of a dataset with recordings that are available in the public domain. This database comprises 1886 songs but only of 10-second duration segments and in mp3 quality. Another effort was made in 2004, when the international symposium for music information retrieval (ISMIR) distributed a dataset for a melody pitch extraction contest [41]. This dataset contains 20 excerpts of music, 10 of which are songs. This dataset has been used only by a small number of researchers [42], mainly because of the limited number of diverse songs. In addition, the problem for using any of these databases for evaluation of an SVS system is the

lack of the *clean*[1] tracks of voice and music accompaniment that would serve as ground truth [43]. As a result, none of these databases is widely used and accepted as a reference point and researchers tend to resort to the use of customised databases (e.g. [44, 45]).

There are suggestions in the MIR/MDL (music digital library) evaluation project regarding the creation of a database for MIR [46], similar to TREC (text retrieval conference) that is used for speech recognition [47]. However, the project seems to have limited activity since 2003 [48].

1.5 Aims and scope of the project

The aim of this research is to develop an effective unsupervised algorithm for isolating[2] the singing voice from a given stereophonic mixture of music accompaniment and singing voice (i.e. a song). The scope of the project is limited to stereophonic (i.e. 2 observations/channels) studio recordings where the singing part is performed by a solo human voice. This is believed to be the most common approach to song production over the last fifty years. However, the scope is not limited to a specific genre or number/type of accompanying instruments.

Given the above aim, the work involves a systematic literature survey into blind source separation in order to establish the state-of-the-art methods that can be useful for SVS. This is envisaged to provide in-depth knowledge of the advantages and disadvantages of the various methods and a clearer aspect of previous work in the field.

As discussed previously, the separation in frequency domain is deemed as the

[1] The term clean refers to a sound signal that contains neither artefacts nor noise. In the case of SVS, this is the clean vocal that is termed *a capella*.
[2] As mentioned here, the aim of this work is the singing voice extraction. However, since most literature refers to this challenge as singing voice separation (SVS), the latter term is preferred in this thesis.

most challenging part of the field. Therefore, a main aspect of this study is to develop methods for the extraction of the singing part from its concurrent music accompaniment. This is in addition to investigating methods for addressing the time-domain segregation issue. The ultimate purpose in this regard is the successful classification of music-only segments and singing voice segments (time domain) from any given song from the literature of popular music.

For evaluation purposes and experimental investigations, the work includes the development of a suitable database for SVS with a variety of samples, ranging from simple/ideal cases to real-world scenarios. This facet of the project will also include the review and selection of meaningful evaluation systems, so that the quality of separation can be measured objectively.

1.6 Organisation of the thesis

The seven subsequent chapters that complete this thesis are organised as follows:

Chapter 2: Literature review

This chapter investigates the anatomy of the singing voice, its role in music and the differences that exist between speech and singing. The existing commercial approaches that exist in SVS are discussed. The chapter also gives a general overview of the fields of *auditory scene analysis* (ASA) and *computational auditory scene analysis* (CASA).

Chapter 3: Blind source separation and independent component analysis

In this chapter, two approaches that attempt source separation based on statistical properties of the signal are described. These belong to the family of blind source separation (BSS) algorithms and are the principal component analysis (PCA) and the independent component analysis (ICA). This chapter also focuses on the differences of the aforementioned methods that are frequently vague in the literature.

Chapter 4: Applied source extraction from polyphonic mixtures

The process of stereophonic music production is described here together with an approach that is important in the context of this thesis. This approach, named azimuth discrimination and re-synthesis (ADRess), has the advantage that exploits properties of the signal that are specific to stereophonic production.

Chapter 5: Singing extraction through modified ADRess and non-vocal independent component subtraction

This chapter proposes a novel approach for the purpose of stereophonic SVS, which combines properties of the ADRess method with ICA. In addition, this chapter includes a thorough description of the dataset of songs that is used for the purpose of evaluation in the current study. The chapter concludes with the experimental investigation and a comparative evaluation of SEMANICS and ADRess.

Chapter 6: Singing extraction through multiband amplitude enhanced thresholding and independent component subtraction

Here, an extension of SEMANICS is introduced and investigated. The main facet of this approach is the exclusion of ADRess from the method introduced in Chapter 4. The merits of SEMANTICS over its predecessor are presented through a comparative analysis.

Chapter 7: Hybrid SEMANTICS

This chapter proposes an alternative integrated approach to SVS. It involves combining novel time-domain segregation procedures with a modified version of the frequency-domain voice isolation techniques used in SEMANICS and SEMANTICS. The performance of the complete system (with each of the considered music pruning methods) is analysed based on a set of experimental investigations.

Chapter 8: Summary, conclusions, and future work

The final chapter provides a summary of the work, and suggests a plethora of ways in which the project can advance.

2
LITERATURE REVIEW

In this chapter, the focus is on human voice with particular emphasis on the singing voice. The anatomy of this special case of instrument is presented together with the critical role that it has in music. In addition, the differences of singing and speaking are detailed, and the existing commercial approaches that attempt to separate voice from music are discussed. The chapter also gives an overview of the field of auditory scene analysis (ASA) and its computational counterpart, which is aptly named computational ASA (i.e. CASA).

2.1 The singing voice

The human voice, while capable of generating an incredibly diverse range of sounds, can be simply described in mechanical terms: It is a machine that consists of a power supply (lungs), an oscillator (vocal folds), and a resonator

(larynx, pharynx, and mouth) [49]. The capabilities of the human voice, which include the production of tones as well as noise, are very similar to those of musical instruments and indeed the human voice has been referred to as the first musical instrument [50, 51].

Formally, the sounds that are within the aptitude of the human voice production are classified into three broad categories: the voiced sounds, the unvoiced sounds, and the mixed sounds. However, in the context of music, the voiced sounds are the most important in this study, and that is where the weight of this section lies.

The voiced sounds are produced by the process of phonation, which finds its most common interpretation on the basis of the *myoelastic aerodynamic principle* [52, 53]: the power supply (lungs) is controlled by the diaphragm, resulting in expansion and contraction. This air reservoir has the ability to maintain pressure above atmospheric levels. The release of stored air pressure in the lungs, chopped by the vocal folds (which act as the oscillator), creates the sound [54]. In practice, the folds are adducted (i.e. constricted) in response to nerve impulses transmitted from the brain to the muscles of the larynx and in turn this provides the necessary condition for vibration [55].

The last stage in this procedure is the movement of the air from the lungs and restricted area of the trachea and subglottic space through the glottis to a bigger space which causes a sudden pressure drop. This loss of pressure takes place exactly at the level of the vocal folds which are consequently drawn together, according to a theory that is known in fluid dynamics as the *Bernoulli principle* [56]. The outcome is a complex tone, which is the initial source for voiced sounds in singing, as well as in speech. For this reason, it has been named the *voice source* [49]. At that point, this sound is analogous to the buzzing sound of a trumpet player when their lips are separated from the trumpet [57].

In order to produce the familiar vocal articulations, the voice source enters the

vocal tract (the resonator). The extensive polymorphism of the vocal tract (i.e. the laryngeal cavity, the pharynx, the oral cavity, and the nasal cavity) [58], which is unique to human beings, renders the human voice one of the most versatile sound production machines among the global fauna. This is because of the dependence on the form of an enclosed space and the acoustic properties that are attributed to this space [55]. An illustration of this three-part mechanism is provided in Figure 2.1.

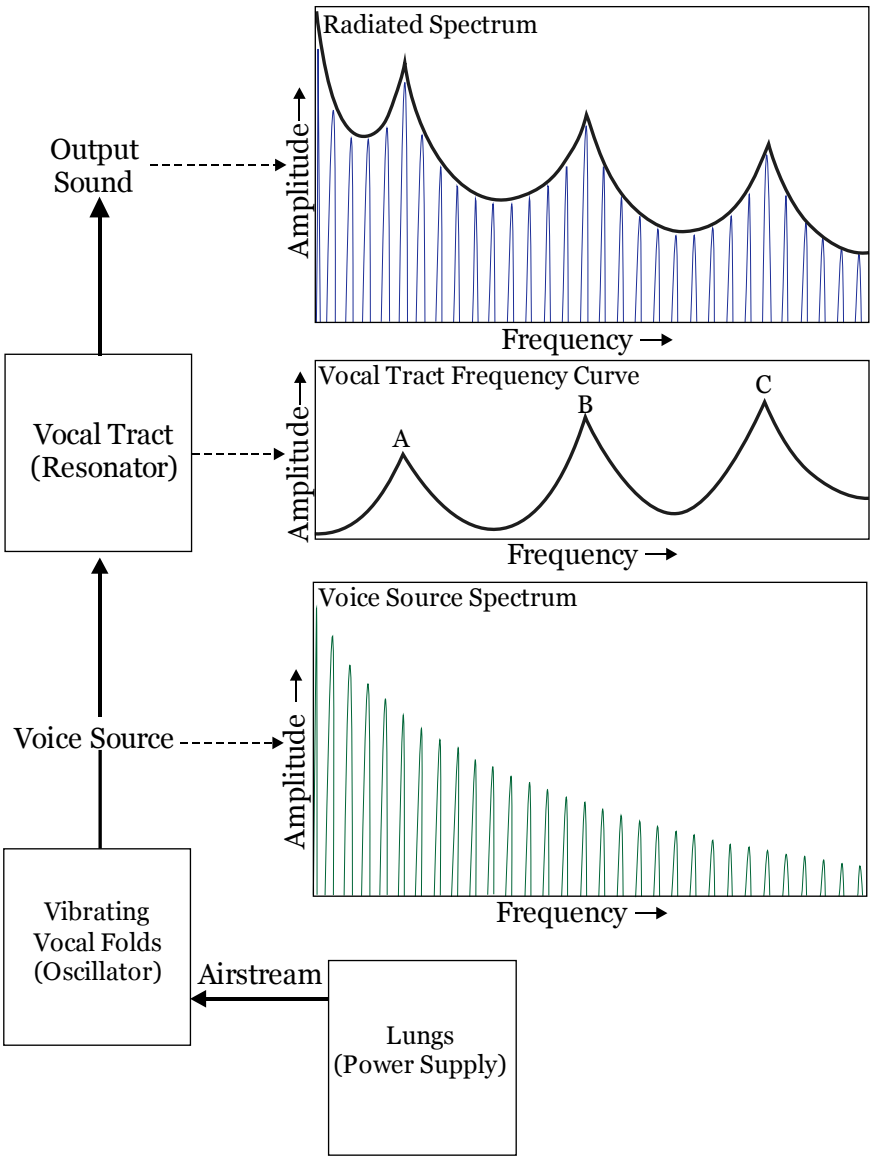

Figure 2.1: Illustration of the voice production mechanism [49]

While vocal folds are the primary factor for one's voice quality (i.e. timbre), the ability to form distinctive sounds (in singing as well as in speech) is accredited to the vocal tract. Therefore, it is important to examine how this unique instrument works.

The vocal tract measures ca. 16.9 cm in males and 14.1 cm in females [59]. Much like all reed instruments, it is a tube open at one end (the mouth) and closed at the other (the vocal folds). Assuming a simplified model of a perfect half-open acoustic tube, the first four partials are illustrated in Figure 2.2. The velocity of sound (i.e. v) is 343.2 m/s at 20 °C.

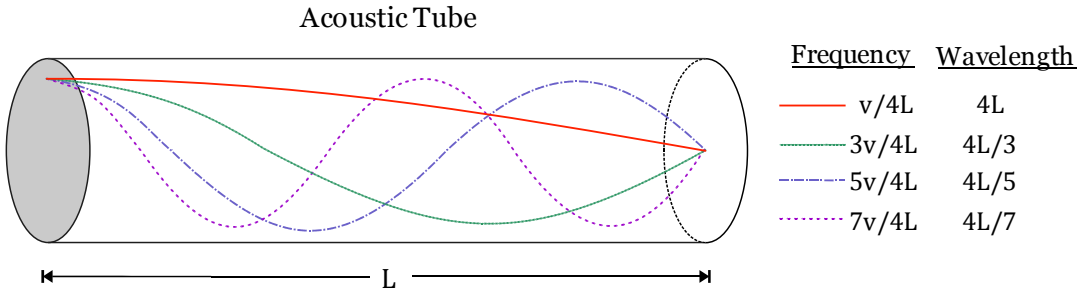

Figure 2.2: The first four partials of a half-open tube

The right end of the tube that can be seen in the figure is the closed end, where all the partials are of zero amplitude as the air has nil volumetric flow rate. The partials will reach maximum amplitude at the opposite end of the tube, which is open. As this is a half-open tube, the partials comprise only the odd harmonics (odd-quarters law). In reality, however, the vocal tract is much more complex and impure.

As its physical characteristics can be changed, the vocal tract can be seen as a dynamic acoustic triple[3]-component filter, which results in resonances at specific frequencies. This is how the vowels that exist in languages throughout the world attain their distinctive acoustic signature. These resonances are

[3] The three parameters are the positions and forms of the tongue, jaw, and lips.

generally known as *formants*. An example of the power spectrum when a soprano sings the vocal 'ah' can be seen in Figure 2.3 where the y-axis represents the ratio of the power spectrum with the mouth open vs. that with the mouth closed.

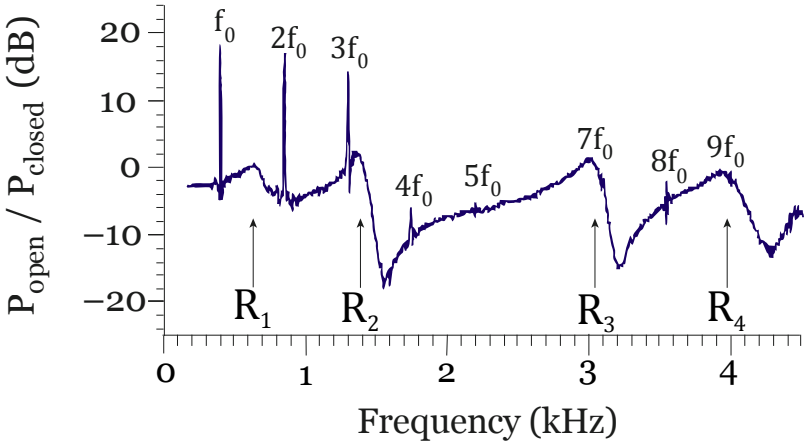

Figure 2.3: Vowel 'Ah'[ɑː] sung at A_4 (i.e. 440 Hz) [60]

In this figure, it is observed that the fundamental frequency f_0 is usually lower than the first resonance R_1. The harmonics of f_0 that exist close to the resonances R_i produce associated peaks (i.e. *formants*). In Western languages vowels are mainly characterised by the positions of R_1 and R_2 [60].

In contrast, for unvoiced sounds (i.e. most consonants) the vocal folds remain in an open position while the increased air supply by the lungs is constricted by the tongue, the soft palate, the teeth, the lips, or a combination thereof. Therefore, the unvoiced sounds are a product of air turbulence. Finally, there are also the mixed sounds (e.g. [v] and [z]) where a combination of turbulence and phonation is required [61].

To summarise, the process of singing can be seen as a source-filter process where the excitation source and the filter are independent (or at least assumed independent) of each other [62]. The source is produced by the release of pressurised air from the lungs that vibrates the vocal folds in a periodic way

(phonation). The vocal tract acts as the filter and simplistically can be modelled as a half open tube. The vocal tract is also responsible for the characteristic resonances, otherwise known as formants of the voiced sounds. The unvoiced and mixed sounds require a different approach where the vocal folds do not vibrate and the sound is produced through constriction of the air (turbulence).

2.2 The role of voice in music

In the field of human science, the suggestion that the first utterance of a human was sung is quite common. Contrary to the belief that "…in the beginning there was the Word…" [63], it is thought that the first humans tried to mimic natural sounds such as the singing of birds [64]. Although this might seem an idea hard to fathom, one should consider that—at the dawn of humanity—information communicated through the human voice was primarily conveyed by means of pitch alteration [65][4]. The idea of language did not exist, and therefore the different sounds that the early human could produce (and perceive) exhibited severe qualitative and quantitative limitations. This is because the muscles involved in speech were underdeveloped due to lack of training, and also due to the evolution-driven descent of the larynx that had not yet taken place [66]. As a result, "men sang out their feelings long before they were able to speak their thoughts" [67].

This kind of "singing", however, should not be confused or likened with the concept of modern singing in a studio or a concert hall. The character of these primal utterances was purely exclamatory and stripped from intellect and sense. Little did the first human understand, that he was laying down the bricks for the foundation of the sole universal language (i.e. music). It is not known at which point the humans actually realised that this process could

[4] In modern humans, this way of communication is common by infants and their quasi-siren cries that, depending on the pitch span, express pain, irritability, tiredness, or even joy.

transcend semiotics, as well as convey ideas and—most importantly—past events; however, the realisation of this notion marked a milestone in the foundation of language, and preceded any embryos of further musical manifestations [68].

Since these early days, the process of singing has developed—as far as is known—globally in each and every culture, suggesting that its inception is inherently linked to human nature, as well as nurture. In parallel, and as music polyphony evolved, people started to accompany the singing voice with various types of musical instruments. Therefore, music became a catalyst of the effect and impact of the singing voice, which rightly starred as the leading musical instrument [69].

In more recent times, the word 'song' commonly refers to a more conventional and concise concept: it denotes the singing of lyrics that is (usually) accompanied by music and lasts for a short period of time (typically a few minutes). In particular, when members of the general public refer to a *song*, they usually mean an actual *recording* of the song. Indeed, the popularity of this manifestation of art finds its ally in the recording technology, which has played a defining role in the distribution and the appraisal of the singing voice discipline over the recent decades. In fact, the distribution of music has found its zenith with the development of digital recording technology and the World Wide Web; without a doubt, there is now a greater amount of music (including popular songs) being distributed than before due to informal channels and ease of access to musical recordings.It is not surprising that singing has maintained or even gained popularity throughout the centuries, while forming an integral part of music culture. Even from a technical point of view, the human ear has the ability to perceive and analyse the frequency span of the human voice much more efficiently than the rest of the audible spectrum [70]. The perceived loudness is also elevated (Figure 2.4) at the frequency range that carries the main body of voice, known in communications as the voice frequency or voice band, i.e. (0.3 to 3.4) kHz [71].

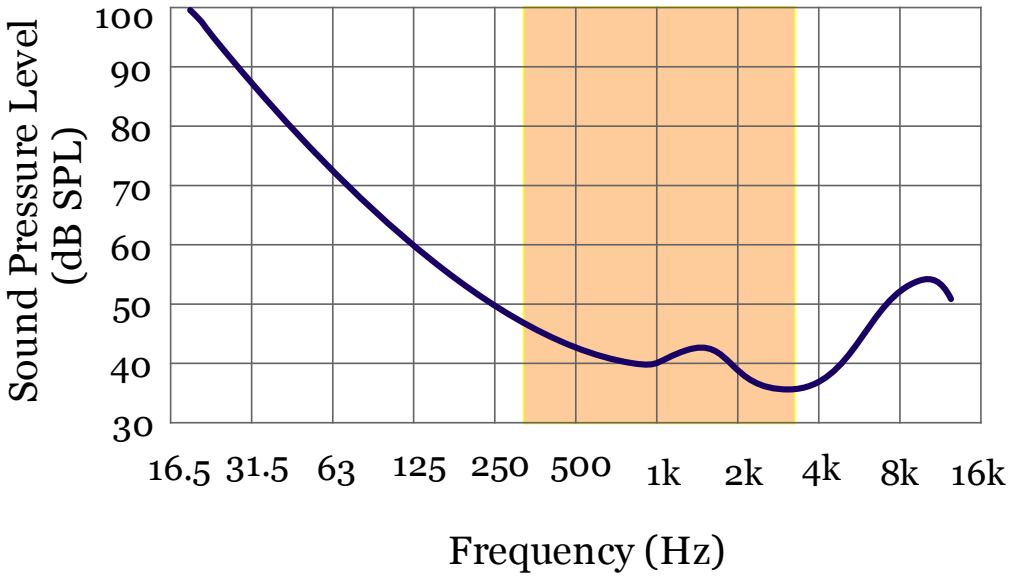

Figure 2.4: Equal loudness contour for 40 phons [72]

As a result, it is expected that—in ordinary circumstances—a listener will mainly focus on the vocal part of a music track that contains voice (i.e. a song[5]). Adding to the latter, the vocal instrument has the ability to carry and communicate a language and, in that sense, it is superior to the other instruments, as they are restricted to pitch and timbre variation. Indeed, it is because of this reason that the singing style, the voice timbre [73], and the lyrics of a sung part not only give essential information about the piece of music, but also define it.

To demonstrate this, one should consider that just the vocal component of a song is enough for a listener to recognise the title, singer, lyrics, and—arguably—genre of a song. Furthermore, in various musical cultures and in particular within popular music, the singing voice usually carries one of the most significant features of a song: the *melody* [74-76].

[5] Hereon, the term song will be used to describe any music track that contains a vocal part, even if it is not sung, for the sake of brevity.

2.3 Existing commercial approaches related to SVS

Due to the aforementioned integral role that voice has within songs, researchers and companies have endeavoured to eliminate, isolate, or enhance the vocal element of a song. A common use for elimination is the *karaoke* [77][6] that was invented in the 1970s [78]: a form of entertainment, where amateur singers perform over a pre-recorded accompaniment. This accompaniment is, on most occasions, the original published song, processed with a simple vocal elimination technique. This technique, more often than not, involves filtering as well as inverting one channel of the stereo track before summing both channels [79]. This processing cancels out the voice, which is typically recorded in mono and panned centrally in the final stereo mix. However, this technique is quite superficial and ineffective, as it also cancels out some important instruments that are panned centrally (e.g. bass, kick drum) and leaves artefacts [80, 81].

Sometimes, music producers are keen to create a new version – arrangement of a pre-existing song by taking the isolated voice track and producing new music to accompany it. There are two ways to approach this: the producers have either the means to access the multi-track session of the song (the individual tracks that the song is comprised of), or they apply band-pass filtering. The latter approach ensures that all music-only frequency bands are attenuated. Finally, the producers make sure that the newly composed/produced material masks[7] the traces of music that have eluded the former process [82].

[6] *Karaoke* is a portmanteau of the Japanese words *kara* and *ōkesutora,* meaning an "empty orchestra".

[7] "Masking" in Acoustics is the psychoacoustic phenomenon where the human perception of a sound is affected by another sound. Usually, these sounds have considerable amount of energy in similar frequency range.

A more sophisticated approach to voice isolation is the feature *Center Channel Extractor* in *Adobe Audition* [83]. While it produces superior results compared to the aforementioned method of karaoke, still fails to separate instruments that co-exist in the central space of the stereo field. In addition, the user has to set the frequency range that wants to be isolated and the size of the FFT window.

A different category to the aforementioned voice isolation is voice enhancement. This includes the effort of sound engineers to make the vocal part of a song "stand out", i.e. be more clear and intelligible by the listener. Usually, though, engineers do have possession of the individual tracks so they can conveniently boost specific frequencies of, or apply dynamic range compression to the vocal track so that it is more "present" in the final mixture. The relevance of this category to the present study will be shown later in Chapter 5.

Another aspect of voice enhancement is pitch correction. The best known commercial systems for this purpose are the software packages *Antares Autotune* and *Celemony Melodyne* [84, 85]. Traditionally, both these industry-standard commercial applications worked with monophonic audio streams and drew their algorithms from the extensive research of monophonic melody transcription [86]. Though far from perfect, these packages give satisfactory and reliable results when the target stream is monophonic, i.e. comprises pitches that do not co-exist in the same temporal space.

Recently, however, the company *Celemony* developed a feature of their application *Melodyne* termed *direct note access* (DNA) which can operate on specific monophonic lines of a polyphonic stream, but it does not allow the isolation of the vocal part from polyphonic audio [87]. The algorithm used in the aforementioned feature is not available to the public as the patent is pending [88]. Similarly, *AudioScanner* [89, 90] is based on a technique that the authors term "human-assisted time-frequency masking" and targets a sole

audio component (i.e. one instrument) from a polyphonic recording in order to manipulate it independently of the rest. For example, in a duet with cello and voice, this application aims to apply a low-pass filter to the voice without disturbing the frequency spectrum of the cello [68]. In a similar manner to DNA, AudioScanner does not extract an individual line of a polyphonic stream, but rather focuses on processing it without affecting the rest of the polyphonic material.

Although not directly related to the aforementioned categories, *Shazam* [91] is noteworthy; not only because it is arguably the most used commercial application in the field of music information retrieval (MIR), but also due to the robustness of its algorithm. In particular, Shazam is able to retrieve the identity of a song (or music track in general), given an excerpt that is captured through a microphone of a mobile phone device. The application functions by extracting a constellation of points from the spectrogram of the tested material and matching it to a member of a precompiled database. This process is termed *combinatorial hashing* by its creators. The significant feature is its outstanding robustness against severely degraded and contaminated testing material [92].

Finally, the (commercial) software that is closest to the aim of the present study appeared towards the end of 2011, and was released under the name *Hit'n'Mix* [93]. This application claims to be capable of unsupervised isolation and classification of musical instruments (e.g. guitar, voice, bass, drums, etc.). In practice, however, the algorithm results in a ca. 50% mismatch with regards to instrument classification while the degradation and contamination of the isolated parts that result from this process renders them not fit for purpose outside the context of the original song [94].

As seen in this section, the growing interest of the industry towards isolation/separation of music sources is manifested through the commercial eagerness of companies to release software as soon as possible—and therefore

gain a market share—even by compromising the reliability and robustness of the various software packages. Also, with the exception of Hit n' Mix, the rest of the aforementioned processes that relate to SVS can only operate with user supervision, which is deemed ineffective towards a generic solution.

At first, this lack of robustness seems rather odd, as the field of content-based information extraction from speech, e.g. speech recognition [95, 96] is well developed and mature. In order to understand the gap between these two fields, the differences between speech and singing are described in the next section.

2.4 Singing as opposed to speaking

One might think that singing and speaking audio signals are very similar as they are both produced by the human vocal tract. However, the reality is quite different: an opera singer is able to cut through the sonic force of a full orchestra while singing, whereas speech can never achieve the same result. This begs the question, what is so different?

Studies by Sundberg [97] indicated that the quality of voice is somehow "darker" in singing, which can be likened to speaking and yawning concurrently. During singing, the larynx moves towards a lower position and the lowest parts of the pharynx and of the laryngeal ventricle are expanded. This physiological change has been given the name *covering* [98]. The formant between (2 to 3) kHz of this singing technique has been described by Sundberg as the *singing formant*.

It is because of this formant that a tenor, for example, can be heard over the orchestra by an opera audience (Figure 2.5). Singers employ covering much more in operatic, rather than popular singing. However, in almost all kinds of singing, there are three fundamental differences when compared with speech.

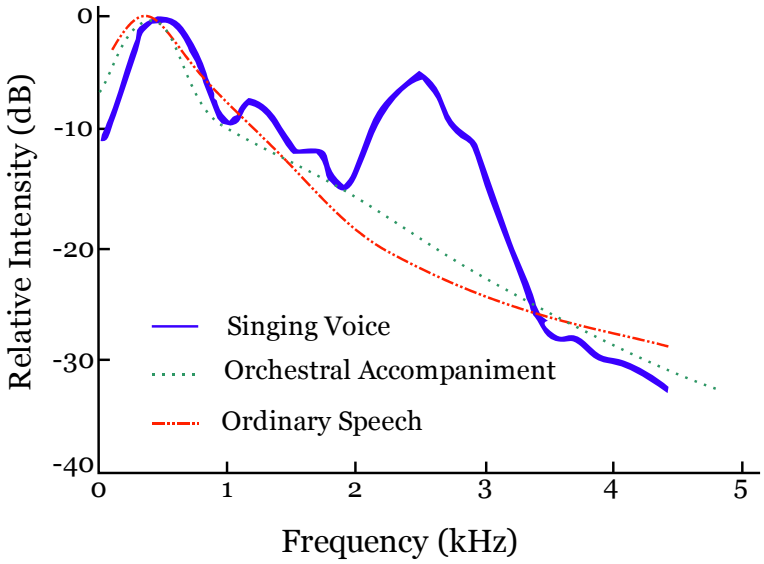

Figure 2.5: Difference between speech, orchestral music, and singing voice [49]

One of the most important differences to consider is the acoustic variation and the ratio of vowels to consonants [99, 100]. In particular, the actual time that is spent on each of these two categories. While the time expended on vowels comprises 60% in a typical English speech utterance [101], it approaches 90% in a common sung phrase of the same language [102]. The latter is easily explained, as vowels are inherently more powerful and they also have the ability to carry the melody within a song. Therefore, singers are taught to sustain vowels over consonants in order to be audible and promote the melodic structure. This phenomenon can be clearly observed in the singing style known as *bel canto*[8] [103, 104].

The second difference is that the dynamic range in singing is broader and reaches the physiological limits of the human voice organ, contrary to the—typically—narrow dynamic range of speech [105]. Speech needs to be produced at a comfortable level that makes the words intelligible to the

[8] Bel canto ("beautiful singing") broadly refers to the style of singing that was developed primarily in Italy during the 17th-19th century in order to account for the increasing density, (and therefore loudness) of the accompanying orchestra.

listener while an additional dynamic range is reserved for conveying (or implying) emotions. In addition, from an interpersonal sociological aspect of many modern societies, surpassing a certain dynamic threshold is not only rare, but also considered rude. On the other hand, singing as an art does not carry these social limitations and conventions. As a result, spontaneous speech usually produces signals up to 84 dB SPL, while a trained singer can reach an astounding sound pressure level of 114 dB [106, 107].

Lastly, the fundamental frequency (f_0) of sung material has a range from 80 Hz to 1400 Hz and exhibits rapid alterations, whereas in speech the typical range is up to 500 Hz and presents lesser temporal variation [108].

For the above reasons, speech separation systems [109, 110] are quite incompetent when applied to singing voice material. In addition to that, the "interference" that usually exists in speech analysis is uncorrelated noise, while in a song it is usually harmonically correlated music accompaniment, presenting, thus, a further challenge to any automated system.

2.5 Auditory scene analysis (ASA)

Over the past years, there has been much investigation in the area that can be broadly described as *auditory scene analysis* (ASA) [33, 111]. Briefly, ASA is a proposed model of the psychophysical procedure, during which the human auditory system receives as an input a combination of audio that is produced within a physical environment, and with the aid of cognitive[9] functions [112] classifies it into acoustic objects that are perceptually meaningful.

Certainly, the most in-depth analysis in this area has been carried out by Bregman [1], where he asked the fundamental question of how do humans

[9] Although in the context of ASA, the human auditory/aural system refers not only to the sensory system but also to the cognitive functions that comprise it, classical psychophysics does not recognise the cognitive validity of the acoustical objects that are modelled in ASA.

classify the complex—in time and frequency domains—mixtures into autonomous acoustical objects. In general, the study of ASA could be considered as a two-fold scheme: the problem of integration, and the process of segregation.

The problem of integration in auditory perception is identified on the basis of two models that describe two principal problems: *simultaneous integration* (or *perceptual fusion*), and *sequential integration*. A paradigm of perceptual fusion takes place when humans are faced with the task of identifying the separate streams of a singer and an instrument performing at the same time in a mixture. On the other hand, sequential integration could be likened to the challenge of cognitively pinpointing the temporal boundaries of words within a time-continuous stream of speech [113].

The process of segregation describes the function of ASA on the decomposition of sounds as a model deriving from the human brain. Two main classifications are regarded for this task, namely *schema-based* or *memory-based* segregation and *primitive segregation*. Primitive segregation is an innate, automatic, and obligatory [114] process, during which streams are parsed according to the similarities of local acoustical cues, such as frequency, timing, or amplitude [115]. By contrast, the memory-based segregation derives from the cognitive function of recognising prior knowledge schemata, i.e. the understanding of segregation that the listener has, based on past experience [116].

Segregation, according to ASA, is governed by rules that dictate the art and the reason of the cognitive formation of acoustic objects out of longer acoustic streams. These principles, clearly inspired from *Gestalt* psychology[10] [117, 118], could be better described as "bonds" that are developed between the stimuli which are perceived and processed not only by the aural, but also by

[10] "Gestalten" means shape in German. The philosophy of the Gestalt psychology is that "the whole is more than the sum of its parts". Gestalt psychology plays a major role in psychoacoustics.

the rest of the senses[11]. Generally, they apply to audible sounds (i.e. speech, music, and noise). Taking into consideration that interpretations in the literature exhibit slight variations, a concise summary is given below [1, 32, 119, 120]*:*

Proximity: Tones that relate to each other in terms of pitch or temporal proximity, exhibit a larger probability of being grouped together in the same acoustic object.

Continuity: A group of frequencies demonstrate a tendency to be grouped together, as long as they form a continuous trajectory or a discontinued but smooth trajectory. This principle is closely related to *proximity.*

Closure: The human auditory system has the ability to apply certain forms of anti-masking: it is likely to perceive a sound as continuous (and therefore as one acoustical object), even if it is interrupted by a broadband noise, provided that this sound is continued after the interruption.

Common Fate: Attributes that are subject to similar alterations will probably be grouped together: frequency components which originate from the same location in space (i.e. the same auditory scene position) share the same 'fate', and correspond to the same object. The same applies to components which are modulated at a similar rate or have simultaneous onsets and offsets (e.g. vocal performance idioms such as vibrato, formant change or pitch bending).

Similarity: According to this principle, grouping also depends on "vertical" (i.e. spectral) similarities of a stream*,* such as timbre. This principle also exhibits the property of being time independent.

In the next section, the efforts to computationally model and recreate the psychophysical task of ASA are described.

[11] The research that was carried out by Bregman focuses also on generalising the perception model across the various different aspects of the human sensory system.

2.6 Computational auditory scene analysis (CASA)

In general, CASA is the study of auditory scene analysis by computational means [121]. This definition is fairly vague, as it is entirely functional, does not make reference to underlying mechanisms, and—thus—lacks boundaries/limits. Without specifying any restrictions, the area of this field would span from modelling the sequence of action potentials produced by neurons of the cochlea [122], to the inference of metrical time-signature from music pieces [123]. Furthermore, without constraining this definition, the problem of, for example, source separation would have been circumvented by assigning an exclusive and isolated microphone for each source in an acoustic scene.

2.6.1. Scope of CASA

A definition that is more descriptive, and closer to what the signal processing community understands by the term CASA is "the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene" [32]. This also provides an intuitive classification of CASA between monaural and binaural methods. In addition, there seems to be a conventional interpretation of the field and the role of CASA that was gradually crystallised by many researchers [123-127] that have worked in this area over the past 20 years. According to this convention, CASA has two primary goals:

1. The development of a computational method, which is autonomously able to track and isolate a target sound source in a cocktail party scenario.
2. The implementation of an adaptive listening system, which is able to automatically compute and group *strands*, which are the equivalent of acoustic objects as described in ASA [1].

The applications of the latter endeavour extend to the aid of hearing-impaired individuals whose auditory system might be missing this capability [33]. Furthermore, the overall field of CASA has become associated with perceptually motivated approaches to sound separation that are distinct from other methods [32]. However, the human perceptual functions are not necessarily slavishly modelled or followed.

It should be noted that most CASA systems in the literature apply to the speech segregation or separation field. The systems that are widely used as reference points are based on common fate, continuity, and training for the case of mono CASA [123, 128-130]. On the other hand, in binaural CASA, the systems use mostly sound localization [131-134].

In the remainder of this section, the main features and tools of these and other similar CASA systems are briefly described. As there is a plethora of methods and variations in the said field, a unified description is attempted.

2.6.2. Data-driven and prediction-driven systems

Depending on the way in which the above processes are used in a higher-level algorithm, two "schools of thought" dominate the CASA field: *data-driven* and *prediction-driven* algorithms. In data-driven algorithms [135] cues are extracted from the spectrum of a sound, and representations of these cues form an abstraction of the original source (Figure 2.6). The information flow is exclusively unidirectional, from concrete to abstract [123].

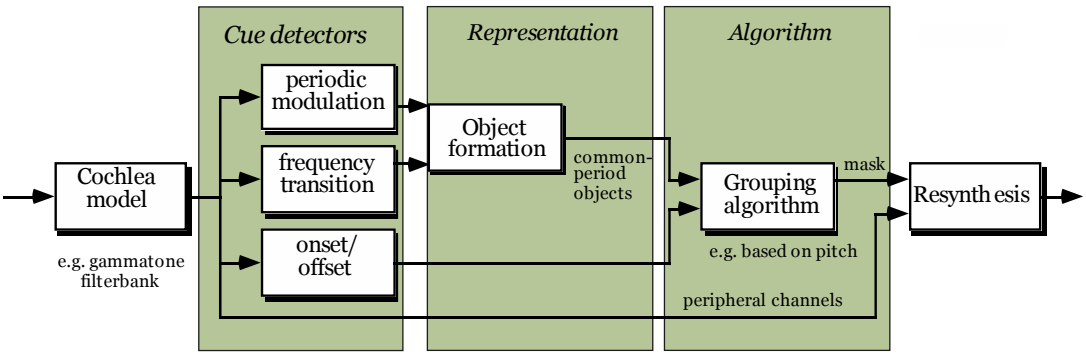

Figure 2.6: Workflow of a typical data-driven CASA system [124]

The first stage involves subjecting the acoustical mixture to a process that could be better described as a peripheral analysis. This is usually a time-frequency representation, such as the cochleagram (see Section 2.6.3) that is based on an approximate model of the human cochlea. Subsequently, features/cues are detected and extracted. The type of features is varied across the literature, but examples of such features include onsets/offsets [136], as well as amplitude and frequency modulation, corresponding to the description of 'common fate' in ASA. Afterwards, the cues/features are represented in a form that is between high and low level, i.e. representations termed 'mid-level' [137]. Again, several of these representations are proposed, such as *sinusoidal tracks* [138], and *synchrony strands* [135]. The most typical mid-level representation, however, is the correlogram (Section 2.6.4). From these representations the target source is selected, forming a time-frequency mask (described in 2.6.5), leading to a re-synthesis of the source (Section 2.6.6).

According to [125], the main problem with the data-driven approach is that it treats all sounds in the same way, regardless of their context. Concerns are also raised about the system's inability to detect masked sources. In order to engage these problems, the study in [123] proposed the prediction-driven approach, where the features extracted from the sound are compared with internal models of the components. In other words, this algorithm takes into consideration predictions for the continuity of the components and that is how

it resolves the masking problem. The modelling process considers the cases of noise, transient clicks, and a correlogram representation of periodic energy called the *weft* [137]. It must be noted, however, that there are significant similarities between the two approaches (data-driven and prediction-driven). In the next sections, the stages that are frequently common between these two categories of CASA systems are detailed.

2.6.3. Cochleagram

The cochleagram is the initial processing stage of most CASA systems [139-141]. It is a time-frequency representation of sound that models the known properties of human frequency selectivity [32]. The mechanics of the basilar membrane are commonly modelled by using a filter-bank which is the basis of the *gammatone filter* as described in [142].

The process starts by applying a pre-emphasis filter on the signal *x*, such as to model the outer and middle ear, that act as a high-pass filter. An approximation to this function is as follows [143, 144]:

$$y(t) = x(t) - \mathrm{A}x(t - \Delta t). \tag{2.1}$$

Here, A is the pre-emphasis factor and Δt is the sampling interval (e.g. 22.675 μs for 44.1 kHz). The pre-emphasis factor is slightly varied across CASA approaches and mostly set empirically [143]. Pre-emphasis is followed by the computation of the impulse response $g_f(t)$ of the gammatone filters:

$$g_f(t) = t^{\mathrm{N}-1} e^{-2\pi t b(f)} \cos(2\pi f t + \varphi) u(t). \tag{2.2}$$

In the equation above, N is the order of the filter, f is the central frequency of the filter, φ is the phase, $u(t)$ is the step function, and $b(f)$ determines the bandwidth using the expression:

$$b(f) = 1.02\text{ERB}(f). \tag{2.3}$$

The distribution of bandwidth is usually set in accordance with the equivalent rectangular bandwidth (ERB). ERB represents an ideal rectangular filter that exhibits equal peak gain across the whole frequency spectrum [145]:

$$\text{ERB}(f) = 6.23e^{-6}f^2 + 93.39e^{-3}f + 28.52. \tag{2.4}$$

Finally, the *energy measure* is used to create short-time energy spectra [144]:

$$e_{f(t)} = \frac{\Delta t}{W}\sum_{k=0}^{W/\Delta t}\left|g_f(t - k\Delta t)\right|^2 e^{-a\Delta tk}, \tag{2.5}$$

where $e_f(t)$ is the energy measure output of the gammatone filter $g_f(t)$ centred at frequency f at time t, while W is the window length over which the energy measure is computed, and α represents the decay of the exponential window. A comparison between a cochleagram that is generated following the process described here and a log-frequency spectrogram is given in Figure 2.7. At such coarse temporal scale, they both look similar; however, it is evident that the cochleagram represents onsets in a much clearer fashion.

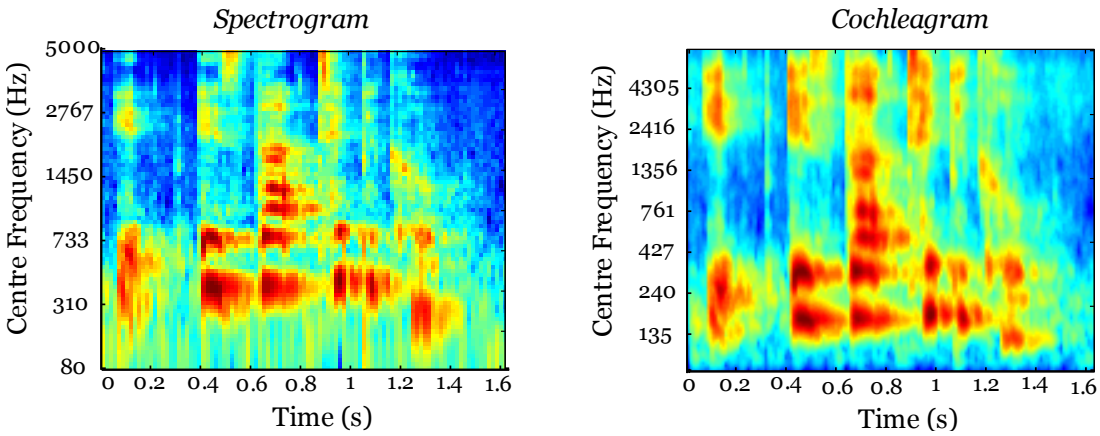

Figure 2.7: Log-frequency spectrogram (left) and cochleagram (right) of the female speech utterance "Is that typical?". The vertical axis of the spectrogram is limited to show the same frequency range as the cochleagram, for the purpose of direct comparison.

2.6.4. Correlogram and cross-correlogram

A form of representation that is used frequently in CASA, but also in other fields such as speech recognition, and geology [146, 147] is the correlogram [148]. For the case of CASA, this is again a perceptually motivated short-term representation of sound, based on pitch perception [149, 150], and it produces symmetric structures that are used in order to estimate groups of spectral components (i.e. components that probably belong to the same acoustical object) for each frame. The "channels" of the correlogram are the different time-frequency components and are represented in such a way that the amplitude in each channel is most likely corresponding to a single periodic source [146]. The correlogram is usually computed in the time domain (although studies have suggested greater efficiency when computed in the frequency domain [151]) using the autocorrelation function:

$$acf(n, c, \tau) = \sum_{k=0}^{\mathrm{K}-1} a(n-k, c)a(n-k-\tau, c)h(k). \tag{2.6}$$

As seen above, $a(n, c)$ models the action potential of the auditory nerve for frequency channel c at time n, τ is the time lag, and K is the length of the—typically—Hann window $h(k)$. A correlogram of a "sawtooth" wave with fundamental frequency 440 Hz after pre-processing with (2.1) can be seen in Figure 2.8. Here, the first peak of the summative correlogram is at 2.27 ms which corresponds to the period of the f_0, i.e. T(440 Hz) ≈ 0.0027 s, as expected.

When cross-correlation instead of autocorrelation is used, the resulting representation is defined as a cross-correlogram [152]. The perceptual motivation this time is the left and the right ears of the listener [153]. This concept forms also the base for the inter-aural time difference (ITD) and inter-aural phase difference (IPD) that are used across many source separation methods and are further discussed later in this thesis.

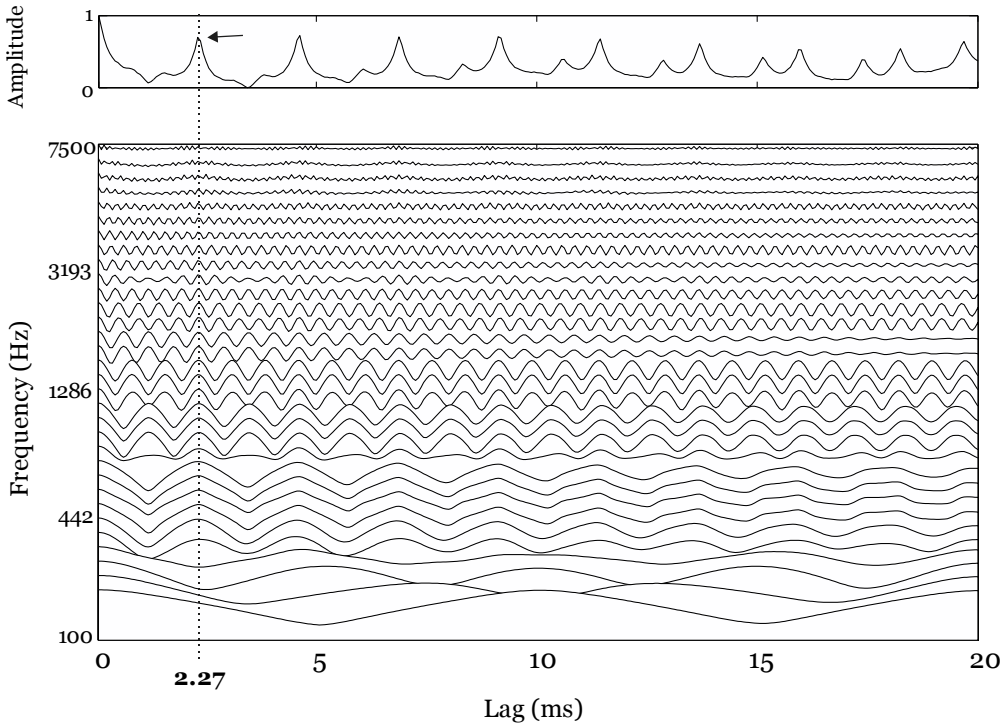

Figure 2.8: Correlogram of a sawtooth-wave tone at 440 Hz (bottom) and summative correlogram with normalised amplitude (top)

In a similar way to (2.6) the cross-correlogram is computed as [140]:

$$ccf(n, c, \tau) = \sum_{k=0}^{K-1} a_L(n-k, c) a_R(n-k-\tau, c) h(k). \tag{2.7}$$

The indices $_L$ and $_R$ represent the left and the right ears respectively.

2.6.5. Time-frequency mask

A time-frequency (TF) mask constitutes the cornerstone of many CASA systems. This mask works usually on a frame by frame basis and its function is to isolate the spectral components that belong to a target source [154]. It can be classified into two categories: the binary TF mask, and the soft TF mask.

The binary TF mask can be better visualised as a two-dimensional logical array. This array is superimposed on the short-term frequency magnitude-spectrum (or energy-spectrum) of the target source and allows only "true" bin values to pass through at certain while the others are cancelled [155]. On the other hand, a soft TF mask works in the same way, except that the array is not binary. Instead, it provides a gain for each frequency-transformed magnitude point in time [156].

When a binary TF mask is included in a system as the last process, it also sets limits on the system in question [157]. Particularly, its maximum isolation effectiveness matches the performance of the "ideal" mask, which is well defined in literature.

Formally, the computation of the ideal binary mask (IBM) is as follows [158]:

$$IBM(t,f) = \begin{cases} 1, & T(t,f) - N(t,f) > \mathrm{Z} \\ 0, & otherwise \end{cases}. \tag{2.8}$$

Here, $T(t,f)$ is the spectrum of the target mixture and $N(t,f)$ is the spectrum of the interference, while t indicates the time index, and f represents the bin index. Z is a binary threshold that is usually set empirically. For the case of the ideal soft mask (ISM) or ratio mask [159], equation (2.8) becomes:

$$ISM(t,f) = \frac{T(t,f)}{M(t,f)}. \tag{2.9}$$

Above, $M(t,f)$ is the spectrum of the target mixture, i.e $T(t,f) + N(t,f)$. An example of these two types of TF mask for an excerpt for the song "Salala" [160] are given in Figure 2.9. In this case, it is obvious that even the ideal binary mask, would let through much interference because of the significant overlapping frequencies of voice and music in the mixture. On the other hand, ISM provides more satisfactory results, given an accurate phase estimate.

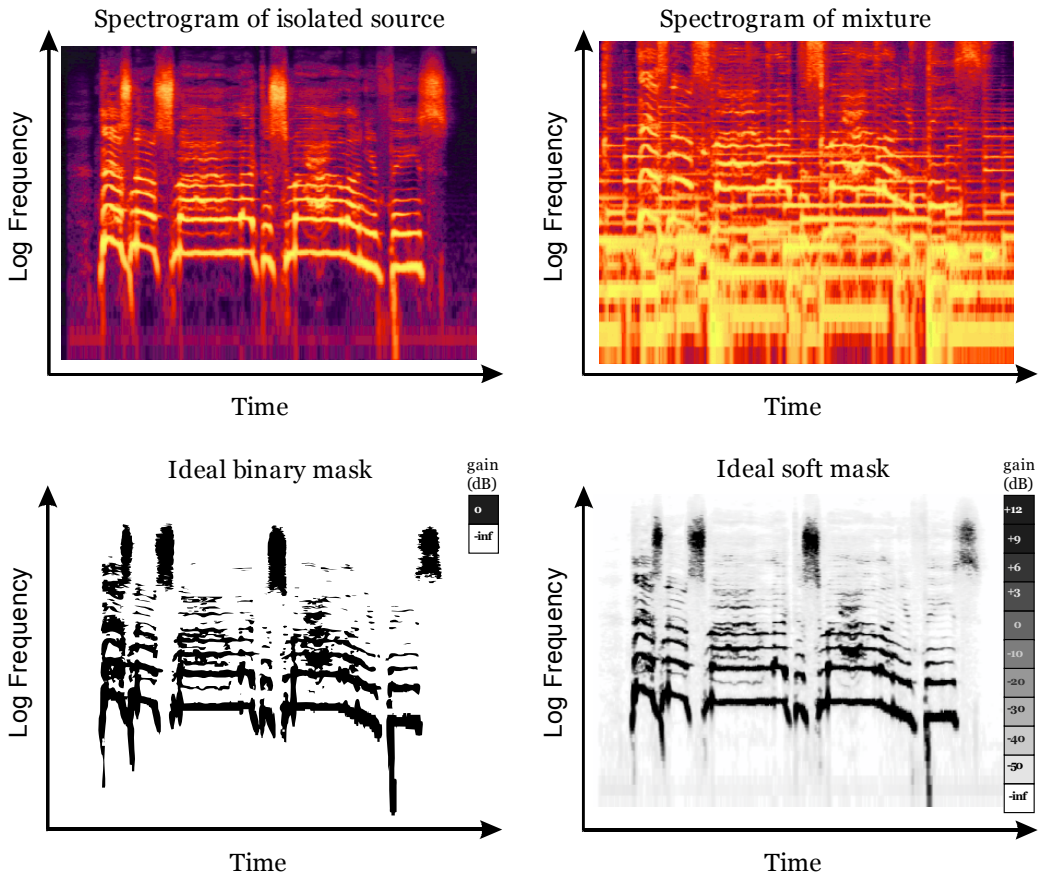

Figure 2.9: Spectrograms and TF masks. Isolated source (voice) is top left and mixture is top right. The ideal binary and soft masks for the target source are shown on bottom left and bottom right respectively. The threshold Z is set to 1 dB. Scaling factors in the TF masks are represented in dB gain.

2.6.6. Re-synthesis

The previous procedures describe different representations and methods that are motivated by the human auditory system. Re-synthesis is the term that is used in this study for the purpose of collectively describing the various transformations from the aforementioned representations back to the time domain. Since some of the representations are already in the time domain (e.g. the correlogram), a more accurate statement is that re-synthesis in this context is the conversion to an acoustic equivalent of the detected source [123]. This stage serves also to provide an output for subsequent evaluation of the effectiveness of a system.

The most intuitive approach for re-synthesis concerns the transformation of the TF-masked spectrogram, which is done with the process of inverse short-term Fourier transform (ISTFT) e.g. [135]. However, there are approaches that suggest transformation from the output of the gammatone filter-bank [130] or even inversion of the correlogram in order to produce a single time-domain signal [161].

2.7 Chapter summary

This chapter has presented an overview of the anatomy of the singing voice given together with the major differences between speaking and singing voice. These differences were highlighted in order to clarify the deviation of the challenges between speech segregation and SVS. The two foundational fields of SVS were analysed: the auditory scene analysis (ASA) and the perceptually based computational auditory scene analysis (CASA). These two fields are in close relationship with each other not only because CASA is motivated by ASA, but also due to the modelling of the human auditory system that CASA aims to achieve. CASA is the oldest attempt in source separation with computational aid and the majority of its approaches concern monaural inputs. The next chapter details the process of stereophonic music production and focuses on stereophonic source separation methods that are important in the context of this thesis.

3

BLIND SOURCE SEPARATION AND INDEPENDENT COMPONENT ANALYSIS

This chapter is introducing the family of the blind source separation (BSS) algorithms. These methods are based on statistical properties of the signals, e.g. principal component analysis (PCA) and independent component analysis (ICA). Due to the reason that these algorithms are closely linked, particular weight is given in highlighting the differences between them. The chapter concludes with the description of the most prominent algorithm in this family (i.e. fast ICA) and the identification of the shortcomings thereof with regards to SVS.

3.1 Blind source separation

The field of blind source separation (BSS) postulates the challenge of *blindly* separating sources. The term was first used by Herault and Jutten in 1986 [162]. In this case, the sources are the original signals, e.g. speakers in a cocktail-party problem. The adjective *blind* is used because only weak assumptions of the original sources can be made. This terminology stems from the field of digital communications, where *blind techniques* were intended to work when the "eye", i.e. the oscilloscope diagram of a synchronised discrete signal, was closed [163].

In literature, the terms BSS and ICA are often used interchangeably [164, 165]; in this thesis, however, BSS is studied as a specific challenge, while PCA and ICA (sections 3.2 and 3.3) are discussed—in the rest of this chapter—as proposed solutions to the BSS problem.

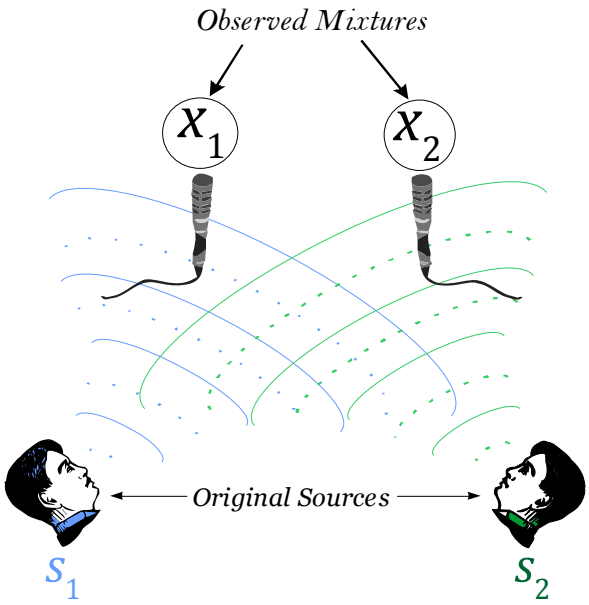

Figure 3.1: BSS challenge model

In its simplest form, the challenge of BSS is to separate two utterances (s_1 and s_2) by solely observing their mixtures x_1 and x_2 (Figure 3.1). In this model, linearity and stationarity of the mixing effects of the environment are assumed. Thus, Figure 3.1 can be formally expressed as [166]:

$$x_1[n] = a_{11}s_1[n] + a_{12}s_2[n], \text{and} \tag{3.1}$$

$$x_2[n] = a_{21}s_1[n] + a_{22}s_2[n], \tag{3.2}$$

where α_{ij} are the mixing parameters that depend on the distance and the axis of each speaker to the microphones, while x_1 and x_2 are the observed mixtures (e.g. the two channels of a stereo recording), and n is the discrete time index.

For N sources and mixtures, the equations (3.1) and (3.2) are rewritten describing—thus—a latent (i.e. hidden) variables model [167]: Let N sources x_i … x_N be modelled as linear combinations of J random variables s_j, i.e.

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{iJ}s_J \qquad a_{ij}, i, j \in \mathbb{Z}. \tag{3.3}$$

Note that index n is dropped from the above equation. This model is called the *instantaneous model,* in which any delay between the sources or interference (noise) that might exist in the observed mixture is not considered [168]. The relation between the number of sources J and the number of observations N defines two categories of BSS: when J>N the class of BSS is (*under*)*complete* or *underdetermined* [169], while J≤N defines an *overcomplete* or *overdetermined* case [170]. The assumption that is common in all the approaches to BSS is the statistical independence of the sources s_J. An explanation of this statistical property is provided in Section A.1 of the appendix.

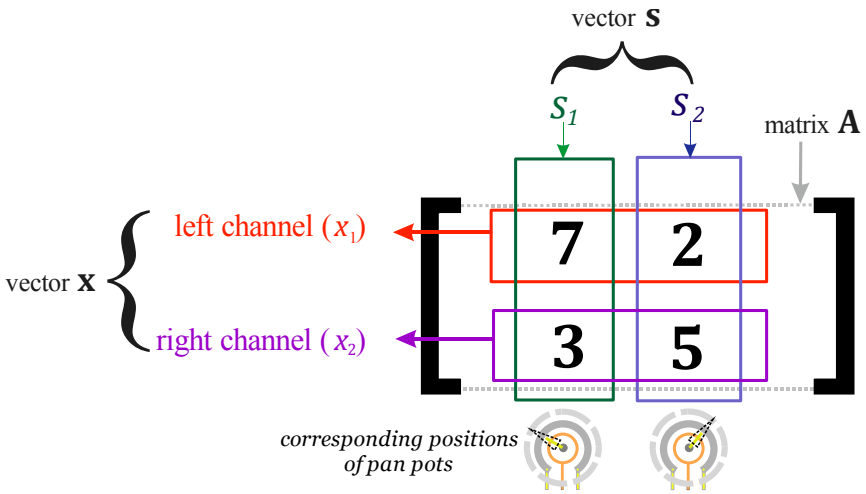

Figure 3.2: Function of the mixing matrix in the instantaneous model. The stereo signal denoted as vector **x** is produced from two sources that constitute vector **s**. This mixture, in terms of stereo localisation, gives the impression that the s_1 is left and the s_2 is right. Provided that $var(s_1) \cong var(s_2)$, s_1 will also be perceived as being louder than s_2.

In vector/matrix notation, equation (3.3) can be simplified. In equation (3.4) vectors are denoted as bold lowercase letters and matrices as bold uppercase letters. Vectors are column vectors unless otherwise stated.

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{3.4}$$

where **A** is the mixing matrix.

Figure 3.2 illustrates how a mixing matrix functions to produce a stereo mixture from two sources. In order to solve equation (3.4) for **s**, the inverse (i.e. $\mathbf{A}^{-1}$) of the mixing matrix is needed. This *de-mixing* matrix is noted as **W**:

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{3.5}$$

The objective of BSS is to estimate the de-mixing matrix using an inverse system (referred to as a reconstruction system), which is commonly based on a neural network and adaptive learning [171, 172]. A unified model of the algorithma that target BSS can be seen in Figure 3.3.

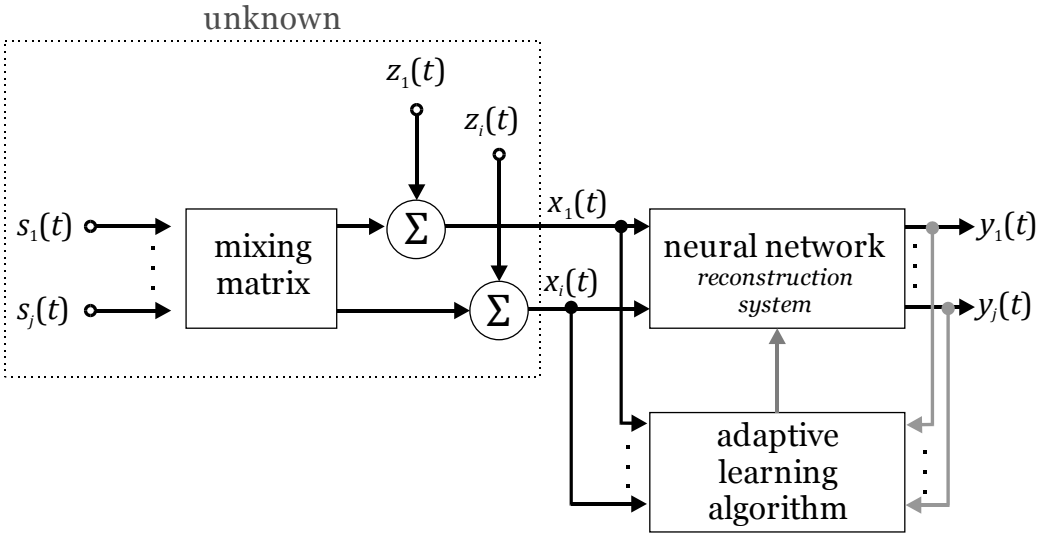

Figure 3.3: BSS generic model [173]. $s_j(t)$ and $y_j(t)$ are the original and estimated sources respectively. The term $z_i(t)$ represents interference (noise) where applicable.

Here, the original sources are mixed with an unknown matrix and contaminated with noise. The target of the system is to derive the decontaminated sources s_i by observing their mixtures x_i.

After formulating the problem of BSS, the rest of this chapter discusses methods that attempt to solve this problem. Particular focus is given to principal component analysis (PCA) and Independent component analysis (ICA). These methods enjoy growing interest in many research areas, such as astrophysics [174], image de-noising, data compression, and magnetoencephalography (MEG) [175]. However, the discussion in this thesis is adjusted to the perspective of their application in audio source separation from stereophonic signals.

3.2 Principal component analysis (PCA)

PCA is a mathematical process of second order statistics, developed in 1901 by Pearson and is often used as a dimensionality reduction technique in multivariate statistical analysis [176, 177]. Its main feature is the

decomposition of the covariance matrix of input data.

The general PCA model usually begins with centring the data: The mean of the data becomes zero by subtracting the initial mean average. This procedure is only for simplifying the model and is not destructive, as the mean can be added back at a later stage [178].

Taking into consideration the problem postulated by BSS in (3.4), PCA aims to linearly transform the vector $\mathbf{x}$ into $\ddot{\mathbf{x}}$ by multiplying it with a matrix $\mathbf{V}$ so that $\mathrm{E}[\ddot{\mathbf{x}}\ddot{\mathbf{x}}^T] = \mathbf{I}$, Where $\mathrm{E}[\cdot]$ is the expected value, and $\mathbf{I}$ denotes the identity matrix:

$$\ddot{\mathbf{x}} = \mathbf{V}\mathbf{x}. \tag{3.6}$$

The aforementioned process is also referred to as *whitening*[12] [179] or *sphering* [180]. From the above equation and equation (3.4), derives a new mixing matrix $\tilde{\mathbf{A}}$ as follows[13]:

$$\ddot{\mathbf{x}} = \mathbf{V}\mathbf{A}\mathbf{s} = \tilde{\mathbf{A}}\mathbf{s}. \tag{3.7}$$

The principal components of the bivariate vector $\ddot{\mathbf{x}}$ are the orthonormal basis unit vectors, that demonstrate the maximised variance of the data [181]. The main advantage of this transformation is that the new mixing matrix $\tilde{\mathbf{A}}$ is orthogonal:

$$\mathrm{E}[\ddot{\mathbf{x}}\ddot{\mathbf{x}}^T] = \tilde{\mathbf{A}}\mathrm{E}[\mathbf{s}\mathbf{s}^T]\tilde{\mathbf{A}}^T = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{I}. \tag{3.8}$$

[12] Generally, whiteness of a zero-mean random vector means that its components are uncorrelated and their variance equals one.

[13] In this thesis, a tilde (~) is used over a letter only as a diacritic, and does not indicate matrix transposition, which is denoted by the superscript italic letter T (i.e. ^{T}).

Orthogonality of the target de-mixing matrix is seen as a way to limit the search to the group of unitary or orthonormal matrices, depending on the case [163]. In fact, this process significantly improves the efficiency of the algorithm as orthogonal matrices have only $n(n\text{-}1)/2$ parameters, instead of n^2. For example, for a two-dimensional matrix, a single angle parameter is sufficient [182, 183]. This is also a significant aid to computational efficiency as whitening is a simple and standard procedure and reduces the complexity of the problem of BSS at least by a factor of two [184].

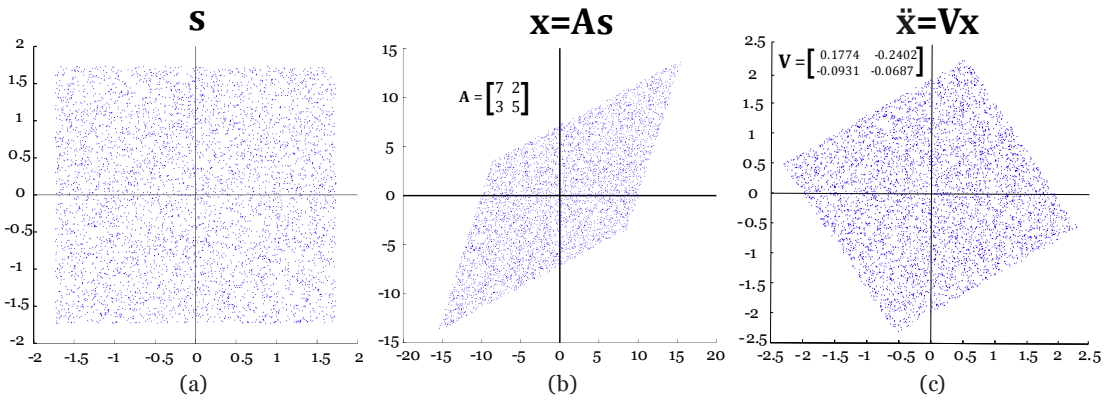

Figure 3.4: Bivariate distributions of a) two sources with uniform distributions ($\mu \approx 0$, $\sigma \approx 1$). horizontal axis: $s_1(t)$ vertical axis: $s_2(t)$ b) two mixtures of the same sources and c) the same mixtures after whitening. The equivalent expressions of the instantaneous model are presented on top of each drawing.

An example of whitening of mixtures deriving from uniformly distributed data can be seen in Figure 3.4. Here, $dist(cov(\ddot{\mathbf{x}}), \mathbf{I}) \approx 5 \times 10^{-15}$, where $dist(\cdot)$ denotes Euclidean distance, and $cov(\cdot)$ denotes covariance [185]. From a visual perspective, whitening initially rotates a matrix and subsequently stretches it. The directions of the mixing matrix are visually so prominent, that it is almost tempting to try to estimate the mixing matrix heuristically. However, as discussed in A.1, such methods are generally not robust.

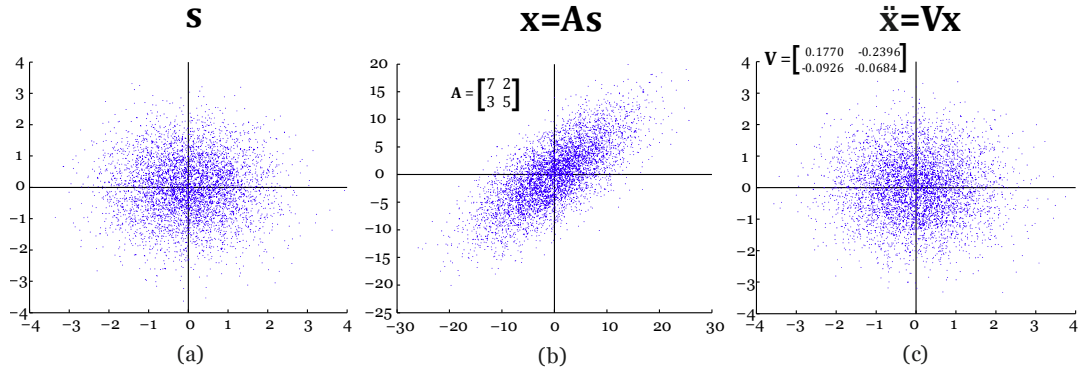

Figure 3.5: Bivariate distributions of a) two sources with Gaussian distributions (μ≈0, σ≈1). horizontal axis: $s_1(t)$ vertical axis: $s_2(t)$ b) two mixtures of the same sources and c) the same mixtures after whitening. $dist(cov(\ddot{\mathbf{x}}), \mathbf{I}) \approx 8 \times 10^{-15}$. The equivalent expressions of the instantaneous model are presented on top of each drawing.

When the sources are normally distributed, whitening does not make the problem any easier and—as can be seen in Figure 3.5—there is essentially no change in their joint distribution.

This is because no amount of rotation and stretching can make the distribution orthogonal. The limitations that Gaussian distributions pose are going to be discussed further in Subsection 3.3.4.

There are several methods [184] that can be used in order to perform PCA. In the next two subsections, two different methods for the purpose of whitening or solving the BSS using PCA are explored. These are the eigenvalue decomposition (EVD), and the singular value decomposition (SVD) [186]. The close relationship of these two will become apparent as the section progresses.

3.2.2. Using eigenvalue decomposition (EVD)

Eigenvalue decomposition is a procedure of linear algebra, whereby the eigenvalues λ_i and the eigenvectors $\mathbf{g}_i$ of a matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ are related as follows [187]:

$$\mathbf{M}\mathbf{g}_i = \lambda_i \mathbf{g}_i \,. \tag{3.9}$$

Here, λ_i is a scalar and $\mathbf{g}_i \neq 0$. The factorisation of $\mathbf{M}$ can be expressed as:

$$\mathbf{M} = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{-1}, \tag{3.10}$$

where $\mathbf{G} \in \mathbb{R}^{N \times N}$ contains the eigenvectors $[\mathbf{g}_i \ldots \mathbf{g}_N]$ of $\mathbf{M}$, while $\boldsymbol{\Lambda}$ contains the corresponding eigenvalues λ_i in its diagonal in such way that $\Lambda_{1,1} \geq \Lambda_{2,2} \geq \ldots \Lambda_{N,N}$.

Each vector of matrix $\mathbf{G}$ is an estimated de-mixing vector, i.e. a $\mathbf{w}_i$, that produces a principal component (PC) when multiplied with the input $\mathbf{x}$ from (3.5) [188, 189]. For the case of solving BSS the usual procedure is to derive PCs by sorting them according to their eigenvalues.

For the purpose of whitening, the matrix $\mathbf{V}$ of equation (3.7) is derived as in (3.11) [190]:

$$\mathbf{V} = \left(\mathbf{G}\boldsymbol{\Lambda}^{\mathbf{0.5}}\right)^{-\mathbf{1}} = \boldsymbol{\Lambda}^{-\mathbf{0.5}}\mathbf{G}^{-1} = \boldsymbol{\Lambda}^{-\mathbf{0.5}}\mathbf{G}^{T} \,. \tag{3.11}$$

3.2.3. Using singular value decomposition (SVD)

The covariance matrix can also be decomposed with the method of singular value decomposition (SVD) [191]. When applied to the mixture, the process of SVD for a matrix $\mathbf{X}$ gives:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{H}^{T} \,, \tag{3.12}$$

where $\mathbf{U} \in \mathbb{R}^{M \times M}$, and $\mathbf{H} \in \mathbb{R}^{N \times N}$ are orthonormal, while $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times N} \geq 0$ is diagonal.

The diagonal values of **Σ** are also called the *singular values* of **X,** and they satisfy the condition $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \ldots \Sigma_{N,N}$ [190]. On the other hand, **U** and **H** contain the *left singular* vectors and *right singular* vectors respectively[14].

The instantaneous model is discarded here so that **X** denotes the matrix of *all* samples of mixtures, i.e. **X = AS**, where the components of **S** are listed below:

$$\mathbf{S} = \begin{bmatrix} s_1[n] & \cdots & s_N[n] \\ \vdots & \ddots & \vdots \\ s_1[N] & \cdots & s_J[N] \end{bmatrix}. \tag{3.13}$$

The number of time samples is denoted as N, while the number of sources as J.

Considering the similarities here with the decomposition of EVD, the left singular vectors are the eigenvectors of $\mathbf{M}\mathbf{M}^T$, the right singular vectors are the eigenvectors of $\mathbf{M}^T\mathbf{M}$ while the diagonal **Σ** contains the square roots of $\mathbf{M}^T\mathbf{M}$ and $\mathbf{M}\mathbf{M}^T$.

By rewriting equation (3.12) the de-mixing matrix can be computed in the following way:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{H}^T \Leftrightarrow \mathbf{X}^T = \mathbf{H}\boldsymbol{\Sigma}^T\mathbf{U}^T \Leftrightarrow \mathbf{U}^T = \boldsymbol{\Sigma}^{-T}\mathbf{H}^T\mathbf{X}^T , \tag{3.14}$$

Thus, the de-mixing matrix **W** of equation (3.5) is identified as:

$$\mathbf{W} = \boldsymbol{\Sigma}^{-T}\mathbf{H}^T . \tag{3.15}$$

[14] In SVD, the bold capital **V** is usually used for the right singular vectors. In this study this is replaced with **H** in order to avoid confusion with the matrix **V** that is used to modify the mixing matrix for the purpose of whitening in equation (3.6) onwards.

Taking into consideration equations (3.5) and (3.15), the columns of the matrix **U** correspond to the de-mixed signals (i.e. the estimated sources **S**) [181].

Accordingly, for the case of whitening the expression is:

$$\mathbf{V} = \mathbf{\Sigma H}^{T}\sqrt{\mathrm{N}}\,, \tag{3.16}$$

where N is the length of mixing matrix **A** as in (3.4).

In the next section, the ICA technique is described, as well as the usage of PCA in its context for the purpose of pre-processing.

3.3 Independent component analysis (ICA)

ICA is a general purpose statistical technique that is closely related to the case of blind source separation (BSS), as the assumptions regarding the nature of the original sources are minimal. In addition, ICA is conceivably one of the most widely used BSS methods [167]. The name of this technique derives from its direct association with PCA [192]. The key difference between PCA and ICA is that the latter uses higher-order statistics: while PCA estimates components by deriving a de-mixing matrix that maximises variance (the second moment), the de-mixing matrix that derives from ICA is based on maximising independence, which is associated with the fourth moment, i.e. kurtosis.

3.3.1. Uncorrelated vs. independent multivariate distributions

At this stage, the difference between the absence of correlation between two random variables and their statistical independence is detailed. This is deemed to be of particular importance as a) it is one of the two critical differences between PCA and ICA, and b) the distinction is not always clear in the literature.

Intuitively, two variables (e.g. *x* and *y*) are independent if *x* does not give any information about *y*.

Although independence implies lack of correlation, this is not commutative. For example, if *x* is the result of rolling one die and *y* is the result of rolling the same die a second time, *x* and *y* are independent (and uncorrelated). On the other hand, if *x* is a card drawn randomly from a deck and *y* is a card drawn subsequently from the same deck *without replacing x*, then *x* and *y* are uncorrelated but not independent, as drawing *x* made drawing *y* more probable). In technical terms, their independence can be determined by their probability densities (i.e. the occurrence of *x* makes the occurrence of *y* neither less nor more probable and vice versa). The relationship of lack of correlation and independence is formally explained below.

Generally, two variables *x* and *y* are deemed to be statistically independent when:

$$p(x \cap y) = p(x)p(y), \tag{3.17}$$

where $p(\cdot)$ denotes probability. Assume that the bivariate distribution of (y_1, y_2) comprises discrete values and have a uniform distribution among the values [0, 1], [0,-1], [1, 0], [-1, 0]. These are uncorrelated, as their covariance is zero:

$$\mathrm{E}[y_1 y_2] - \mathrm{E}[y_1]\mathrm{E}[y_2] = 0, \tag{3.18}$$

but they are not independent because:

$$\mathrm{E}[y_1^2 y_2^2] = 0 \neq \frac{1}{4} = \mathrm{E}[y_1^2]\mathrm{E}[y_2^2]. \tag{3.19}$$

3.3.2. Basic concept of ICA

According to the central limit theorem (CLT), the distribution of a sum of independent random variables tends toward a Gaussian distribution, provided that the random variables have finite variance [193]. The estimation of the de-mixing matrix **W** in ICA is based on the reverse of this theorem, i.e. different values of **W** are applied in iterations, and converge when the rows of **s** (the estimated sources) reach maximally non-Gaussian distributions.

3.3.3. Maximisation of non-Gaussianity

In order to examine how the maximisation of non-Gaussianity is carried out, assume sources s_1 and s_2 where their probability distribution functions are identical, i.e. $pdf(s_1) \equiv pdf(s_2)$. These two sources are mixed with an unknown mixing matrix **A**, in order to produce the observed mixtures x_1 and x_2. After the centring and whitening of these mixtures, the target is to estimate initially one of the sources (say $\check{s}$). Consider **w**, which is a row of the de-mixing **W** from equation (3.5), as the de-mixing vector for this source.

Formally expressed, the above becomes:

$$\check{s} = \mathbf{w}^T\mathbf{A}\mathbf{s} = \mathbf{w}^T\mathbf{x} = \sum_i w_i\, x_i, \tag{3.20}$$

where i is the source index. Hence, $\check{s}$ is a linear combination of s_1 and s_2 with coefficients $\mathbf{w}^T\mathbf{A}$, that should be either [1, 0] or [0, 1].

At this point the theoretical foundation of the central limit theorem (CLT) is employed. As mentioned before, according to CLT the distribution of a sum of independent random variables tends toward a Gaussian distribution, provided that the random variables have finite variance [193]. So, the sum of two independent values (in this case **x**) is more Gaussian than s_1 or s_2. As a result, $\mathbf{w}^T\mathbf{x}$ will be the least Gaussian when it equals, in fact, one of the sources.

Therefore, the aim here is to find a vector **w** that maximises the non-Gaussianity of $\mathbf{w}^T\mathbf{x}$; this vector cannot be determined, but it can be estimated.

This is one of the core ideas of ICA, namely that *maximising the projected non-Gaussianity* gives us *one* of the independent components. However, **w** can have 2*n* local maxima, i.e. two for each source (this is the reason that one of the ambiguities of ICA is the signal phase of the obtained ICs as described in Section 3.5).

After the estimation of the 'first' **w**, the rest of these components are estimated by finding all the local maxima. In addition, since the sources are assumed uncorrelated, the search for the rest of the estimates is restricted to the ones that exhibit poor correlation with the first IC.

3.3.4. Assumptions of ICA

In this subsection, the main restrictions for estimating the de-mixing matrix using ICA are listed. However, it should be noted that all these restrictions have been proven quite relaxed, as there have been studies on tackling each one of them. These studies are mentioned here but not described, as this review is mainly concentrated on the main idea of ICA. Later in this thesis, the effect of these restrictions is investigated, particularly for the case where the observed mixtures are music recordings.

Square mixing matrix: One of the critical requirements is that the number of observed mixtures x_i equals or outnumbers the number of sources s_j, i.e. N≥ J. Otherwise, the problem is called underdetermined and the matrix cannot be inverted. However, during the last few years, there are studies suggesting that, even when the observed mixtures are fewer than the original sources, components can still be computed [194-196].

Non-Gaussian sources: The variables s_j are not allowed to have Gaussian distributions. If they do, that means that the observed mixtures x_i are also Gaussian. As it is explained in 3.4, higher-order statistical information diversity

is essential for the estimation of **W** in ICA; in the case of Gaussian distribution, skewness and kurtosis are zero [197]. Nonetheless, it is not assumed that the pdfs of s_j are known (if they were known, the problem would be much simpler). In some instances of ICA, though, the assumption of non-Gaussianity can be replaced by assumptions on the time structure of the signals [198].

Statistical independence: The sources s_j must be statistically independent. This is explained in 3.3.1 and it is perhaps one of the fundamental restrictions of ICA, as the acronym itself suggests. At this point, it is important to emphasise that this restriction concerns *only* the sources s_j and *not* the mixtures x_i that are—in every case—dependent (see A.1). The assumption of independence is not unrealistic in most cases and—surprisingly—it doesn't need to be exactly true [199, 200].

Stationarity: The instantaneous model in equation (3.3) has the restriction that the statistical properties of the sources as well as of the mixing matrix, must not change in time. In music terms, the equivalent statement is that a source should not change in pitch, timbre, amplitude, or panning location. For example, a sine wave is stationary, while a note from a piano is not (the sound fades out, and, at the risk of being pedantic, changes in timbre and decreases in frequency after the hammer strikes the string [201]). Generally, a stochastic process is said to be *stationary in the strict sense* only if its joint density doesn't change between time shifts. Once again this restriction has been proven—in practice—to be able to be bypassed [202, 203].

Noise omission: In the instantaneous model of ICA, the noise term is omitted. For many purposes of ICA, this would be an unrealistic assumption; nevertheless, it is made, since the estimation of the noise-free model is difficult enough in itself. Despite this, this assumption seems to be effective in many applications [204], such as fetal signal reconstruction, where the noise is treated as a separate source [205]. However, there are some approaches that include a noise term in their problem formulations and, with appropriate

modifications, the cancellation of noise as well as the separation of sources are claimed to be successful [206, 207].

Linearity: By linearity of the mixing process, it is implied that any delays[15] or reflections that would occur in a real-world environment are disregarded. This is because the estimation of independent components from non-linear mixtures poses a fundamental problem: solutions always exist and they are highly non-unique [208].

Filtering: For the purpose of solving BSS time signals (e.g. audio signals) using ICA, frequency filtering can be occasionally very useful [204], as discussed later in Chapter 6. Although each case is different and general filtering parameters cannot be generalised it is important to show that the ICA model still holds after the filtering.

Hence, equation (3.4) is rewritten for the case of continuous signals, i.e. considering the composition of **S** as in (3.13) and **X** in a similar manner, where columns of **X** represent individual observations in time:

$$\mathbf{X} = \mathbf{AS} \tag{3.21}$$

When filtering (represented as matrix **F**) is applied the expression becomes:

$$\acute{\mathbf{X}} = \mathbf{XF} = \mathbf{ASF} = \mathbf{A}\acute{\mathbf{S}} \tag{3.22}$$

This means that the model for the mixing matrix is still the same, but the target sources are filtered.

[15] As seen in Figure 3.1, s_1 is not equidistant from the two microphones. This results in a time/phase delay between the presence of s_1 in the two mixtures.

3.3.5. Effect of ICA assumptions in stereophonic studio recordings

Since the object of this study is source separation from polyphonic observations, it is important to examine how restrictive the aforementioned assumptions are for this case.

The restriction[16] of time delay usually does not pose an obstacle, as the phase of the signal in studio recordings is coherent between the left and the right channels and only its intensity varies, as discussed in Chapter 4.

On the other hand, statistical independence and stationarity can form a strong challenge in this field as, music sources in a song do correlate and they are generally not stationary [209].

Noise omission from the model of ICA is not deemed to pose a significant challenge as every source in a music mixture is considered rather a music instrument (even if in a different context it would be noise).

Non-linearity in stereophonic recordings is quite often the case, as non-linear effects, such as compression, limiting, and gating, are commonly used. When these effects do not exist, the case is deemed to be plausible for ICA, as the mixing is linear and therefore a de-mixing matrix **W** exists.

The assumption of non-Gaussianity holds for the case of music as the music sources are generally not Gaussian. However, convolution during mixing makes the sources more Gaussian. This is because in cases such as mixing (as well as others), convolution increases entropy [210]. As a result, the more processing a source has, the more difficult is for ICA to find its de-mixing matrix. This is especially evident for processing that has long filter kernels, e.g. reverberation and delay.

The most restrictive assumption for the case of stereophonic music recordings

[16] An exception to this is the addition of stereo delay-based effects in music production, which is nevertheless rare for the case of the main vocal part (though quite common for the backing vocals).

is the square mixing matrix. Clearly, most stereophonic tracks of music contain more than two sources [211] and can reach tens or even hundreds in the extreme. With regards to this, Section 4.1 details mixing procedures in stereophonic recordings and mixing matrices thereof.

These restrictions indeed suggest against the feasibility of the use of ICA on its own in the case of music. As can be seen in subsequent chapters, however, it can still be used as a pre-processing step. In the next section a specific algorithm that performs ICA is reviewed and analysed.

3.4 Fast independent component analysis (FICA)

There are several algorithms that attempt to solve the model of BSS by using ICA. These include the *Infomax* [212] algorithm that uses the maximum likelihood estimate, and *Jade* [213] that uses EVD. Of particular interest to this study is the fixed-point algorithm of fast ICA (FICA) because of its speed and robustness [214]. In this section, FICA is described, with emphasis given on the metric of Gaussianity that is used.

3.4.1. Pre-processing

Centring: Similarly to the pre-processing used for PCA in Section 3.2, the mixtures are linearly 'shifted' so that their means are zero. This process is reversed after the de-mixing matrix **W** is estimated: The expression $\mathbf{W}\Delta\mu$ (where $\Delta\mu$ is the mean that was initially subtracted for centring) produces the mean vector of **s** so that it can be added back to the ICs.

Whitening: PCA is used here for the purposes of whitening. Although PCA cannot recover non-orthogonal sources [178], it has the advantage that the analysis can be based on second-order statistics only and can be a useful pre-processing step for ICA [214]. The use of PCA as a spatial whitening technique has commonly been viewed as a way to simply limit the space of de-mixing

matrices to orthogonal ones. In practice, though, it also demonstrates the property of reducing the unwanted contribution of additive noise in the mixtures [190].

3.4.2. Measuring Gaussianity in FICA

One of the most challenging steps of ICA is to measure the Gaussianity of the resulting vector, after each iteration/trial of a de-mixing matrix **W**. Here, two standard metrics for this purpose, namely *kurtosis* and *negentropy* (negative entropy), are described and reviewed since most algorithms that perform ICA are based on these methods or a combination of them. The section continues to detail how they are combined for measuring Gaussianity in the fixed-point algorithm FICA.

3.4.3. Kurtosis

The most intuitive measure of Gaussianity is kurtosis, the fourth-order cumulant. Kurtosis measures the curve "narrowness" of the probability distribution shape. If the curve is narrower than Gaussian, the kurtosis is positive[17] and the distribution is called *leptokurtic* or *super-Gaussian*. If the curve is wider than Gaussian, the kurtosis is negative and the distribution is called *platykurtic* or *sub-Gaussian.* Gaussian distribution has a kurtosis of zero and is also known as mesokurtic [215].

Typical examples of mesokurtic (Gaussian), leptokurtic (Laplacian) and platykurtic (uniform) distributions are given in Figure 3.6.

[17] The correction factor of excess kurtosis is assumed (-3).

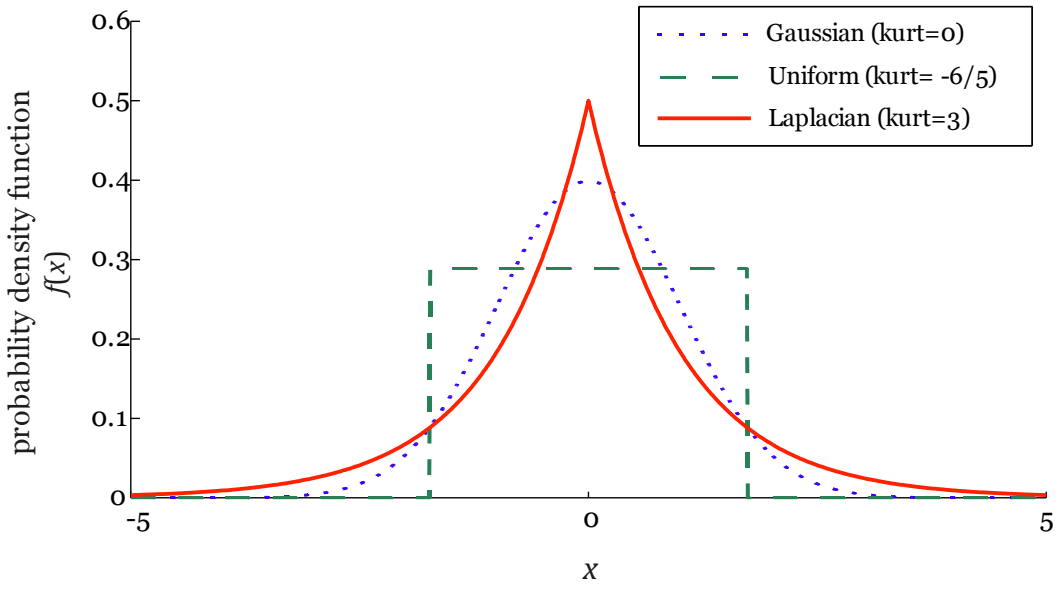

Figure 3.6: Density functions of three typical distributions. The range of values of the selected functions has been selected so that $\mu = 0$ and $\sigma = 1$.

Measuring kurtosis is a computationally light process and is obtained from the fourth moment:

$$kurt(x) = \mathrm{E}[x^4] - 3(\mathrm{E}[x^2])^2. \tag{3.23}$$

However, this method has significant disadvantages when the distribution is computed based on observed samples (rather than a function) and can be very sensitive to outliers, i.e. the estimation may be based only on a few observations from the tail of a distribution. This can result to a misleading measurement [216], because expectations of polynomials such as these of the fourth order are more affected by data that are far from zero than data that are close to zero (or generally the mean). This is often referred to as the *fat tail* problem [217]. For example, a sample of 1000 random values that has unit variance and zero mean, contains one value equal to 10, so the kurtosis equals at least $10^4/1000-3 = 7$, which is very large (in Figure 3.6, the kurtosis of the very narrow curve of the Laplacian distribution is only 3). As a result, kurtosis cannot be absolutely reliable for non-Gaussianity measurement.

3.4.4. Negentropy

Another method of measuring Gaussianity (or to be precise, non-Gaussianity) is the negative entropy, or *negentropy*. In information theory, entropy is a classic mean of measuring the uncertainty of a random variable [218]. Its value is closer to zero when the value of the variable is very predictable, and acquires its maximum value when the result is unpredictable (i.e. all the possible values of the variable have the same probability). For example, a coin that is biased will have a value closer to zero than a fair coin, which will have entropy of one. On the other hand, a fair die roll will have a greater value, because the possible outcomes are six ($\log_2 6 \approx 2.6$).

Formally, entropy $H(X)$ for discrete random variables is given by:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i), \tag{3.24}$$

where x_i are the n possible values of X.

Entropy can be generalised to continuous random variables. This case is called *differential entropy* [219]. Assume a random variable vector $\mathbf{x}$ with a probability density function $pdf(\mathbf{x})$. The differential entropy $h(\mathbf{x})$ is given by:

$$h(\mathbf{x}) = -\int pdf(\mathbf{x}) \log pdf(\mathbf{x}) d\mathbf{x}\,, \tag{3.25}$$

where $d\mathbf{x}$ is the difference between any two adjacent values of $\mathbf{x}$. The base of the logarithm defines the units of the entropy (e.g. if it is $\log_2$ the units are bits).

In contrast to entropy, differential entropy gives a relative measure of randomness and can have negative values: a continuous variable that has high

peak(s) in its probability density shape will have small differential entropy (i.e. negative with large absolute value). It has been proven that a normally distributed variable has the largest entropy among all random variables of unit variance [220, 221]. Therefore entropy can indeed be used as a non-Gaussianity measure.

Negentropy $J(\mathbf{x})$ is the normalised version of differential entropy [222] and is given by:

$$J(\mathbf{x}) = h(\mathbf{x}_{gauss}) - h(\mathbf{x}), \tag{3.26}$$

where $\mathbf{x}_{gauss}$ is a normally distributed random vector and has the same correlation and covariance matrix as $\mathbf{x}$. The significant advantages of negentropy are that it is robust in relation to kurtosis and that it is well justified within the statistical theory. However, it is computationally very expensive as it requires estimation of the pdf. Hence, computationally simpler methods are usually used in order to derive an efficient ICA algorithm.

3.4.5. Approximation of negentropy

As a 'lighter' alternative to the approximation of negentropy, the description in [223], involves the use of higher-order cumulants. For variable x with $\mu = 0$, $\sigma = 1$:

$$J(x) \approx \frac{1}{12}\mathrm{E}[x^3]^2 + \frac{1}{48}kurt(x)^2. \tag{3.27}$$

Unfortunately, this estimation method involves kurtosis and comes with all its aforementioned weaknesses, especially if it is considered that the *skewness* (the 3rd order cumulant) is zero for symmetric distributions. This would make $J(x)$ solely dependent on kurtosis.

Taking into consideration all the aforementioned advantages and disadvantages of measuring Gaussianity, FICA [224] generalises the estimation model (3.27), using non-quadratic functions, defined therein as "non-polynomial moments". The polynomial functions x^3 and x^4 are estimated by other functions: the simplest case with two functions $G_1(x) = x^3$ and $G_2(x) = x^4$ is expressed as:

$$J(x) \approx k_1(\mathrm{E}[G_1(x)])^2 + k_2(\mathrm{E}[G_2(x)] - \mathrm{E}[G_2(v)])^2 \,, \tag{3.28}$$

where k_1 and k_2 are positive constants, and v is a Gaussian variable with zero mean and unit variance. The significant advantage of this model is that, even if the approximation is not accurate, it produces a non-negative value and is consistent, giving zero for Gaussian variables. Furthermore, if this model is considered for symmetrical distributions (i.e. of zero skewness), then it can be rewritten using only one non-quadratic function:

$$J(x) \propto (\mathrm{E}[G(x)] - \mathrm{E}[G(v)])^2 \,, \tag{3.29}$$

where $G(x) = x^4$, and the symbol '$\propto$' indicates proportionality. Although this seems at first like an approximation of kurtosis that would lead to the same problems, $G(x)$ can be changed as follows so that it does not grow too fast [224]:

$$G_1(x) = \frac{\log(\cosh a_1 x)}{a_1}, \text{or} \tag{3.30}$$

$$G_2(x) = -e^{\frac{-x^2}{2}}, \tag{3.31}$$

where $1 \leq a_1 \leq 2$ is often equal to 1. This approximation of negentropy is fast and constitutes a good compromise between the classic methods of measuring Gaussianity by kurtosis and negentropy.

3.4.6. FICA for one component

After examining the approach of FICA on the measures of non-Gaussianity, this subsection describes the steps of the algorithm for the purpose of estimating a vector that maximises the contrast[18] [225] as in equation (3.20). Initially, the algorithm for the estimation of the first IC is described. As described in Subsection 3.3.3, the vector that is important here is **w**. Thus, the algorithm must "test" (or, in technical terms, update by learning rule) several values of **w** in order to maximise the non-Gaussianity of the projection $\mathbf{w}^T\ddot{\mathbf{x}}$.

There are several optimisation methods in order to approach the value of **w**, such as gradient descent [226], Newton iteration [227], and fixed-point iteration [228]. The latter optimisation is employed by FICA, as it is acknowledged as an acceptable trade-off between efficiency and effectiveness [229]. The measure for updating **w** is negentropy, as, the higher the index of negentropy, the further the distribution from Gaussian.

Since the approach is fixed-point optimisation, the corresponding derivatives of equations (3.30) and (3.31), defined as $g_1(x)$ and $g_2(x)$ respectively are used. Adding to these the derivative from kurtosis as described in 3.4.3 results in three choices:

[18] A contrast function F(**x**) is any non-linear function which permutation and scaling tolerant and estimates the level of statistical independence between the components of **x**.

$$g_1(x) = \tanh(a_1 x), \tag{3.32}$$

$$g_2(x) = xe^{\frac{-x^2}{2}}, \text{or} \tag{3.33}$$

$$g_3(x) = x^3, \tag{3.34}$$

where $1 \leq a_1 \leq 2$ is a scalar constant that is empirically set usually to 1 [227]. Table 3.1 gives the description of the algorithm. In practice, the expectations are estimated as an average over the available data sample [230].

1. Centre the data so that $\mu = 0$
2. Whiten the data to give $\ddot{\mathbf{x}}$
3. Choose an initial vector $\mathbf{w}$ randomly with the restriction $\|\mathbf{w}\| = 1$
4. Let $\mathbf{w} \leftarrow \mathrm{E}[\ddot{\mathbf{x}} g(\mathbf{w}^T \ddot{\mathbf{x}})] - \mathrm{E}[g'(\mathbf{w}^T \ddot{\mathbf{x}})]\mathbf{w}$
5. Let $\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\|$
6. If not converged, go back to Step 4

Table 3.1: FICA Algorithm for the first IC

As can be seen in Table 3.1, step 4 is the update of $\mathbf{w}$ by evaluating the projected negentropy as in (3.29). Following each iteration of step 4, $\mathbf{w}$ is normalised, i.e. it is divided by its norm, in order to remain on the unit sphere and keep the projected variance of $\mathbf{w}^T\ddot{\mathbf{x}}$ constant. In other words, restraining the norm of $\mathbf{w}$ to unity is equivalent of restraining the variance of $\mathbf{w}^T\ddot{\mathbf{x}}$ to unity,

because $\ddot{\mathbf{x}}$ is whitened. The function g can be any of the equations (3.32)-(3.34). Convergence in step 6 means that the old and new $\mathbf{w}$ will have the same direction, that is a dot-product equal (or almost equal) to 1. Note that the vector could converge to two points, since $\mathbf{w}$ and $-\mathbf{w}$ have the same direction. Hence, the sign of the IC is ambiguous in ICA (see 3.5).The functions $\breve{g}_1$, $\breve{g}_2$, and $\breve{g}_3$ that correspond to g_1, g_2, and g_3, are given as:

$$\breve{g}_1(x) = a_1(1 - \tanh^2(a_1 x)), \tag{3.35}$$

$$\breve{g}_2(x) = (1 - x^2)\exp\left(-\frac{x^2}{2}\right), \text{ and} \tag{3.36}$$

$$\breve{g}_3(x) = 3x^2 \tag{3.37}$$

The derivation of the above is given in [167].

3.4.7. FICA for more components

The same algorithm can be used in order to estimate different components. The way that this can be implemented is to start from different initial values in step 3 in Table 3.1 in order to obtain different components.

However, the faster approach is that the outputs $\mathbf{w}_1^T\ddot{\mathbf{x}} \ldots \mathbf{w}_N^T\ddot{\mathbf{x}}$ are orthogonalised after each iteration. Since $\ddot{\mathbf{x}}$ is whitened during the pre-processing stage, they can easily be orthogonalised by decorrelation. There are two different ways of doing this, and, because this is an important step of FICA, each way gives considerably different results when applied on the same data set [204]. These two different approaches on orthogonalisation, namely deflationary and symmetric are given in A.2.

3.5 Ambiguities of ICA

ICs can be estimated up to an arbitrary permutation: In order to demonstrate this problem, equation (3.4) is rewritten so that the target is to estimate the columns of the matrix **A** (say $\mathbf{a}_i$):

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{a}_i s_i \,. \tag{3.38}$$

Obviously, the sums in this equation are interchangeable, so there is no possible way of finding the "order" of the ICs. Although this might sound like a minor problem, as essentially there is no such thing as "order" in a recording environment, the difficulty that this weakness of ICA exhibits is that consistency in the results cannot be guaranteed. That is, the same algorithm applied on the same mixtures, may come with different outcomes every time an experiment is performed. Furthermore, if ICA was to be applied on a short-term basis, the IC segments cannot be easily reconnected.

Magnitudes of the ICs cannot be found: In equation (3.4), the matrix **A** (which consists of the a_{ij}) and **s** are unknown. Therefore, the individual s_i (which are scalar) can be cancelled by dividing the corresponding column $\mathbf{a}_i$ of **A** by the same scalar, e.g. a_i :

$$\mathbf{x} = \sum_{i} \left(\frac{1}{a_i} \mathbf{a}_i \right) (s_i a_i) \,. \tag{3.39}$$

Therefore, the mixing matrix can be estimated up to an arbitrary magnitude.

Phase of the ICs cannot be determined: As explained in Subsection 3.3.3 , the vector **w** can have two values that maximise the non-Gaussianity (i.e. **w** and –**w** have the same direction). That means that the sign of the ICs cannot be

determined and the signal might have a rotation of 180°. Again, this would seem unimportant in the case of an audio signal, as the acoustic result is the same no matter the phase of a signal. However, in a similar case of segmenting and reconstructing a signal as mentioned above, this ambiguity can prove troublesome.

3.6 Chapter summary

In this chapter, the BSS problem has been formulated and two possible solutions for this purpose were analysed. These two are closely related as they both implement projections of the de-mixing matrix in order to maximise statistical properties of the signal. This chapter has also attempted to give an in-depth description of the statistical model of Independent component analysis (ICA) with emphasis on FICA, a popular fixed-point implementation. Moving away from the BSS formulation of the source separation problem, the next section describes an approach that is specifically designed for the case of stereophonic recordings and is based on inter-channel differences.

4

APPLIED SOURCE EXTRACTION FROM POLYPHONIC MIXTURES

The field of SVS, which is of particular interest in this thesis, can be seen as a subset of audio source separation. The latter has taken mainly two paths (mono and stereo) depending on the nature of the observed mixture. The title of this chapter, source extraction from polyphonic mixtures, uses the term polyphonic with its dual meaning: polyphony as many individual pitches and sources that are simultaneously present in the same stream [231], and multi-channel that means more than one signals of recorded audio, designed to be listened concurrently. In line with the above, the chapter gives an overview of stereophonic music production and encompasses the polyphonic path of audio source separation with emphasis on a method that is developed for stereo recordings and is important in the context of this thesis, namely the azimuth discrimination and re-synthesis (ADRess).

4.1 Music recording and stereophonic production techniques

Since the first [232][19] music recording of *The Lord's Prayer* in 1884 by Emile Berliner on an Edison cylinder machine [233], music recording techniques have developed dramatically in terms of process and sophistication. Although, initially at least, the focus was on developing and enhancing the media and reproduction of recordings, it soon became apparent that the system of the era (termed retrospectively as *monophonic*) had reached its limitations [234]. Although the term *stereophonic* (from the Greek words *στερεός* (stereos), which means solid, and *φωνή* (phōnē), which means voice [235]) was introduced in 1880 by Graham Bell [236], it was not until the 12th of March 1932 when Leopold Stokowski performed Scriabin's *Poem of Fire* for the purpose of recording the first stereophonic disc [237, 238][20]. Following the initial development stages of this novel technology, the stereophonic reproduction system became the predominant choice for the purpose of listening to music recordings until today.

4.1.1. Stereophonic or binaural?

It is important at this stage to clarify the terms *binaural* and *stereophonic*, because the distinction is not always clear in the literature. A binaural system makes use of two recording sources/microphones, and usually two independent amplifying channels. This process duplicates normal listening. On the other hand, a stereophonic system results in an idiomatic sound pattern at the listener's ears and results in indication of direction which is perceptually located *between* the spatial difference of the loudspeakers [239]. In addition, a

[19] In fact, the very first audio recordings were performed with the *phonautograph* invented by Édouard-Léon Scott de Martinville in 1857. However, these recordings were not supposed to be played back but rather serve as a laboratory measurement of amplitude and waveforms.
[20] It is claimed that stereophonic discs were produced prior to this date. However, none of them has survived.

significant difference between these two systems is the medium that is used for reproduction. In binaural systems, the intended reproducer is a pair of headphones, while in a stereophonic system the equivalent is a pair of loudspeakers (although often listened to on headphones by consumers). The latter system results in the capability of the listener to turn their head and "face" a different source. The use of loudspeakers also introduces the shadow/ghost effect, i.e. each loudspeaker output arrives at both ears of the listener from different angles and at different times, even filtered through the obstacle of the head [240, 241].

4.2 Multi-track recording and stereo mix-down

Although stereophonic and quadrophonic [242] productions coexisted briefly, stereo quickly gained popularity and soon it became the industry standard. In parallel with the growth of stereophonic production, multi-track recording was developed. The motivation behind this type of production was Les Paul, a jazz songwriter who needed to record himself playing multiple instruments. As the machines of the era (late 1950's) did not allow him to accomplish that without significant degradation in sound quality, the company *Ampex* provided him with the first multi-track recorder which he nicknamed *Octopus* [243]. The development of multi-track recording together with the novelty of stereophonic production signified a milestone in the production of music recordings; after the late 1960's, monophonic production had mostly ceased and the majority of commercial music adhered to multi-track recording and stereophonic mix-down[21].

Briefly described, the aforementioned process involves N sources that are recorded individually and are subsequently electrically summed and

[21] Mix-down is the term in audio engineering, when many tracks (considered to be sources) are mixed "down" to fewer channels (e.g. mono, stereo, surround)

distributed across two channels using a mixing console [244, 245]. The spatial localisation (termed *panning* by mixing engineers [241]) between the two channels (which correspond to a left and a right speaker) is achieved by the use of a panoramic potentiometer (usually known as a "pan pot" [246]) that divides a single signal into two continuously variable intensity ratios [247], as seen in Figure 4.1. This method of assigning sources to a left and a right channel constitutes the basis of what is commonly referred to as the *stereo field*.

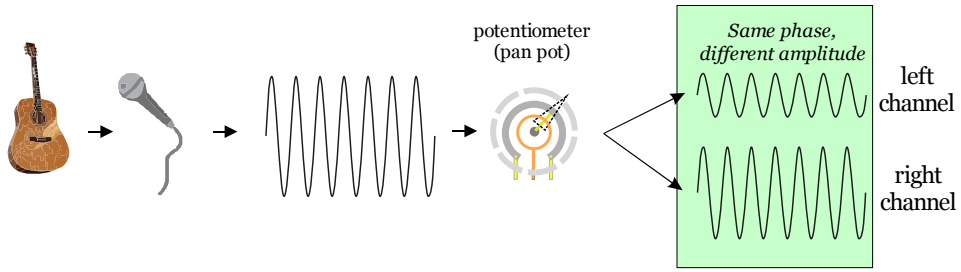

Figure 4.1: Simplest scenario of stereophonic mixing for one source.

Usually, the simple case described above constitutes only a part of the mixing process, as it is common that some linear, as well as non-linear processing is applied. This processing frequently involves frequency filters (known as equalisation), amplitude compression, and reverberation. They can be applied either on the individual sources that are subsequently summed (Figure 4.2), collectively, or both individually and collectively (Figure 4.3) which is the most typical scenario [246].

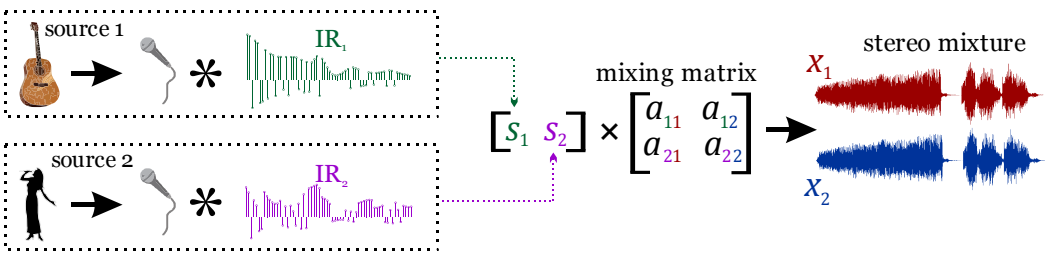

Figure 4.2: Stereophonic mixing for two sources with individual processing. The processing might comprise several functions (e.g. equalisation, chorus effect, reverb) but is represented here as a single impulse response (IR) for each source. The asterisk indicates convolution.

By examining the latter scenario, which can be seen in Figure 4.3, it is observed that the signals of the sources take two paths. One path (usually referred to as an auxiliary bus) consists of a 'monophonic' mixing matrix[22] and its own IR convolution. The other path follows a convolving procedure and gets summed with a 'stereophonic' matrix before the signal is finally convolved with another IR. In this case, the mixing matrix is not square as the number of sources exceeds the number of channels. To be more clear, by the first two scenarios (Figure 4.1 and Figure 4.2) an ideal de-mixing matrix produces the individual sources with their processing without cross-contamination. In Figure 4.3, however, the same does not apply.

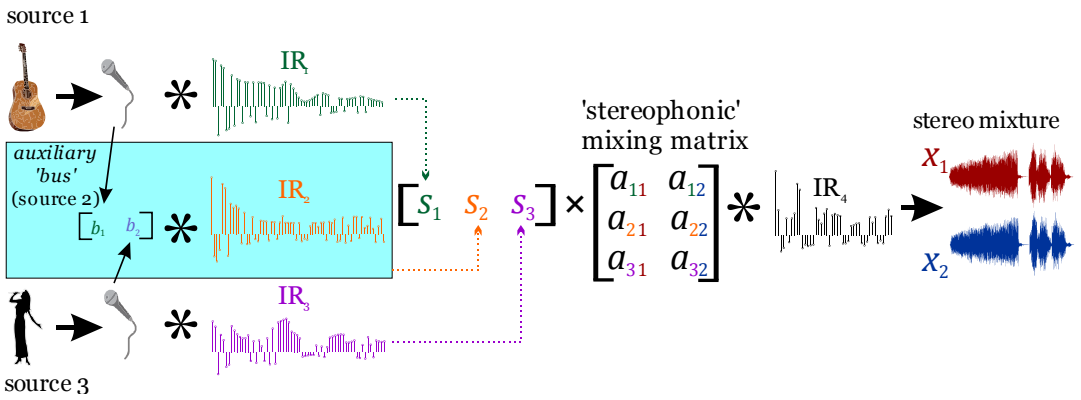

Figure 4.3: 'Real-world' mixing scenario for two sources. The auxiliary bus is created with a 'monophonic' matrix and is treated as a separate source. The additional processing creates, in practice, a very long IR for each source.

4.2.1. Mixing conventions

Over the years of stereophonic mixing, audio engineers have developed certain conventions while mixing and mastering[23] [248]. Such conventions exist, for example, with regards to panning. During the first steps of stereo in the 1960's

[22] In some cases, there are also stereophonic auxiliary buses. The result, however, does not increase the complexity of the mixing scenario, as the output of a stereophonic auxiliary bus is just treated as two sources instead of one.

[23] Audio mastering is usually the final step of audio post-production and comprises the preparation of recorded audio and its transfer from the state of the final mix to a data storage device.

the panning of instruments was either hard left or hard right (e.g. in the first stereo production of *The Beatles* [249]). After this era of experimentation, the panning of the voice is traditionally located at the centre, together with part of the drum set (i.e. the bass and the snare drums) and the bass [240]. The crystallisation of the artificial location of these three components is not arbitrary: On one hand, the vocal as the most important 'instrument' of a music track needs to be equally present in the left and the right channel. This also stems from the use of popular music in various venues. As the audience is not always located in the middle of the stereo field in such places, a vocal drifting from the centre would result in a part of the audience missing the vocal component of a song [240]. On the other hand, bass frequencies (which are also the main component of the bass drum) do not exhibit good localisation due their long wavelengths [57, 250] and, as a result, panning deviation from the centre does not have an effect.

Another convention is that, with rare exceptions, a song's lyrics are intended to be intelligible. Mixing engineers use a number of techniques and processes to manipulate the contributing components and ensure the vocal part is clearly audible. This usually involves dynamic range compression to impose artificial stability in the amplitude of the vocal signal, and filtering to enhance the frequency spectrum of the singing voice in bands where masking occurs [246]. Finally, engineers tend to impose artificial loudness on music tracks (see "loudness war" [251]). One way of achieving this is by adjusting the spectrum of music tracks in order to match the Fletcher-Munson curve of equal loudness (see Figure 2.4), in addition to dynamic range compression.

Having examined the three scenarios of mixing in Section 4.2, the next section describes the method that is termed ADRess, which exploits the inter-channel or inter-aural intensity difference (IID) that naturally occurs in multi-track recordings that are mixed down as stereo.

Although the focus of this section is ADRess as it is important in the context of this thesis, it should be noted that all algorithms that involve the exploitation of IID [252-254] work in a similar manner.

4.3 Azimuth discrimination and re-synthesis (ADRess)

ADRess is an algorithm for source separation that exploits prior information usually found in stereophonic music recordings [255]. The ultimate target of ADRess is to extract one music source from the others. However, in practice, ADRess extracts a panoramic audio *subspace*, rather than a specific music source. To make this clearer, the stereo field of a stereo recording is visualised as a semicircle (Figure 4.4). All music sources included in the recording are distributed across this area.

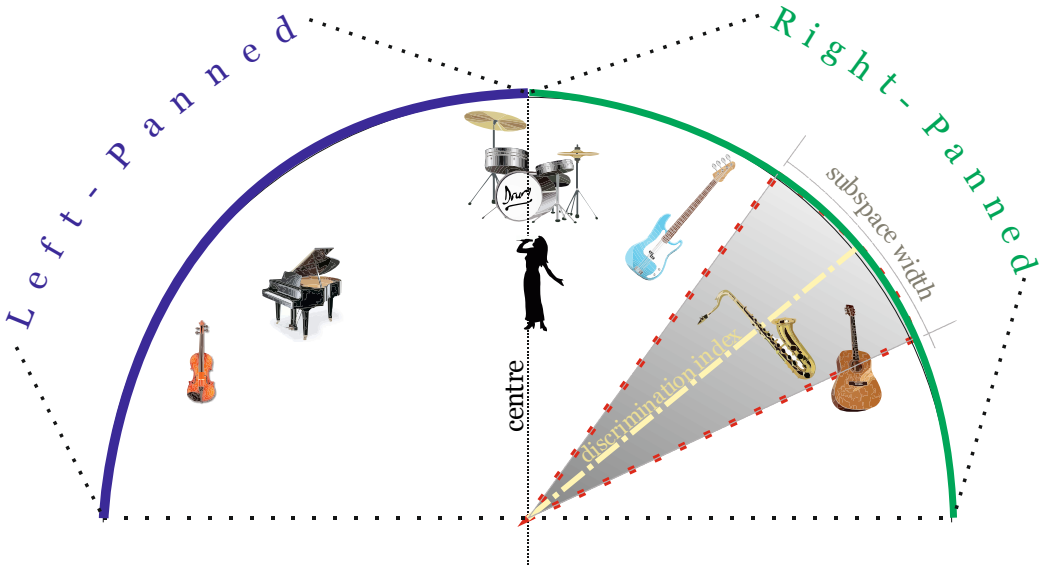

Figure 4.4: Illustration of the functionality of the ADRess algorithm

In geometric representation, ADRess aims to isolate a central angle of this semicircle and, in fact, *every* music source that is contained in this central angle is going to be isolated. In ADRess, this central angle is named the *azimuth*

subspace (illustrated by the greyed out area) and has two parameters. The first parameter is the *discrimination index*, which could be represented in the figure as the bisector of the central angle, and determines the panoramic position (i.e. the angle) of the azimuth subspace. The second parameter is the *subspace width*, which can be represented as the arc of the central angle, and determines the radial span of the azimuth subspace. These two parameters must be set manually by the user by means of trial and error, as ADRess does not have a way of detecting them for the target source.

The figure represents schematically how ADRess functions in stereophonic mixtures. The two parameters (i.e. discrimination index and subspace width) are set so that the saxophone can be isolated from the rest of the music sources. This graphical example also demonstrates the main disadvantage of ADRess: when two or more music sources are "close", in terms of panoramic positioning (panning), ADRess will not be able to completely isolate one from another. For instance, although the aim is the extraction of the saxophone solely, some material (i.e. some of the frequency material) of the guitar is inevitably included. As explained before, this happens because the two music sources are located in close panoramic proximity. In order to avoid the smearing from the guitar frequencies, the subspace width should be reduced. However, this would cause some of the frequency components that constitute the saxophone part to be left out. It should be noted that, in principle, ADRess is not specifically meant for SVS, but rather for general music source separation, as it essentially targets a subspace, which is set by the user, and not a specific source such as the vocal element.

4.3.1. Basic concept of ADRess

Moving on to the theoretical concept that ADRess incorporates, every music source has a panoramic position that can be expressed as an intensity ratio between the two stereo channels [256]. ADRess, exploits this principle that is termed inter-channel intensity difference (IID) [257].For example, the piano in

Figure 4.4 is 75% left and 25% right, the guitar is 80% right and 20% left, and the singer is 50% right and 50% left (i.e. exactly at the centre). Hence, if for example the right channel is multiplied by 25%, the guitar is going to be equally distributed between the left and the right channel (20% left and 20% right). A simple subtraction (left channel – right channel) can then cancel out (i.e. zero) the guitar. The algorithm will then be able to reconstruct the eliminated guitar, because it can pinpoint its frequency components, which are zeroed after the previous subtraction.

There are two fundamental challenges in this concept. Firstly, there is no prior information about the panoramic positioning of the music sources in each song. Hence, the correct factor (in the previous example 25%) that will bring a music source to equilibrium between the two channels is unknown. The user must manually select the position of the targeted subspace, and therefore the correct factor. In practice, this is achieved by means of trial and error. The second challenge is that every music source consists of many frequency components (partials) that are usually shared. For example, the guitar in Figure 4.4 might, at a given time-point of a song, share a partial with the singer. Assuming the panoramic positions mentioned earlier and that the partial is equally shared between these two music sources, its panoramic position will be the average between the positions of the guitar and the singer, i.e. (50+80)/2 = 65% right (and 35% left). So, although the majority of the frequency material of the guitar will be 80/20 right, there will be one partial that will be found 65/35 right. Therefore a 25% right channel gain will not cause the subtraction "left channel – right channel" to cancel out this particular frequency, which means that it will not be included in the reconstruction of the guitar. To tackle this problem, the user input is needed again in order to heuristically select the subspace width, which is effectively the width of the panoramic area that will be extracted.

4.3.2. ADRess on a stereo signal

Formally, the description of ADRess on a stereo signal is as follows: Let $L(t)$ and $R(t)$ be the audio signals in the left and right channels of a commercial stereo recording respectively. These can be expressed as:

$$L(t) = \sum_{j=1}^{J} Pl_j S_j(t) \text{ , and} \tag{4.1}$$

$$R(t) = \sum_{j=1}^{J} Pr_j S_j(t) \text{ ,} \tag{4.2}$$

where S_j are the J independent sources, while Pl_j and Pr_j are the left and right panning coefficients for the j^{th} source. The intensity ratio between $L(t)$ and $R(t)$ for the j^{th} source [258] can be expressed as:

$$g_j = \begin{cases} \dfrac{Pl_j}{Pr_j}, & if\ Pr_j > Pl_j \\ \dfrac{Pr_j}{Pl_j}, & if\ Pr_j < Pl_j \end{cases} . \tag{4.3}$$

The above derives $Pl_j = g_j \times Pr_j$ or $Pr_j = g_j \times Pl_j$. Because $L(t)$ and $R(t)$ are linear combinations of the same independent sources, the j^{th} source could be cancelled out by using one of the following expressions:

$$L(t) - g(j) \times R(t), \quad \text{or} \tag{4.4}$$

$$R(t) - g(j) \times L(t). \tag{4.5}$$

The choice between expressions (4.4) and (4.5) depends on which channel contains the j^{th} source more prominently (i.e. whether Pr_j is greater or smaller than Pl_j respectively). Consequently, $g \in \{0, 1\}$, which is needed to avoid possible distortions in the signal [244]. In case $Pl_j = Pr_j$, either expression (4.4) or (4.5) can be used. Therefore, the equations that are presented hereinafter describing ADRess follow two paths depending on equation (4.3): if the user has determined that the source is louder in the right channel (i.e $Pr_j > Pl_j$), ADRess follows the equations with subscript index "*R*". Otherwise the equations with index "*L*" are followed.

Further to the magnitude spectral information obtained above, g works as a scaling factor, and can help towards the extraction of a target source from one of the channels. The following details how ADRess helps to determine the value of g and eventually recover the source after it has been cancelled out.

Initially, the signals from the left and right channel of a stereo mixture are broken down into segments. Each segment is then shaped using a Hann window before being subjected to a fast Fourier transform (FFT) process, which is given as:

$$Lf(k) = \sum_{t=0}^{N-1} L(t) e^{-\frac{j2\pi}{\mathrm{N}} kt}, \qquad \text{for } k = 0,1,2 \ldots \mathrm{N} - 1, \tag{4.6}$$

$$Rf(k) = \sum_{t=0}^{N-1} R(t) e^{-\frac{j2\pi}{\mathrm{N}} kt}, \qquad \text{for } k = 0,1,2 \ldots \mathrm{N} - 1, \tag{4.7}$$

for the left and the right channel respectively, where N is the number of FFT points, and *j* is the imaginary unit. In [244] it has been recommended to apply a 4096-point FFT at 1024-point intervals.

The determination of g is based on applying different scaling values to the Fourier Transform of one channel and subtracting it from the Fourier Transform of the other channel. This is performed in order to establish which frequency bins get cancelled (zeroed) by different values of *g.* The number of equally spaced scaling values (gains) is termed the *azimuth resolution*, is represented as β, and is related to g as:

$$g(i) = \frac{i}{\beta}, \qquad \text{for } i = 1,2,3 \dots \beta \tag{4.8}$$

The authors in [244] use an azimuth resolution β = 100. Hereon it is assumed that β = 100 unless otherwise stated.

Therefore, depending on which channel contains the target source more prominently, one of the following equations is used to construct a frequency-azimuth spectrogram (defined as *azimugram*):

$$Az_R(k,i) = \mid Lf(k) - g(i) \times Rf(k) \mid \tag{4.9}$$

$$Az_L(k,i) = \mid Rf(k) - g(i) \times Lf(k) \mid \tag{4.10}$$

The azimugram shows, in fact, the frequency bins (i.e. rows k) that get cancelled out (i.e. ≈ 0) at specific scaling factors (i.e. columns i). To demonstrate the concept of the azimugram, let us consider a very simple stereophonic mixture consisting of:

Source 1: A sum of 5 sinusoids of equal amplitude and frequency of 2540 Hz,

5080 Hz, 7620 Hz, 10160 Hz, and 12700 Hz respectively. This is perceived as a tone[24] with fundamental frequency (i.e. pitch) of 2540 Hz, and *Source* 2: A sum of 5 sinusoids of equal amplitude and frequency of 4350 Hz, 8700 Hz, 13050 Hz, 17400 Hz, and 21750 Hz respectively. This is perceived as a tone with pitch of 4350 Hz. These sources are mixed in a way so that Source 1 is distributed 75% left and 25% right and Source 2 is distributed 54% left and 46% right.

In Figure 4.5, the resulting azimugram can be seen as in equation (4.10), for a frame of 2048 samples. The arrows show the null points, where each source gets cancelled out.

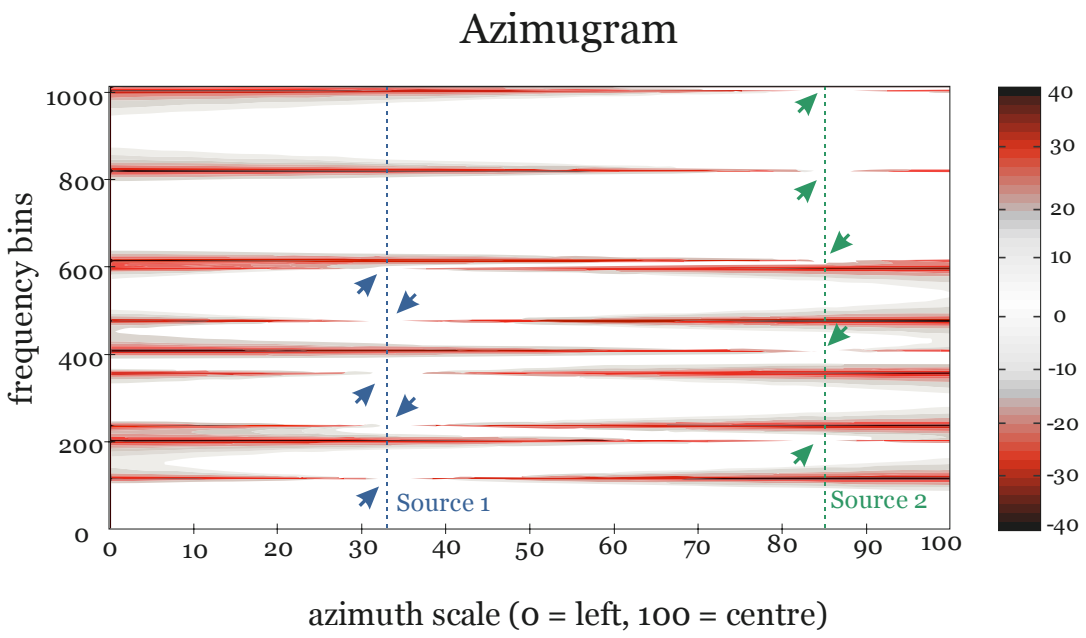

Figure 4.5: Azimugram of the left channel for two non-overlapping sources.

It is shown above that Source 1 would get cancelled out when the STFT of the left channel is multiplied by the value deriving from equation (4.8) for i = 85

[24] It is perceived as one tone because all the sine waves that comprise it have frequencies that are multiples of the fundamental frequency (i.e. 2540 Hz).

(i.e. $g(85) = 85/100$) and then subtract it from the right channel. For Source 2 the respective values is $g(32) = 32/100$. It is worth mentioning that the azimuth scale refers only to the channel that is selected from the equation (4.3) onwards. If, unlike Figure 4.5, the right channel is selected, the value of zero on the azimuth scale will translate to right, while the value of β will translate to centre.

After the computation of the azimugram, the nulls need to be turned into peaks. This procedure will help recovering the target source by using the inverse fast Fourier transform (IFFT). Unfortunately, the magnitude of the peaks is unknown and will need to be estimated.

$$Az_R(k,i) = \begin{cases} Az_R(k)_{max} - Az_R(k)_{min}, & if\ Az_R(k,i) = Az_R(k)_{min} \\ 0, & \text{otherwise,} \end{cases} \quad (4.11)$$

$$Az_L(k,i) = \begin{cases} Az_L(k)_{max} - Az_L(k)_{min}, & if\ Az_L(k,i) = Az_L(k)_{min} \\ 0, & \text{otherwise,} \end{cases} \quad (4.12)$$

for $i \in \{1, 2, 3... \beta\}$, and $k \in \{1, 2, 3 ... N/2\}$,

where N is the number of FFT points. Effectively this process turns the nulls from equations (4.9) and (4.10) into estimated maxima, and sets all other frequency bins to zero. Based on these equations, the magnitudes are reconstructed on the frequency-azimuth plane as shown in Figure 4.6(a).

In an ideal situation, where the sources have no overlapping frequency content, only a single column vector from the above matrix would be needed in order to recover the targeted source by using IFFT. For example, in Figure 4.6(a), where an Az_L matrix with dimensions 1024×100 is shown, Source 2 is obtained from the vector that has an *x* axis value equal to 32, and that includes all of the target frequency bins. However, this case is extremely rare, as the two sources have no overlapping frequency content whatsoever. In the cases where

overlapping harmonics (or any overlapping partials) exist, then these appear as peaks *between* the azimuth location of Source 1 and Source 2. This challenge is demonstrated in Figure 4.6 (b), where the frequency-azimuth domain plane is presented for the case of two synthetic sources with five harmonics each. In this case though, the two sources do share one harmonic, which appears in the middle. The reason is that this particular harmonic is affected by panning coefficients from Source 1 and Source 2 at the same time. Hence, 75% of Source 1 is distributed left and 54% of Source 2 is distributed left. That will mean that their shared harmonic will have an intensity of (0.75+0.54)/2 = 64.5% in the left channel, which subsequently is cancelled when the left channel was multiplied by a $g(55) = 0.55$ ($\approx 0.355/0.645$) in equation (4.10). As a result, this particular peak has a column index $i = 55$ as shown in Figure 4.6(b). Indeed, all songs tend to have overlapping frequency material between their music sources. This problem must therefore be addressed. For this purpose, the *subspace width*, H, is defined, such that $H \in [1, \beta]$.

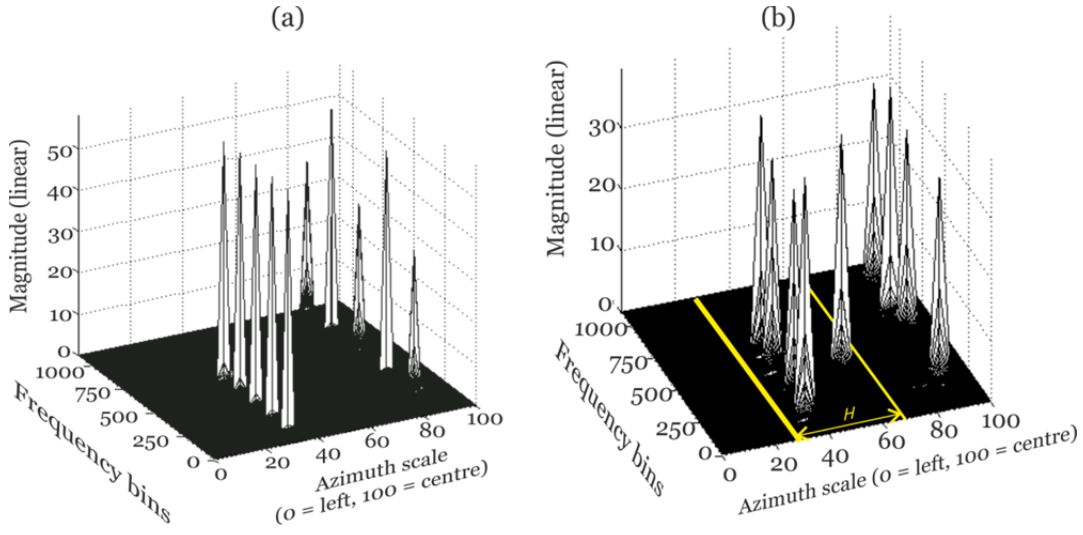

Figure 4.6: Frequency-azimuth plane showing the reconstruction of the magnitudes for two tones (a) without overlapping frequency content and (b) with an overlapping harmonic. The subspace width is denoted by H.

It is evident that, a large subspace width (H) rejects fewer nearby sources, whereas a narrow H includes less frequency material from the target source. In addition to H, the discrimination index d is defined as the centre of the azimuth subspace, so that the subspace spans from d − H/2 to d + H/2. In Figure 4.6 (b), the discrimination index (d) is 48 and the azimuth subspace width (H) is 38. Hence, the selected azimuth subspace spans from 29 to 67.

In practice, the user *manually* selects the values of d and H, by means of trial and error, in order to achieve optimal results. This is one of the inherent disadvantages of ADRess, as there is no provision for the automatic detection of H and d [259][25].

Subsequently, by using H and d, the peaks are extracted, i.e. isolated, from the rest of the frequency-azimuth domain plane, by using one of the following equations:

$$Y_R(k) = \sum_{i=d-H/2}^{d+H/2} Az_R(k,i), \text{ for } 1 \le k \le N \text{ , or} \tag{4.13}$$

$$Y_L(k) = \sum_{i=d-H/2}^{d+H/2} Az_L(k,i), \quad \text{for } 1 \le k \le N \ , \tag{4.14}$$

where $Y_R(k)$ and $Y_L(k)$ are magnitude spectrograms.

From the above equations, a short time magnitude spectrum of the estimated source is obtained. However, as discussed earlier, if two or more sources are panned to the same horizontal location of the stereo field, the obtained

[25] Although methods have been developed for automated detection of H and d, these require an exhaustive search of all possible combination, which is deemed inefficient.

mixtures will contain *all* of them. In other words, if two or more music sources appear in the selected subspace, ADRess will not separate one from the others.

After the equations above, the phase of the signal is also needed, in order to re-synthesise the target source signal using IFFT. In [255], it is claimed that the phase information from the FFT of the original channel—equations (4.6) and (4.7)—is adequate [260]. To restore the phase information, i.e. Φ, polar must be converted to complex form:

$$\Phi_R(k) = \angle(Rf(k)), \tag{4.15}$$

$$\Phi_L(k) = \angle(Lf(k)). \tag{4.16}$$

The real and imaginary parts of the spectrum of the target source signal are estimated as:

$$\hat{S}(k) = \begin{Bmatrix} \mathbb{R}\,\hat{S}(k) = Y(k)cos\Phi(k) \\ \mathbb{I}\,\hat{S}(k) = Y(k)sin\Phi(k) \end{Bmatrix}, \tag{4.17}$$

where $\hat{S}(k)$ is a complex spectrogram. Each time frame is then re-synthesised using IFFT:

$$\hat{s}(t) = \frac{1}{\mathrm{N}}\sum_{k=1}^{\mathrm{N}} \hat{S}(k)e^{\frac{j2\pi}{\mathrm{N}}tk}\ \text{, for } t = 1 \ldots \mathrm{N} \tag{4.18}$$

The re-synthesised time frames are then recombined using a standard overlap and add scheme [261].

As is described in this section, ADRess is an algorithm that usefully exploits prior information especially for the case of commercially produced songs, but has two major drawbacks: the manual user input of critical parameters (i.e. discrimination index (d), subspace width (H), channel where the targeted source is louder), as well as the algorithm's inherent incapability to extract closely positioned sources. The latter can be a significant challenge for the case of SVS, as most of the songs have the singer panned at the centre, together with other music sources, such as parts of the drum kit and the bass.

4.4 Chapter summary

In this chapter, the stereophonic mixing techniques as they generally apply today have been discussed. The chapter has concluded with the review of a method, termed ADRess, which is specifically developed for stereophonic recordings. ADRess, as with many algorithms that make use of IID, is based on the assumption that the sources exhibit same phase characteristics across the two channels of the stereo mix. After reviewing in the previous chapters three different categories of methods for source separation (CASA, BSS, and IID) the next chapter continues to introduce a novel method for SVS.

5

SINGING EXTRACTION THROUGH MODIFIED ADRESS AND NON-VOCAL INDEPENDENT COMPONENT SUBTRACTION

In this chapter, a new approach for the purpose of SVS is presented. The approach, termed singing extraction through modified ADRess and non-vocal independent component subtraction (SEMANICS), combines properties of the azimuth discrimination and re-synthesis (ADRess) method with independent component analysis (ICA). The proposed method is developed and optimised specifically for the case of SVS from stereophonic recordings, which form the majority of commercially distributed music tracks. This chapter also presents the dataset that was developed for the purpose of evaluation. Finally, the experimental investigations that are based on the *bss_eval* [43] metrics are analysed.

5.1 ADRess as the basis of a singing extraction system

As described in Chapter 4, ADRess is an algorithm that exploits prior information especially for the case of commercially produced songs, but has two major drawbacks: the crucial user settings (i.e. discrimination index *d*, subspace width H, and determination of the channel in which the target source is louder), and the algorithm's inherent incapability to extract closely positioned sources. The latter can be a significant challenge for the case of SVS, as most of the songs have the singer panned at the centre, together with other music sources, such as parts of the drum kit and the bass guitar. The next section investigates how ADRess can be modified and optimised for the case of SVS.

5.2 Modification of ADRess

For the case of SVS, a further assumption can be made that will help the azimuth discrimination, and has not been sufficiently considered in ADRess:

As described in Subsection 4.2.1, the lead vocal component of a stereophonically mixed song is traditionally placed at the centre. Although this does not eliminate the need for an azimuth subspace, it is sufficient to minimise user input. To be specific, the upper limit of the subspace as described in 4.3 is set equal to β and the lower limit to β − H. Thus, only one scalar needs to be defined. For example, if H = 4 and β = 100, the value of d(H) will be 98, and the subspace will span from 96 to 100. The difference is that, although the value of d(H) can vary, it is determined solely by the width of H as d(H) = β−H/2, where H and β are by definition even integers. As mentioned before, depending on the value of H, a trade-off applies: a large value includes more vocal components but results in poor separation, whereas a small value provides better separation, but excludes some of the target bins.

The user input in the original ADRess includes the determination of d and H,

and the selection of the channel that contains the target source more prominently. As voice will be equally present in both channels, the latter parameter choice can also be eliminated. In other words, the choice between (4.4) and (4.5) is not necessary, as either is sufficient when the target is the central position of the stereo mix. Thus, user input is reduced.

As described in Section 4.3, one of the main limitations of ADRess is its inability to separate sources that are close in panoramic position. However, the central position of the mix, which is where the vocal element usually exists, is arguably one of the most occupied spaces of the stereo field. In an SVS system, ADRess on its own will only help to isolate the central panoramic subspace. This subspace will include the voice mixed with all the instruments that exist in the centre, but in order to further isolate the voice, further processing needs to be employed.

Having identified the advantages and disadvantages for ADRess towards unsupervised SVS, a new approach is proposed here termed "singing extraction through modified ADRess and non-vocal independent component subtraction (SEMANICS)".

5.3 SEMANICS

SEMANICS is a SVS system that could be better described as a fusion between ADRess and ICA. The modification applied to ADRess for this purpose consists mainly of an approach that is described in this chapter, and is termed "amplitude discrimination". The ICA part of SEMANICS exploits the application of ICA to stereophonic mixtures.

Figure 5.1 shows a schematic representation of the algorithm. The novelties introduced in SEMANICS involve two main stages: amplitude discrimination (which operates based on the threshold computation), and the non-vocal independent component subtraction process, which requires the non-vocal

independent component (NIC) determination shown in the figure.

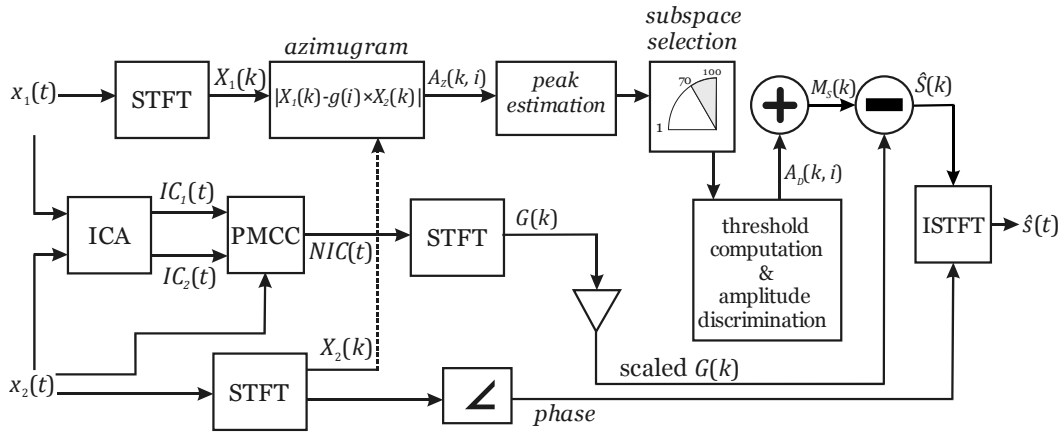

Figure 5.1: Structure of the proposed *SEMANICS* approach to SVS

The following subsections detail the process involved in these two stages.

5.4 Modified ADRess – amplitude discrimination

Equations (4.9) and (4.10) provide a matrix *Az*, whose rows *k* are frequency bins that contain peaks at specific Azimuth values (i.e. columns *i*). Peaks that are near the end of the Azimuth (i.e. ~β) contain the music sources that are at the centre of the original mix. As discussed before, in the considered case, the voice will always appear at the centre of the mix. However, other instruments (e.g. bass, bass drum) are traditionally placed also at the centre of the mix, and the algorithm of ADRess is unable to separate them from the voice.

The herein proposed "amplitude discrimination" (AD) is motivated by the premise that the existence of a singer in a music track often implies that the singing part is the leading music source of the mix. Moreover, the lyrics that are sung usually need to be intelligible, even when the singing voice overlaps tonally with other music sources. Therefore, mixing engineers tend to process the vocal part, such that it is not masked by the accompanying instruments.

The process often includes the enhancement of the frequency ranges of the voice where significant overlap occurs [246]. The aforementioned statements lead to the the assumption that the vocal component will be more dominant than the other music components in the estimated magnitude spectrogram *Az*. By more dominant, it is implied that the magnitude of each of the individual bins that contain the vocal frequencies is generally higher than the mean of the frequency bins within designated frequency bands. Based on this assumption, four amplitude discrimination sub-bands (i.e. 0, 1, 2, 3) are defined. The mean magnitude is then calculated for each of the sub-bands 1-3 (shown as "threshold computation" in Figure 5.1), and only the individual bins that exceed the mean within their corresponding sub-band are extracted. It is worth noting that sub-band zero, in effect, represents a frequency range that the human voice cannot extend to. Therefore, all the frequency bins in sub-band zero are discarded.

The number of sub-bands has been inspired by the four-band dynamic compression (also known as multiband compression) that is applied to most commercial songs during mastering. Preliminary experiments also confirmed this number to be effective.

In order to calculate the threshold for each sub-band, the frequency bins that correspond to the selected sub-bands need to be calculated. This task is performed as follows:

$$\mathbf{U}_0 = [1,2,\ldots \left\lfloor \frac{U_1(start)}{\mathrm{S_R}} \right\rfloor \mathrm{N}-1], \text{and} \tag{5.1}$$

$$\mathbf{U}_m = \begin{bmatrix} \left\lfloor \frac{U_m(start)}{\mathrm{S_R}} \right\rfloor \mathrm{N} \\ \vdots \\ \left\lfloor \frac{U_m(end)}{\mathrm{S_R}} \right\rfloor \mathrm{N} \end{bmatrix}, \text{for } m \in \{1, 2, 3\}, \tag{5.2}$$

where $\mathbf{U}_m$ and $\mathbf{U_0}$ are vectors of integers, m is the index of the sub-bands to be processed, $U_m(start)$ and $U_m(end)$ are the starting and ending frequency of each sub-band m measured in Hz, N is the number of FFT points used, and $\mathrm{S_R}$ is the sampling frequency. Subsequently, the process involves the calculation of the mean average of the magnitudes for the chosen sub-bands 1-3 of the resulting matrix from (4.9) or (4.10):

$$\mu_m = \frac{1}{\mathrm{Q}} \sum_{k \in \mathbf{U}_m} Az_{(k,i)} \quad , \text{for } (\beta - \mathrm{H}) \leq i \leq \beta, \quad \text{and } m \in \{1, 2, 3\}, \tag{5.3}$$

where μ_m is a scalar, i is the discrimination index, m the index of the sub-band, β is the azimuth resolution, Q is the number of elements that are summed, and H is the subspace width. The mean for m = 0 does not need to be calculated as sub-band zero is discarded.

The amplitude discrimination is then applied as follows:

$$A_D(k,i) = \begin{cases} A_z(k,i), & \text{if } A_z(k,i) > \mu_1 \quad \text{and } k \in \mathbf{U}_1 \\ A_z(k,i), & \text{if } A_z(k,i) > \mu_2 \quad \text{and } k \in \mathbf{U}_2 \\ A_z(k,i), & \text{if } A_z(k,i) > \mu_3 \quad \text{and } k \in \mathbf{U}_3 \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for } (\beta - \mathrm{H}) \leq i \leq \beta \tag{5.4}$$

This algorithm functions as a brick wall, allowing only the bins with magnitude higher than their respective sub-band thresholds to pass. It should be noted that the amplitude discrimination is applied after the subspace selection but before the subspace summation (as seen in Figure 5.1).

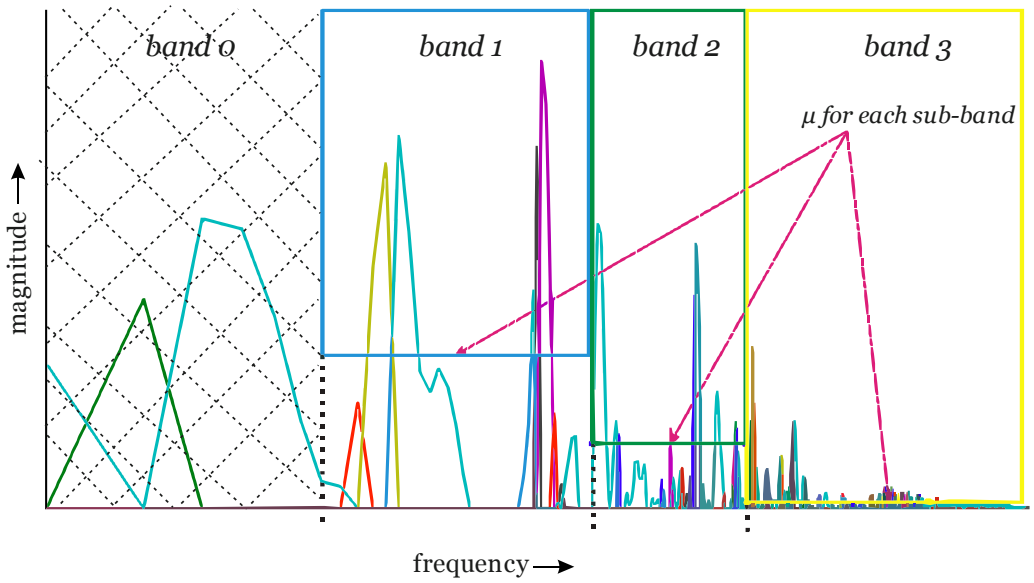

Figure 5.2: Example of amplitude discrimination on the FFT magnitudes. In this example, H = 10, therefore ten different FFT moduli are present in the graph.

The example presented in Figure 5.2 is a magnitude spectrum obtained after selecting a subspace of H = 10 (therefore $d(10) = 95$), for a Hann window of 4096 samples, and overlap of 87.5%. The figure also illustrates how (5.4) operates on a time frame (sampling frequency of 44.1 kHz, sample resolution of 16 bits). The rectangles show the discrimination between the bins that have a magnitude higher than the mean for each of the sub-bands, and are thus included in the estimation of the target source. For each sub-band, the computed mean value operates as a threshold. Each bin that is not included in the rectangles will be set to zero. Here the four sub-bands chosen are: a) 0-0.14 kHz b) 0.14-1 kHz, c) 1-3 kHz, and d) 3-22.05 kHz. In fact the information of the range of the four sub-bands can be limited to their 3 crossing points, in this case [0.14, 1, 3 kHz]. The issue of choosing the crossing points will be discussed in Subsection 5.7.4.

5.5 Non-vocal Independent Component (NIC) Subtraction

The process of amplitude discrimination significantly helps towards the voice isolation, but it is not able to completely filter out all the frequency components originating from other sources. In this section, it is described how SEMANICS uses the Fast ICA algorithm [204] in order to achieve further voice isolation after the amplitude discrimination is applied. Initially, the properties of ICA are exploited in order to obtain a mixture of music instruments from the unprocessed song. Subsequently, this mixture of music instruments is "subtracted", in the frequency domain, from the summed result of the amplitude discrimination process. The motivation for using ICA is as follows.

When the ICA algorithm is applied to underdetermined mixtures, it separates the mixtures into subspaces (in the case of stereo mixture they are two) that are as independent as possible [262]. Some of the source signals will be mainly in the first output while the other sources will find place in the second output [263]. Hence, one of the outputs will contain the vocal element mixed together with some of the sources, while the other will contain only a mixture of the remaining sources, with much less vocal. The latter mixture can be further processed in order to achieve further voice isolation. In the case of this study, the latter is referred to as the non-vocal independent component "NIC".

5.6 Using NIC to further isolate the singing voice

As seen in Figure 5.1, the NIC determination takes place after the fast ICA is applied on the original mixture. As described in Chapter 3, one of the weaknesses of ICA is its ambiguity regarding the order of the independent components. In order to automatically choose which one of the two outputs does not contain the vocal part, each of the ICA outputs is cross-correlated with the original mixture. For this operation, the Pearson product moment correlation coefficient is used [264]:

$$\rho_n = \left| \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \left(\frac{IC_n(t) - \mu_{ICn}}{\sigma_{ICn}} \right) \left(\frac{V(t) - \mu_V}{\sigma_V} \right) \right|, \qquad n \in \{1,2\}, \tag{5.5}$$

where $IC_n(t)$ is the n^{th} ICA output , $V(t)$ are the samples of either the left or the right channel of the stereo mix (depending on which channel has been selected throughout the algorithm), and T is the number of samples in each of $IC_n(t)$ and $V(t)$. μ_{ICn} and μ_V are the sample means of $IC_n(t)$ and $V(t)$ respectively, and σ_{ICn}, σ_V are their standard deviations.

The benefit of using the absolute value of PMCC is that the correlation index has fixed boundaries, i.e. $\rho_n \in [0, 1]$, where the upper limit indicates strong correlation (as in this case it is not significant if it is positive or negative). As the vocal is usually the dominant part of a song, the ICA output containing the vocal will give a higher correlation index, whereas the other (i.e. NIC) outputs a lower value. This is because of the statistical independence of the components as well as the dominance of the vocal part over its accompaniment. The latter is used as follows to enhance the vocal separation process.

Initially, all the columns of the matrix given by (5.4) are added in order to obtain a magnitude spectrogram in one vector, i.e.

$$M_S(k) = \sum_{i=\beta-H}^{\beta} AD_{(k,i)}. \tag{5.6}$$

$M_S(k)$ is a single column vector, and has the same length as NIC. This operation is illustrated as summation in Figure 5.1. Despite the magnitude of the NIC being arbitrary (due to ICA limitations [265]), the magnitude ratio between the sources that are contained in NIC will be similar to that in the original mixture. Therefore, $G(k)$ is defined as the Fourier transform of NIC, and then it is scaled

to match the sample mean of the magnitude spectrum $M_S(k)$. By subtracting the scaled absolute of $G(k)$ from $M_S(k)$, attempts are made to further reduce some of the music sources, i.e.

$$\hat{S}(k) = M_S(k) - \frac{\mu_{MS}}{\mu_G} G(k), \tag{5.7}$$

where μ_{MS} and μ_G are scalars. Subsequently, all the negative elements of $\hat{S}(k)$ are set to zero:

$$\hat{S}(k) = \begin{cases} \hat{S}(k), & if\ \hat{S}(k) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.8}$$

The subsequent procedure in the original ADRess is the signal reconstruction in the time domain using the phase information from the original mixtures together with the magnitude spectrum $\hat{S}(k)$.

5.7 Experimental investigations

During the initial testing of SEMANICS on commercial songs, subjective evaluations showed significant improvement over the ADRess algorithm. However, in order to demonstrate this improvement objectively, this section presents the evaluation method adopted, as well as the dataset used.

5.7.1. Evaluation metrics

In any attempt to separate the singing voice from a song, there is the inevitable challenge of finding an objective testing method, and comparing a new algorithm's performance with that of existing ones. The *bss_eval* system proposed by [43] appears an appropriate choice, not only because it is targeted

specifically to source separation, but also because its results are read in three different values, namely source to distortion ratio (SDR), source to interference ratio (SIR), and source to artefacts ratio (SAR). The downside of this approach is that it requires both the clean source track—in this case the isolated singing voice track, hereinafter *a capella*—and the music accompaniment track, hereinafter *instrumental*.

The *bss_eval* metrics system takes the estimated source $\hat{s}_j$, the *a capella* and the instrumental as input, and decomposes $\hat{s}_j$ as follows:

$$\hat{s}_j = s_{\text{target}} + e_{\text{interf}} + e_{\text{artef}}\,, \tag{5.9}$$

where e_{interf}, e_{artef} are the interference, and error terms respectively. The additional advantage of *bss_eval* is that it allows for a time-invariant gain deformation of the s_{target}, as matching the gain of the input source is usually not important.

The measures (expressed in dB) that are used to evaluate the quality of the separated source are computed as follows:

Source to distortion ratio:

$$\text{SDR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}+e_{\text{artef}}\|^2}\,, \tag{5.10}$$

Source to interference ratio:

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}\ \text{, and} \tag{5.11}$$

Source to artefacts ratio:

$$\mathrm{SAR} = 10\log_{10}\frac{\left\|s_{\mathrm{target}}+e_{\mathrm{interf}}\right\|^2}{\left\|e_{\mathrm{artef}}\right\|^2}. \tag{5.12}$$

The SIR and SAR, can be regarded as valid performance measures with regards to two different goals, namely the rejection of interferences, and the absence of "burbling" artefacts (also known as "musical noise") respectively. The SDR can be seen as a global performance measure [266, 267]. It should be noted that, recently, a modified version of *bss_eval*, called *bss_eval_images* [268] includes an additional factor, namely the source image to spatial distortion ratio (ISR). The ISR is of little significance to separation, but is important for applications that use phase cancellation (e.g. karaoke) [268]. Furthermore, the gain of the estimated output plays a significant role on ISR. Therefore, it has not been used in this study.

5.7.2. Dataset

As discussed above, *bss_eval* requires a set of pre-existing data, namely the music accompaniment and the a cappella track, which are very difficult to retrieve for commercial music. However, artists have recently started to release their songs online in a multi-track form, prompted by the growing interest for music remixing. Although these multi-tracks are sufficient for the assembly of a stereophonic mixture, an *a capella*, and an instrumental track, this assembly is a laborious procedure and the amount of samples that a customised database can contain is subject to human resources and time availability.

As there is lack of a widely available dataset that fulfils the requirements of the *bss_eval* metrics for the case of SVS, a database comprising songs that are available in the above mentioned multi-track format is created for the purpose

of evaluating the proposed system. Two of the songs [269, 270] were taken from [268] while most of the multi-tracks are licensed under CC BY-NC-SA 2.5 and 3.0 [271, 272], and can be acquired from the World Wide Web. During the mixing and mastering process, common types of convoluted reverberation as well as equalisation and compression. The details of the song excerpts [160, 269, 270, 273-280] are presented in Table 5.1.

Title	Artist	Duration (s)	Year	Genre	Format
Salala	Angelique Kidjo Peter Gabriel	8.04	2007	Afrobeat / reggae / worldbeat	PCM (44.1 kHz, 16 bit)
Nude	Radiohead	13.80	2008	Experimental rock	PCM (44.1 kHz, 16 bit)
Kunlarim Sensiz	Sevara Nazarkhan	7.99	2007	Uzbek folk	PCM (44.1 kHz, 16 bit)
Help Me Somebody	Brian Eno David Byrne	8.91	1981	Experimental / art rock	PCM (48 kHz, 16 bit)
Only	Nine Inch Nails	8.30	2005	Industrial rock / electronica	PCM (44.1 kHz, 16 bit)
Resistencia	Los de Abajo	6.03	2005	Reggae	PCM (44.1 kHz, 16 bit)
Tu Vuò Fà L'Americano	Renato Carosone Nicola Salerno	7.70	2007	Jazz / swing	PCM (44.1 kHz, 16 bit)
Shock the Monkey	Peter Gabriel	4.65	1982	New wave	MP3 (160 kbps, 44.1 kHz, 16 bit)
Roads	Bearlin	14.00	2009	Rock	PCM (44.1 kHz, 24 bit)
Que Pena / Tanto Faz (1st excerpt)	Tamy	13.00	2007	Bossanova	PCM(44.1 kHz, 24 bit)
Que Pena / Tanto Faz (2nd excerpt)	--//--	23.00	--//--	--//--	--//--
Don't Know[26]	Suicide Sports Club	7.26	2005	Indie Hip Hop	PCM (44.1 kHz, 24 bit)

Table 5.1: Description of the dataset. All the excerpts are stereophonic.

[26] The multi-track was kindly offered by Bruce Aisher for the purpose of SVS evaluation.

5.7.3. Baseline

Extensive experiments are performed on the aforementioned dataset, measuring the performance of original ADRess and SEMANICS. Initially, *bss_eval* is used to evaluate the baseline performance of the system. This is necessary as the vocal to music ratio presents significant variation between different songs. This ratio can also be regarded as one of the "difficulty" factors that each song presents in terms of vocal isolation, i.e. a lower ratio means that the vocal is more difficult to isolate. For this purpose, the baseline performance is shown in Figure 5.3.

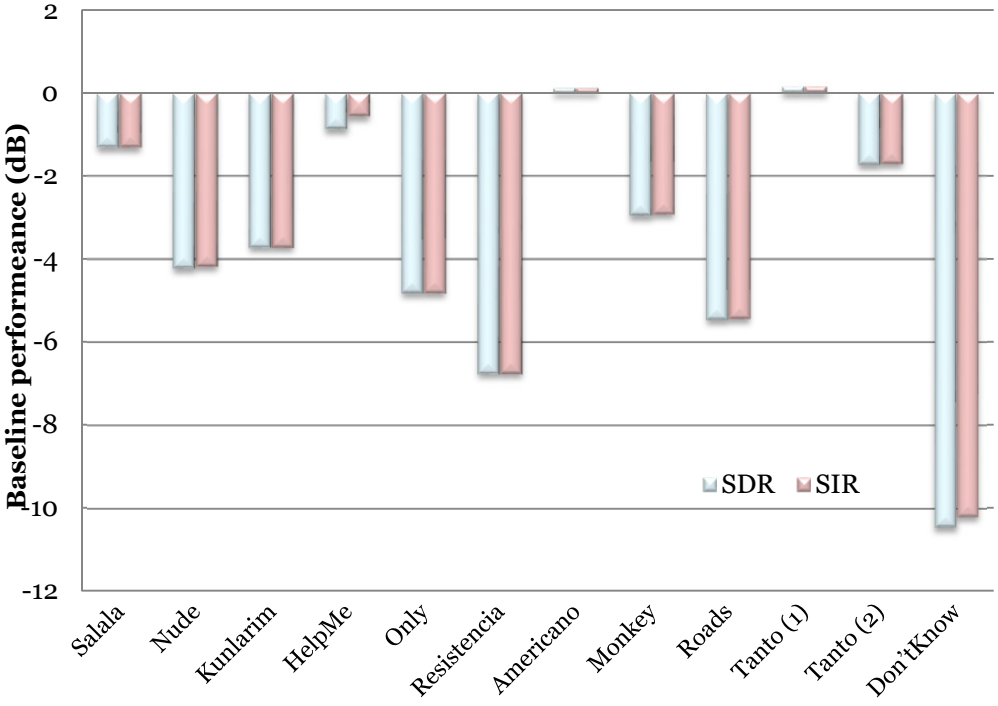

Figure 5.3: Baseline performance of the dataset, i.e. SDR and SIR when no separation is performed.

The results are essentially a signal-to-noise ratio (SNR), where *signal* is represented by the vocal part and *noise* by the music accompaniment. Yet, in order to be consistent with the rest of the experiments, *bss_eval* was also used for measuring the baseline, resulting in the three aforementioned metric

factors, i.e. SDR, SIR, and SAR. The baseline with *bss_eval* is obtained in practice by replacing the estimated source $\hat{s}_j$ in (5.9) with the original mixture. As there cannot be any artefacts (since there is no processing on the mixture during this experiment), SAR is expected to be infinite and therefore is not shown in the figure. However, as can be observed, the SDR and SIR present small fluctuations, which are attributed to the algorithm of *bss_eval* not being able to fully distinguish the difference between interference and artefacts, as well as rounding errors. In other words, taking into consideration (5.10)-(5.12), the SDR should match the SIR when there are no artefacts. The differences shown in the figure, however, are minute and are not considered to significantly affect the evaluation of the results.

5.7.4. Experimental setup

In order to observe the overall performance of the combined algorithms and determine the dependence of improvement upon different algorithm settings, experiments are conducted using all permutations of the settings below:

1. Azimuth subspace width (H), set to 4, 10, 20, and 30
   Four values are chosen, ranging from very narrow width (i.e. H = 4) to a wide/relaxed one (H = 30). This approach will help to examine the relationship between the azimuth subspace width and the amount frequency content which is excluded.

2. Window size set to 512, 1024, 2048, 4096, 8192, and 16384 samples
   In [255] a window size of 4096 is chosen. However, as the window size will directly influence time-frequency transformations, different window sizes are tested in this experiment.
3. Window overlap of 75%, and 87.5%
   The overlap process is an essential component of the time-domain windowing before applying FFT to the raw signal, and is also used prior to the "overlap and add" procedure deployed for audio re-synthesis. In

the original ADRess an overlap of 75% was used. In this study, the potential benefit of an increased overlap (i.e. 87.5%) is also examined.

4. Three different choices of sub-band sets: crossing points in the individual cases are [0.14, 1, 3], [0.14, 2, 5], and [0.1, 4, 8], all in kHz. In each case, the three crossing points provide four sub-bands. For instance, in the case of [0.14, 1, 3], the sub-bands are 0-0.14 kHz, 0.14-1kHz, 1-3 kHz, and 3 kHz to the end of frequency spectrum.

The first choice of frequency sub-bands is based on Sundberg's work [49, 61, 281]. As described in Section 2.4, there is a change in the energy of the singing voice at around 1 kHz and at around 3 kHz. Thus, the first set of crossing points are based on this observation. The other two sets of crossing points are empirically set. The main purpose of the testing with different permutation of parameters is to establish if the system can work for all the songs with these settings fixed without losing significantly in performance.

5.7.5. Results

According to the experimental results, the decreased overlap size (i.e. 75%) does not lead to a significant improvement in the outcomes (i.e. - 0.1 dB approx.), compared with the 87.5% overlap, although the former enables slightly faster processing. The discrimination bands as well as the azimuth width proved to be idiomatic, in the sense that the best results are achieved when these parameters are adjusted individually for each song.

However, in order to directly evaluate the performance of SEMANICS against ADRess, the same parameters for both systems are presented here. These are the settings with which ADRess provided the most satisfactory results, namely azimuth subspace width, i.e. $H = 20$, window size of 4096 overlapping for 3584 samples, and the sub-band crossing points set at [0.14, 1, 3] kHz. Nonetheless, independently of the settings, the proposed system showed major improvement in SIR and significant improvements in SDR over the original ADRess.

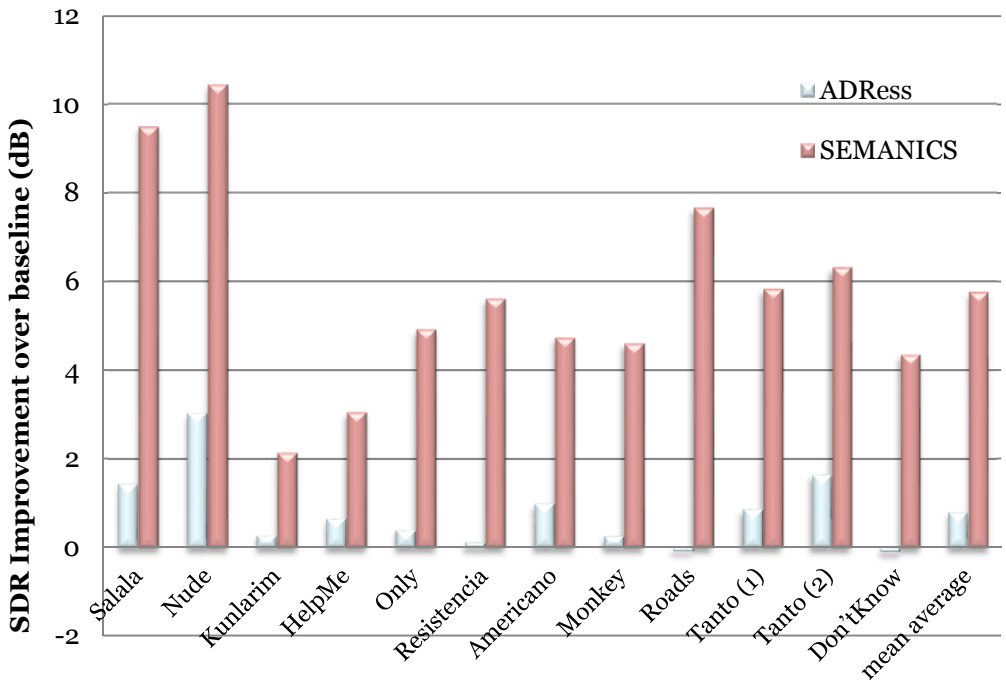

Figure 5.4: SDR performance of ADRess and SEMANICS

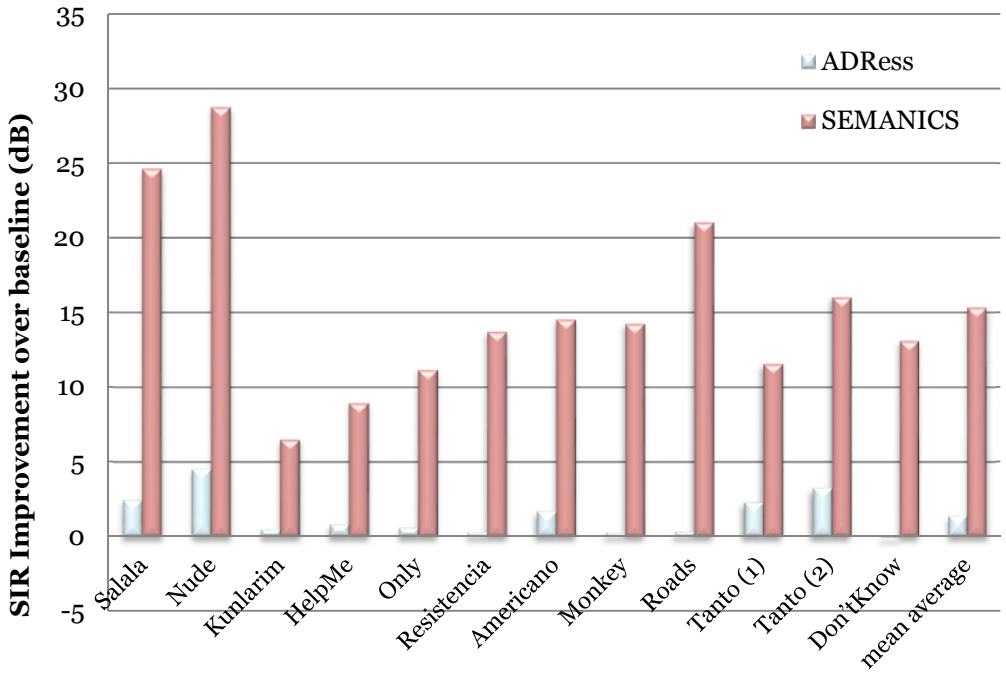

Figure 5.5: SIR performance of ADRess and SEMANICS

In the figures above, SEMANICS is compared with ADRess in terms of SDR and SIR. The metrics are shown in terms of improving isolation over the baseline (Figure 5.3). In these experiments, the only allowed deformation of s_{target} as in (5.9) is a time invariant gain.

For presentation reasons, the illustrations in this section offer only the measurements with the parameters mentioned in the start of this section. The fluctuation of the results according to these parameters and a way to tackle this will be discussed further in Chapter 6.

It is clearly shown that the proposed method increases the SDR and SIR for all the songs of the dataset. The average SIR improvement over ADRess and the baseline (effectively represented in the figures as the 0 dB line) achieved with SEMANICS are seen to be around 14 dB and 15 dB respectively. For the case of SDR the improvement over ADRess and baseline are approximately 5 dB and 6 dB respectively. The full table of numeric values for these figures is shown in the appendix (Section A.3). It should be noted that the relatively poor performance of ADRess (in terms of SDR and SIR) is mainly due to the target subspace being densely populated by various music sources, as well as the reverberant character of the mixtures. This is thought to be an inevitable situation in a variety of commercially produced recordings. In addition, ADRess is not designed to work in an unsupervised way, i.e. the parameters should be set by means of trial and error for each song.

Figure 5.6 shows that ADRess produces fewer artefacts compared with SEMANICS. This is expected as ADRess fails to perform any significant separation (see Figure 5.4), effectively leaving most of the signal intact, and therefore not producing any artefacts. This is also supported by the fact that in songs where ADRess achieves its best performance in terms of separation (e.g. *Nude* and *Salala*), the respective SAR is also relatively low. In contrast to the previous figures of SDR and SIR the values in the figure of SAR are not relative to the baseline; as mentioned earlier, the baseline is—in theory—infinite.

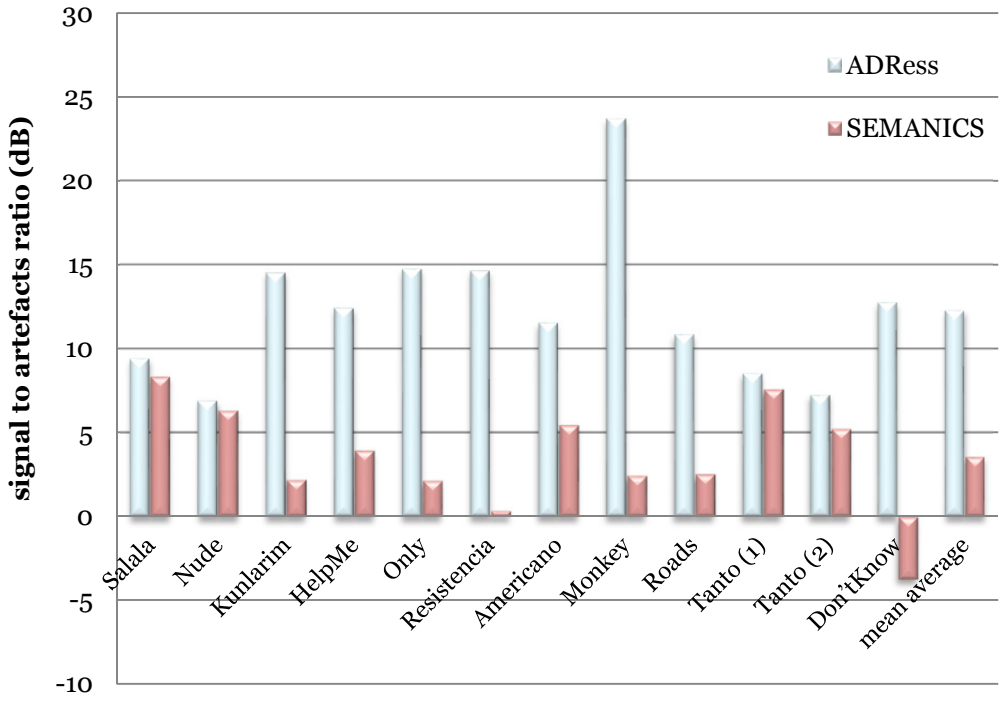

Figure 5.6: SAR performance of ADRess and SEMANICS

It should be stressed that SDR and SIR are much more important to the objective of this study, which is that of audio separation and minimising interference (i.e. achieving a high SIR). This is supported by the study in [266], where the metrics of SDR and SIR are found to be correlated well with the perception of source separation. A high value of SIR has proven to facilitate certain important areas of MIR, such as singer identification [16].

The results for the songs that are taken from the first stereo source separation evaluation campaign (i.e. *Tanto* (1) and *Roads*) compare very well with the rest of the algorithms that were tested in [268, 282]; a comparison is provided in in Section A.3 of the appendix.

5.8 Chapter Summary

In this chapter, a novel algorithm for the separation of singing voice (vocal component) from the accompanying music has been introduced. The proposed method, termed SEMANICS, is specifically for commercially produced stereophonic recordings. It is based on the fusion of independent component analysis (ICA) with a modified version of ADRess. The modification of ADRess is through the incorporation of an appropriate amplitude discrimination procedure.

The experimental evaluation has been based on the use of the *bss_eval* approach together with a varied dataset. The experimental results have clearly illustrated the superior performance of SEMANICS over ADRess in terms of SIR and SDR.

The next chapter involves further exploitation of ICA principles in order to improve the vocal separation effectiveness, and to reduce the dependence on the setting of parameters in the system.

6

SINGING EXTRACTION THROUGH MULTIBAND AMPLITUDE ENHANCED THRESHOLDING AND INDEPENDENT COMPONENT SUBTRACTION

The previous chapter described a system that is based on *azimuth discrimination and re-synthesis* (ADRess) and can extract the singing voice from reverberant stereophonic mixtures. This chapter details an extension to the previous method that is not based on ADRess and exploits both channels of the stereo mix more effectively. In addition, the exclusion of ADRess renders the proposed system less susceptible to performance fluctuation due to parameter settings. For the evaluation of the system the same dataset and evaluation method is used, which enables a direct comparison of the two systems.

6.1 Motivation

As described in Section 4.3, ADRess achieves separation by exploiting the inter-channel intensity difference (IID) that occurs in stereophonic studio recordings. However, ADRess as well as most of the stereo methods that utilise either IID or inter-channel phase difference (IPD) face significant difficulties when processing reverberant mixtures [283]. In addition, because of its supervised nature, ADRess relies heavily on user-determined settings. As a result, SEMANICS, which was introduced in Chapter 5 and is based on ADRess, suffers from the same dependence on parameters. This dependence was mentioned in the previous chapter, and it was observed after testing SEMANICS with permutations of the parameters as described in Subsection 5.7.4. The influence of various parameters on the performance of SEMANICS can be observed in Figure 6.1.

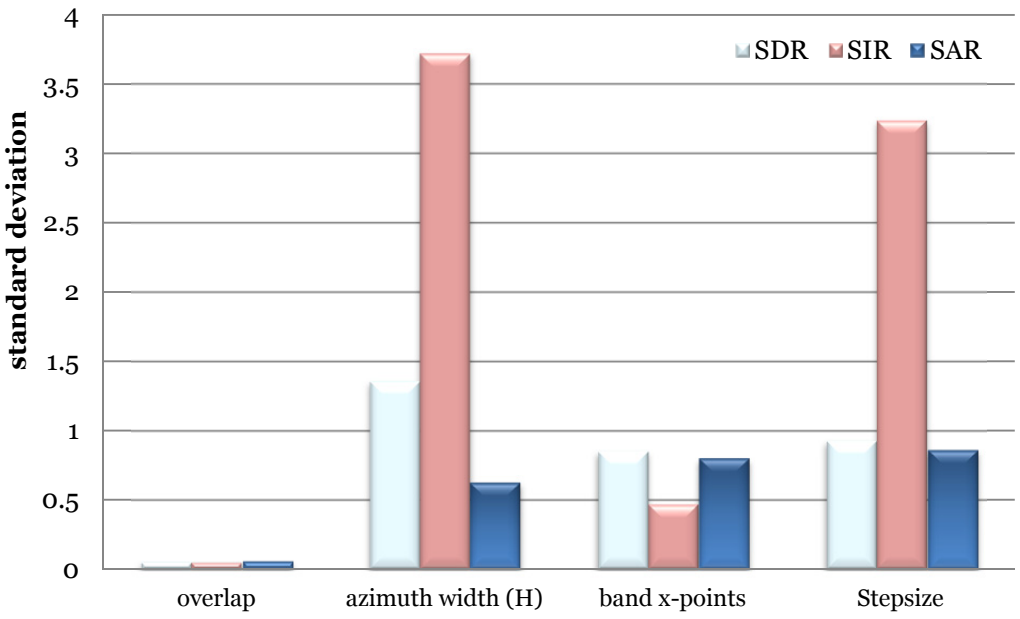

Figure 6.1: Effect of different parameters

The standard deviations in the figure represent the variation that is observed in the metrics when one of the four parameters changes while the others remain fixed. In particular, it can be observed that different overlap settings do not affect the results, while the selection of cross-points of bands presents some, but not extensive fluctuation. Conversely, azimuth width and stepsize (i.e. window size) alter the separation results considerably, especially with respect to interference rejection.

On the one hand, the explanation for the variation that the window size introduces to the metrics stems from the inherent assumption of STFT. In essence, STFT defines a window during which the signal is expected to be quasi-stationary. Having a large window voids this assumption, while a very small window introduces significant spectral leakage. The process of ADRess is hindered by this leakage as it finds false "null" points during the azimugram construction, i.e. (4.9)-(4.10). Therefore, bins are cancelled (and subsequently reconstructed) in positions that appear only due to spectral leakage. On the other hand, the azimuth width is the arc size of the central angle that is extracted from the panoramic hemisphere (see Figure 4.4) and thus defines the actual 'points' on which ADRess applies the subtraction. Hence, both of these parameters that significantly affect the performance of the system are attributed to ADRess. Furthermore, approaches that attempt the automated estimation of the azimuth width require an exhaustive search of all possible combinations for all sources [259].

In order to circumvent the above challenges, a system is introduced here which is a modified version of the algorithm described in Chapter 5, and involves removing the ADRess part. In addition to running unsupervised and utilising the novel approach of non-vocal independent component (NIC) subtraction that was introduced previously, the modified system presented here exploits both channels much more effectively. The new algorithm termed SEMANTICS (singing extraction through multiband amplitude enhanced thresholding and independent component subtraction).

6.2 Overview of the proposed method

As can be seen in Figure 6.2, the proposed system splits the original mixtures into two streams for each channel. The first stream is subjected to Fast ICA and subsequently the non-vocal independent component (NIC) is determined with the help of cross-correlation (PMCC). The second stream is transformed to the frequency domain after it is filtered with a high-pass filter. The NIC is also transferred to the frequency domain, scaled and subtracted from both the STFTs of the processed mixtures. The threshold estimation is calculated from both the channels in order to provide a threshold (i.e. Z_m) to perform the amplitude discrimination. Finally, the signal is transformed back to the time domain after a binary mask is applied with the help of the right and the left channels.

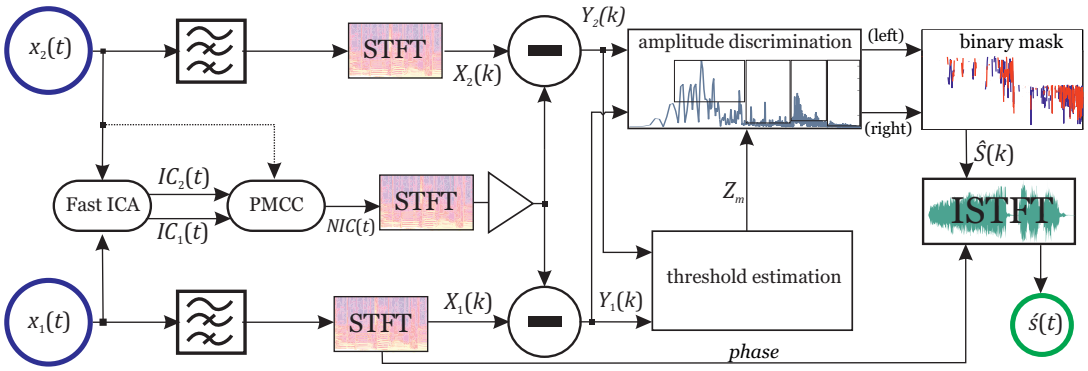

Figure 6.2: Overview of the proposed system (SEMANTICS)

For the purpose of ISTFT, the phase information from the original mixture is used. The rest of the chapter describes the mechanics of the modified algorithm, with reference to the parts that are replicated from the system in Chapter 5.

6.3 NIC subtraction as pre-processing

The previous system, i.e. SEMANICS can be summarised as the application of modified ADRess, followed by the amplitude discrimination and the NIC subtraction. The amplitude discrimination helps significantly towards the voice separation; however, it is the modified ADRess that enabled the latter, as it produces signals that have 'enhanced' presence/dominance of the vocal part. In order to remove ADRess but still retain the efficiency of the amplitude discrimination, the NIC subtraction takes place immediately after the FFT.

In order to be able to perform the NIC subtraction, the process of SEMANTICS starts with the NIC determination (like in the case of SEMANICS). The process is thoroughly detailed in Section 5.6. In brief, it involves the application of Fast ICA to the two channels of the original stereo mixture in order to acquire a time signal whose vocal content is less than either of the original channels. The latter mixture, referred to as NIC, is determined by cross-correlating (using PMCC) each of the ICs with one of the two original mixtures.

Following the previous process, the right, i.e. $x_1(t)$, and left, i.e. $x_2(t)$ channel of the original mixture are subjected to a high-pass filter, for reasons outlined below. Subsequently, $x_1(t)$, $x_2(t)$, and NIC(t),are segmented and transferred to the frequency domain, using a Hann window of 4096 samples at 512-point intervals (i.e. 87.5% overlap). This value of overlap was chosen as it provides slightly better results, while the additional computational cost is not of concern at this point.

6.4 Filtering requirements

Due to the gain tolerance of the de-mixing matrix of ICA [227], the magnitude of G(k) (modulus of the Fourier transform of NIC) is unknown at this stage. In SEMANICS, this problem is tackled by scaling G(k) to the output of the

amplitude discrimination. Since the NIC subtraction is performed here as a pre-processing, the scaling procedure is modified.

Based on the premise that the relative magnitudes of the sources captured in G(k) are similar to those in the original mixture, an intuitive solution would be to scale G(k) in order to match one of the frequency transforms of the original mixture. However, due to the way that ICA operates on complex mixtures, it usually cancels out most low frequency components in the output that correlate poorly with the original audio mixture (i.e. NIC). Furthermore, in audio mixtures, the lower region of the frequency spectrum usually contains most of the energy in the mix. In fact, this is closely in line with the sensitivity of the human ear to different frequencies (i.e. lower sensitivity to lower frequencies [284]).

This means $G(k)$ will not contain much bass frequency. Therefore, this approach would bias the scaling factor towards a smaller value. In order to tackle the aforementioned problem, the proposed method subjects the signals $x_1(t)$ and $x_2(t)$ to high-pass filtering in order to remove components that occupy the lowest part of the frequency spectrum. This filtering is not expected to cause significant loss of the vocal component, as vocal parts are typically high-pass filtered during the mixing process in commercial recordings. During initial investigations, the cut-off frequency providing a good compromise between correct scaling and voice is around 140 Hz. The adopted high-pass filter is a first order IIR system with maximally flat magnitude response (i.e. Butterworth) because of its computational efficiency. The cut-off frequency of 140 Hz is not considered to cause significant degradation to the voice as the pitches below that (i.e. E_2 to $C\sharp_3$) are deemed rare.

After the high-pass filtering and the STFT process, the NIC subtraction proceeds as follows: $G(k)$ is scaled in order to match the mean of the magnitude spectrum $X_i(k)$. By subtracting the scaled $G(k)$ from $X_i(k)$, an initial reduction of the music sources is achieved as follows:

$$Y_i(k) = X_i(k) - \frac{\mu_{Xi}}{\mu_G} G(k), \qquad \text{for } i = 1, 2, \tag{6.1}$$

where i is the channel index, μ_{Xi} and μ_G are the means of $X_i(k)$ and $G(k)$ respectively, and k is the index of the FFT bins, while $Y_i(k)$ is the magnitude spectrum of the processed signal that contains a mixture of the voice and the reduced music sources. It has been observed that, because of the impurity of $G(k)$, (6.1) can sometimes produce negative values for $Y_i(k)$. In such cases, a very low positive value is assigned to $Y_i(k)$.

6.5 Mel sub-band characteristics

The amplitude dominance of the voice is evident in the obtained magnitude spectrogram following the NIC subtraction, especially since many of the music sources are reduced by the aforementioned process. Hence, it is assumed that the magnitude of each of the individual bins that contains the vocal frequencies is generally higher than the mean of the frequency bins within designated frequency bands. This is similar to the assumption made in Chapter 5.

Based on this assumption, M amplitude discrimination sub-bands are defined. In order to limit the parameters that the system depends on, the cross-points are not manually set. Instead, preliminary investigations suggested that an acceptable trade-off is the selection of crossing points such that each sub-band spans an equal number of mels. Thus, the set parameters of the system have been limited to only one scalar, namely M. In this case, $3 \leq M \leq 5$ is found to lead to satisfactory results.

Formally, the thresholds are computed based on both spectrograms that are obtained after the NIC subtraction:

$$Z_m = \frac{1}{\mathrm{Q}} \sum_{i=1}^{2} \sum_{k \in \mathbf{b}_m} Y_i(k,l), \qquad m \in \{1,2,\dots \mathrm{M}\}, \tag{6.2}$$

where Q is the number of elements that are summed, **b** is the sub-band vector, and i is the channel index. It is noteworthy that the threshold is calculated across both channels and not individually in contrast to SEMANICS. This provides a much more accurate estimation of the thresholds, since it evens out peak transients of music sources that might occur in one channel but not the other.

In addition to the M number of sub-bands, a sub-band $\mathbf{b_0}$ is defined, such that it matches the frequencies that were attenuated during the high-pass filtering, but were re-introduced due to STFT errors.

Subsequently, the amplitude discrimination is applied to the result of (6.1). In this system though, this process results in a binary mask, allowing only the bins with magnitude higher than their respective sub-band thresholds across both channels to pass. The bins that do not pass this threshold are set to a very low value (i.e. $0 < \varepsilon \ll 1$) instead of zero. This is also in line with the spectral subtraction method in [285]. Formally, the resulting spectrogram $\hat{S}(k,\ l)$ contains the averaged bins from the two channels as follows:

$$\hat{S}(k,l) = \begin{cases} \frac{1}{2}\sum_{i=1}^{2} Y_i(k,l) & \text{if} \begin{cases} Y_1(k,l) > Z_m \\ Y_2(k,l) > Z_m \\ k \in \mathbf{b}_m \end{cases} m \in \{1,2,\dots \mathrm{M}\} \\ \varepsilon, & \text{otherwise,} \end{cases} \tag{6.3}$$

where ε represents the machine epsilon (in this case $\varepsilon = 2^{-53}$).

Figure 6.3 presents the binary decision (i.e. binary masking) that occurs on the left-hand side of equation (6.3) on the FFT modulus of a Hann-windowed frame of 4096 samples from a song for M = 3.

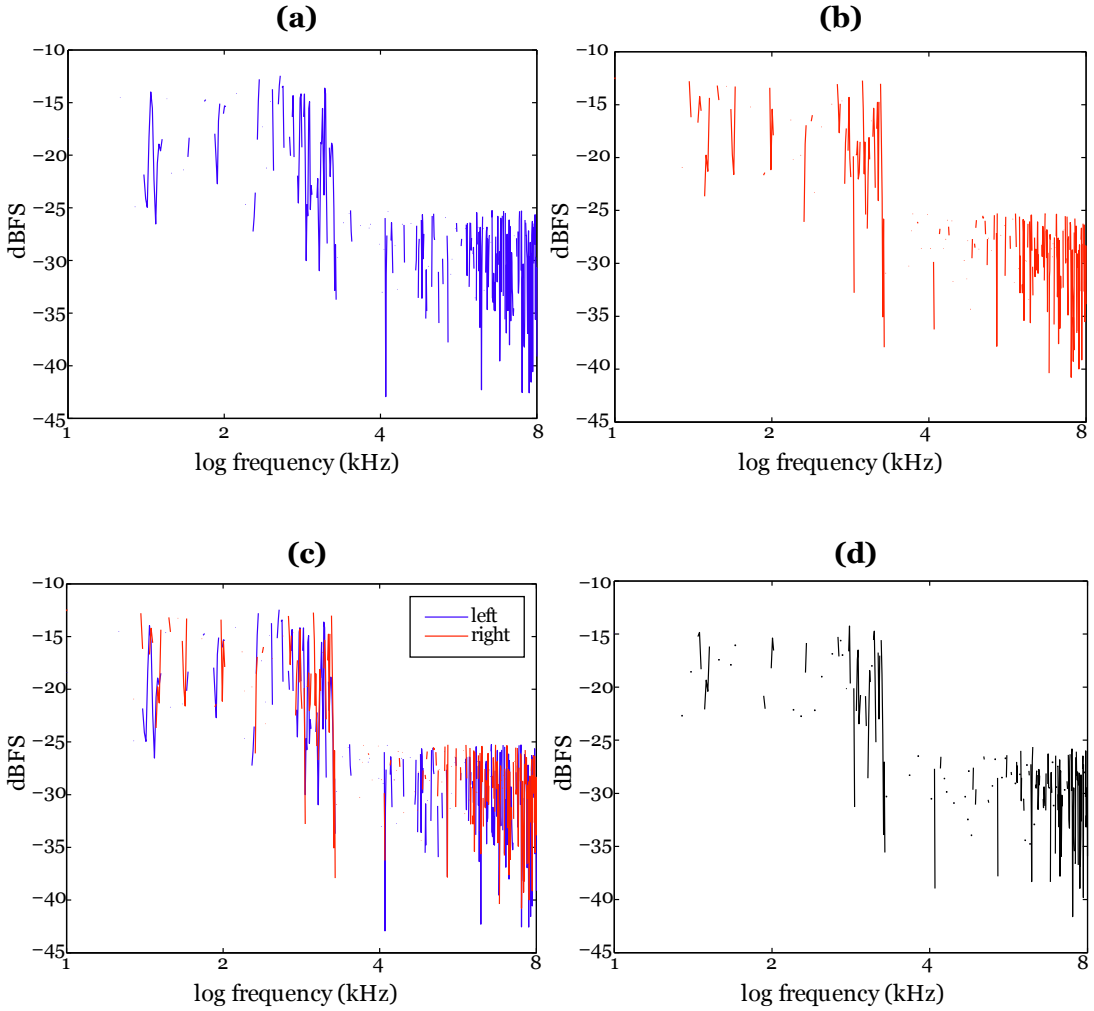

Figure 6.3: (a) Resulting FFT moduli after the amplitude discrimination for one frame of an excerpt of the song *Nude* [279] . Left channel (a), right channel (b), and superimposed (c). The result after the binary mask is applied is shown in (d). The bins that are set to the value of ε are omitted here for illustration purposes.

It is observed in the figure that the two signals (i.e. left and right) that are a result from the amplitude discrimination exhibit substantial differences. In fact, this is the case in most of frames of the tested database. This is because of

transients that occur in the music accompaniment in each channel and, as a result, they leak through the thresholds of AD. With this procedure, these transients are attenuated. In addition, the singing voice usually exists equally in both channels (as it is panned on the centre) and does not get reduced by the aforementioned process. Hence, the importance of using both channels for this process aids substantially towards the voice separation. It should be noted that the binary masking is applied here on each frame separately and is different from binary masks that are part of several CASA systems as presented in Figure 2.9.

The final stage of SEMANTICS involves the use of the phase information from the original mixtures in order to transfer $\hat{S}(k)$ with ISTFT to the time domain.

6.6 Experimental investigations

For the objective evaluation of the separation performance of SEMANTICS *bss_eval* is used with the same database as described in Chapter 5. This gives the flexibility of direct comparison between the SEMANTICS, SEMANICS, and ADRess. The results for signal to interference ratio (SIR) for a Hann window of 4096 samples, overlap of 87.5%, and M = 3 are shown in Figure 6.4. The full table of the numeric results is provided in Section A.3 of the appendix.

It is observed that, in terms of SIR, SEMANTICS provides consistently better interference rejection over its predecessor for all the songs in the database. The improvement obtained is in the range of 0.82 dB to 6.29 dB (μ = 3.33dB). As discussed, this metric together with SDR are the main indicators for the perceptual efficiency separation of the singing voice from a given song [266]. The metric of SAR (Figure 6.5) shows overall slightly better performance for SEMANTICS (μ = 1.28 dB), albeit with a few exceptions, namely 2 out of the 12 songs show slightly worse performance by 0.2 dB to 0.77 dB.

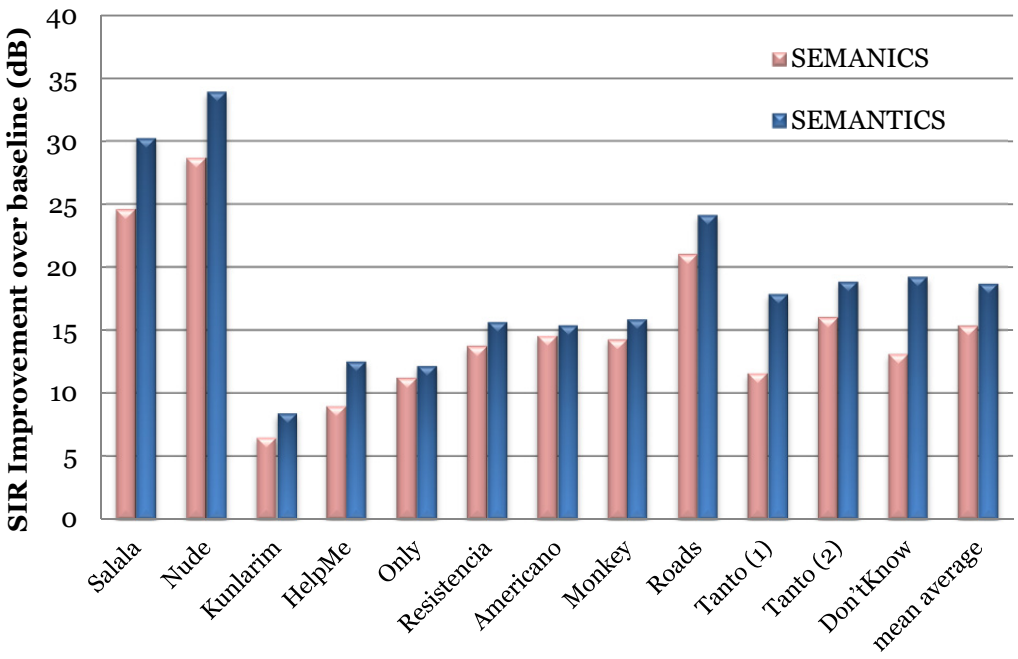

Figure 6.4: SIR performance of SEMANICS and SEMANTICS

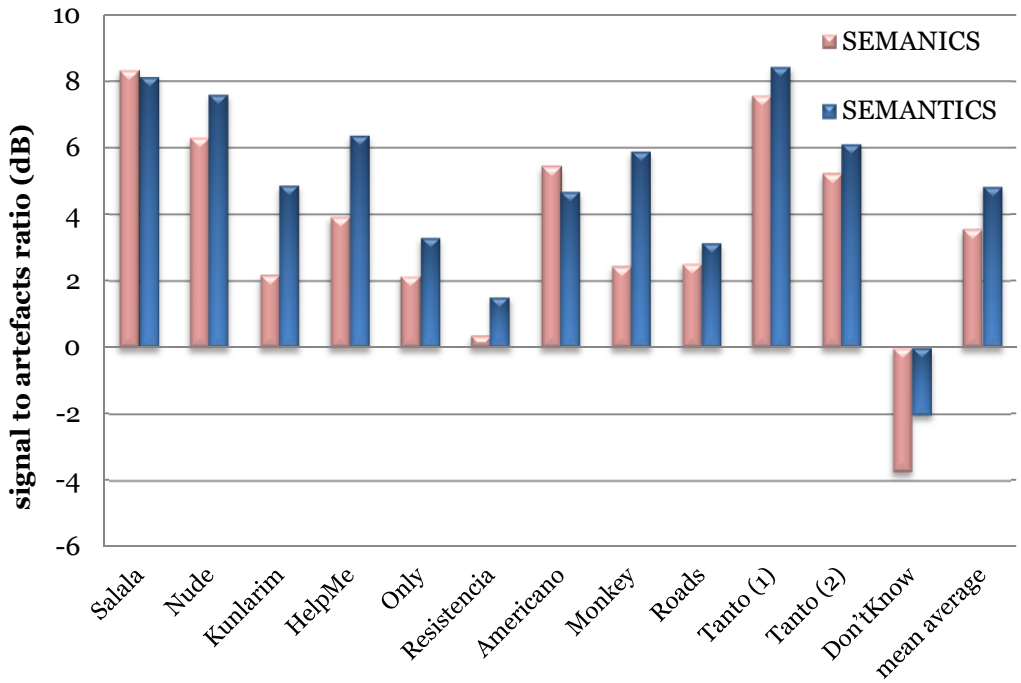

Figure 6.5: SAR performance of SEMANICS and SEMANTICS

The better performance in SAR is attributed to the use of both channels and the binary masking which reduces the leftover transients from music accompaniment and therefore decreases the noise floor.

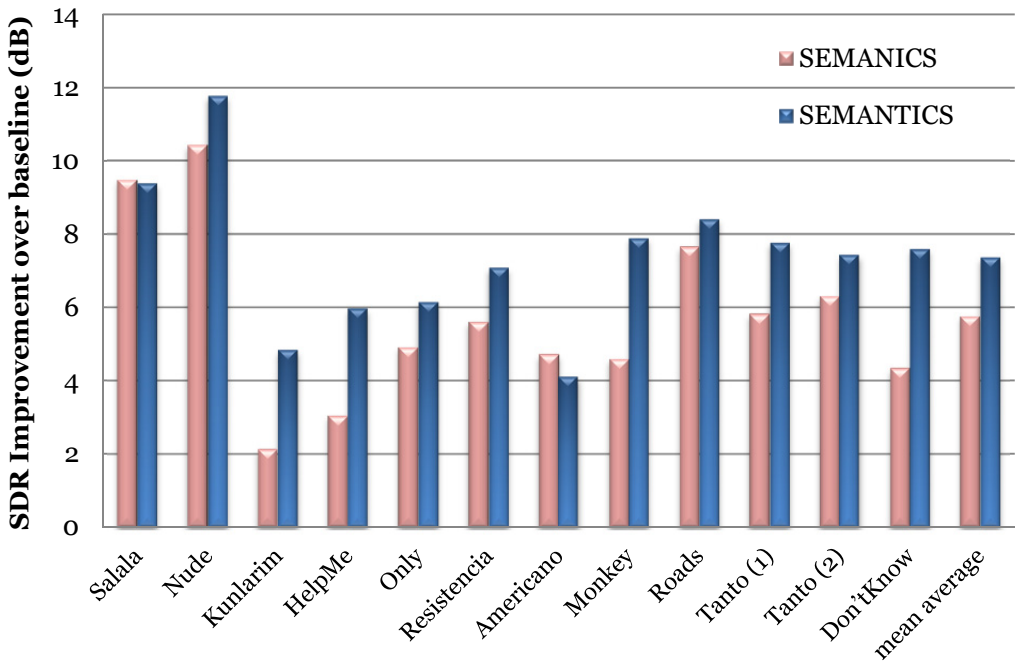

Figure 6.6: SDR performance of SEMANICS and SEMANTICS

Finally, as seen in Figure 6.6 based on this dataset, SEMANTICS provides generally better results also in terms of SDR (μ = 1.61 dB), with the exception of the two songs that are slightly lower also in SAR.

Similarly to the experimental setup that is discussed in 5.7.4, the system is tested with different parameters. The results of the variation that these parameters introduce to the performance of the system can be seen in Table 6.1. In the same manner as in Figure 6.1, the standard deviations for the mean average of the performance are shown when one set of parameters changes while the others remain fixed. The results from SEMANICS are offered for direct comparison.

	SEMANICS				SEMANTICS		
	x-points	stepsize	overlap	width (H)	M	stepsize	overlap
SDR	0.856	0.931	0.056	1.361	0.813	0.885	0.053
SIR	0.473	3.242	0.051	3.725	0.463	3.108	0.041
SAR	0.805	0.863	0.061	0.629	0.779	0.857	0.060

Table 6.1: Standard deviations of the results with different permutations of parameters.

In the table, the strongest advantage of SEMANTICS is demonstrated: except for improving interference rejection it also provides consistency of separation efficiency and reduction of the user set parameters to a minimum. This is attributed to the overall modifications proposed here, but mainly to the removal of ADRess from the system. As discussed earlier, SEMANICS includes ADRess and therefore suffers from dependence on its parameters, particularly the azimuth width(H). SEMANTICS on the other hand has only one parameter (the number of sub-bands M) which presents low variation when set between 3 and 5. It should also be noted that although the window size of the STFT affects the system (as expected), a length of 4096 or 8912 samples consistently provides the best results.

6.7 Chapter summary

In this chapter the system SEMANTICS has been introduced and analysed. This system can be described as an extension of the previous method presented in Chapter 5. The proposed method does not rely on ADRess and makes more efficient use of both channels of the stereo mix.

In particular, the modifications include the readjustment of NIC subtraction in order to be used as a pre-processing stage, which also involves the application

of high-pass filters on both channels of the original mixture. The use of both channels takes place in amplitude discrimination where a binary mask is also applied. The results indicate overall improvement over the previous method in all three aspects of *bss_eval*, especially in the area of SIR, which is important in the context of this study. A facet of the modified system is that minimises the user settings while improving consistency of the separation performance. This is mainly attributed to the removal of ADRess. While this as well as the previous chapter have dealt with the separation of voice when music and voice overlap in the frequency domain, the next chapter details methods that lead to successful removal of music-only time segments.

7
HYBRID SEMANTICS

The previously introduced systems, i.e. SEMANICS and SEMANTICS, address the frequency-domain separation through the combination of NIC subtraction and a frequency-domain separation method denoted as AD. In this chapter, a hybrid approach to SVS is proposed. The method, which is termed *H-SEMANTICS* (*Hybrid SEMANTICS*), is based on complementing previously introduced frequency-domain separation algorithms with time-domain separation.

7.1 Overview

For the purpose of time-domain separation (i.e. music-segment pruning) two different approaches are proposed here. The first approach is termed energy and NIC correlation (i.e. ENIC) and can be viewed as a post-processing separation stage in the time domain while the other, termed MFCC of IC Euclidean distance (i.e. MICED) forms an integrated part of H-SEMANTICS by operating on cepstral parameters.

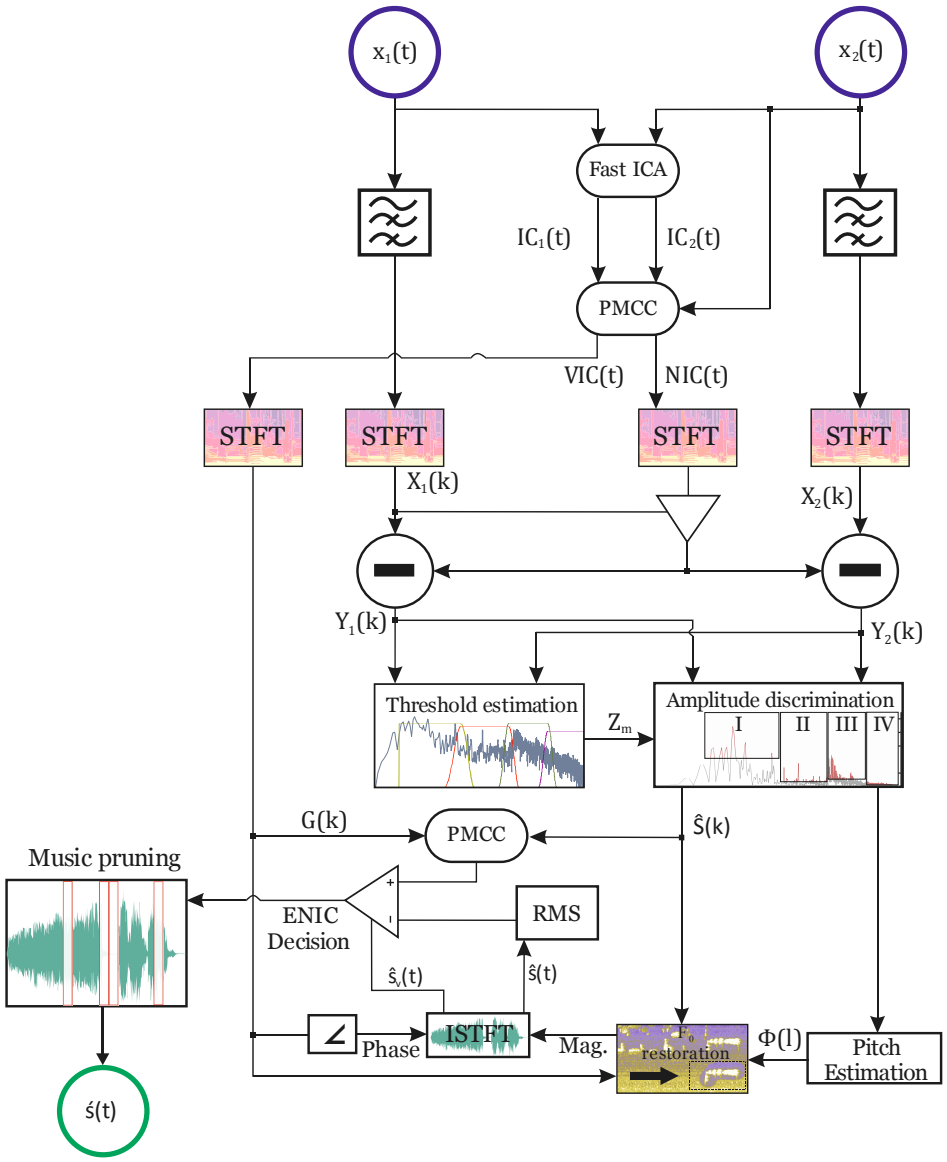

Figure 7.1: Overview of H-SEMANTICS with ENIC

In addition, H-SEMANTICS enhances the approach for separation in the frequency domain (where music and voice overlap) by introducing several modifications. These modifications involve the estimation of the thresholds in AD, the approximation of the phase information for the reconstruction of the signal, and the reconstruction of lost voice material due to high-pass filtering. An overview of the system with the first music pruning method, i.e. ENIC, is presented in Figure 7.1.

Briefly described, the system operation involves subjecting the original mixtures to high-pass filtering and subsequently to STFT. The NIC subtraction then follows as described in Section 6.3. Both estimated signals from the left and the right channels are afterwards used to estimate the thresholds of the amplitude discrimination (AD). Finally, the f_0 restoration is performed before the signal is pruned in the time domain.

The remainder of the chapter is organised as follows. The modifications to the previously introduced SEMANTICS (Chapter 6) are detailed in sections 7.2 and 7.3. Section 7.4 discusses the enhancement of the extracted singing voice through the restoration of f_0 (fundamental frequency). In sections 7.5 to 7.7, the proposed approaches to pruning the music-only time segments are described. Finally, Section 7.8 presents the objective evaluation method used in this study and analyses the experimental results for the proposed system.

7.2 Modifications in the frequency-domain separation

In contrast to SEMANICS and SEMANTICS, where the PMCC resulted in the determination of NIC and the discarding of the other IC, both components are used in H-SEMANTICS for different purposes. While NIC is used for an initial suppression of some of the music sources, the vocal independent component (i.e. the other ICA output, denoted as VIC) is used for fine-tuning of the estimated vocal and reconstruction of its phase.

As reported in Section 6.4, applying a high-pass filter to $x_1(t)$ and $x_2(t)$ before computing their Fourier transforms can result in more accurate scaling during the NIC subtraction. Depending on the cut-off frequency (f_c) of the high-pass filter, some frequency components of the voice will also be lost but, as described later in this chapter, they can be recovered afterwards. It should be noted that f_c must be set to such a value that the singing voice loses at most its f_0, but not its f_1 (first harmonic). This is important in order to successfully

estimate the pitch of the singing voice. For that purpose, the high-pass filter that is used in this method is a high-order, linear-phase FIR filter with f_c = 200 Hz. The reason for the choice of the higher order of the filter over the low-order Butterworth filter used in Section 6.4 is that it was observed during further experimentation with SEMANTICS that a non-linear phase filter can affect the accuracy of the STFT process that follows.

7.3 Threshold estimation

As discussed in previous chapters, mixing engineers use a number of techniques and processes to manipulate the contributing components and ensure the vocal part is clearly audible. This usually involves dynamic range compression to impose artificial stability in the amplitude of the vocal signal, and filtering to enhance the frequency spectrum of the singing voice in bands where masking occurs [246]. This standard mixing approach has been the main motivation behind the introduction of the amplitude discrimination (AD) process as described in Section 5.4. However, the study in Section 6.5 has shown that for maximising the effectiveness of such an approach, the AD should be localised by dividing the full frequency spectrum into a number of sub-bands. The other important consideration for the purpose of AD is the value of the threshold to be deployed in the case of each sub-band.

The motivation behind the modification of the aforementioned process is that the estimation of the thresholds is critical to the performance of the system. For this purpose, the thresholds are calculated dynamically using mel-frequency overlapping Tukey windows, that are used as filters. Formally, for each l frame of audio, for a given sub-band m, the threshold Z_m is defined by:

$$Z_m = \frac{1}{c_4^m - c_1^m} \sum_{i=1}^{2} \sum_{k \in [c_1^m, c_4^m]} Y_i(k,l) w(k,m), \qquad m \in \{1, 2, \ldots \mathrm{M}\}, \tag{7.1}$$

where M is the number of equally spaced sub-bands, $Y_i(k, l)$ are the spectrograms of the left and the right channels resulting from the NIC subtraction, k is the index of the FFT points, i is the channel index, and $w(k, m)$ is a unit weight window described in (7.2).

$$w(k,m) = \begin{cases} \dfrac{c_4^m - c_1^m}{c_4^m - c_1^m + c_3^m - c_2^m} \left(1 + \cos\left(\pi \dfrac{k - c_1^m}{c_2^m - c_1^m} - 1\right)\right) & when\ c_1^m \le k < c_2^m \\ \dfrac{2(c_4^m - c_1^m)}{c_4^m - c_1^m + c_3^m - c_2^m} & when\ c_2^m \le k \le c_3^m \\ \dfrac{c_4^m - c_1^m}{c_4^m - c_1^m + c_3^m - c_2^m} \left(1 + \cos\left(\pi \dfrac{k - 2c_3^m + c_4^m}{c_4^m - c_3^m} - 1\right)\right) & when\ c_3^m < k \le c_4^m \end{cases} \tag{7.2}$$

The variables c_1^m to c_4^m represent the corners of the frequency bands which are defined by:

$$c_p^m = \frac{700\mathrm{K}\left(\left(1 + \frac{\mathrm{S_R}}{1400}\right)^{\frac{m - H(p)}{\mathrm{M}}} - 1\right)}{\mathrm{S_R}}, m \in \{1, 2, \ldots \mathrm{M}\}, p \in \{1, 2, 3, 4\}, \tag{7.3}$$

where $\mathrm{S_R}$ is the sampling rate, K is the FFT size, and $H(p)$ = [1+α, 1, 0, -α]. The parameter α defines the percentage of overlap of sub-bands in the mel-frequency domain. The derivation of each of (7.2) and (7.3) is described in Section A.4 of the appendix. The reason that the corners are defined in such a

way is that they overlap in the mel-frequency spectrum. As can be seen, the formula used for the linear- to mel-frequency conversion is the one suggested by O' Shaughnessy [286]. It should also be noted that c_1^1 is adjusted appropriately so that it excludes the band that has been filtered out during pre-processing.

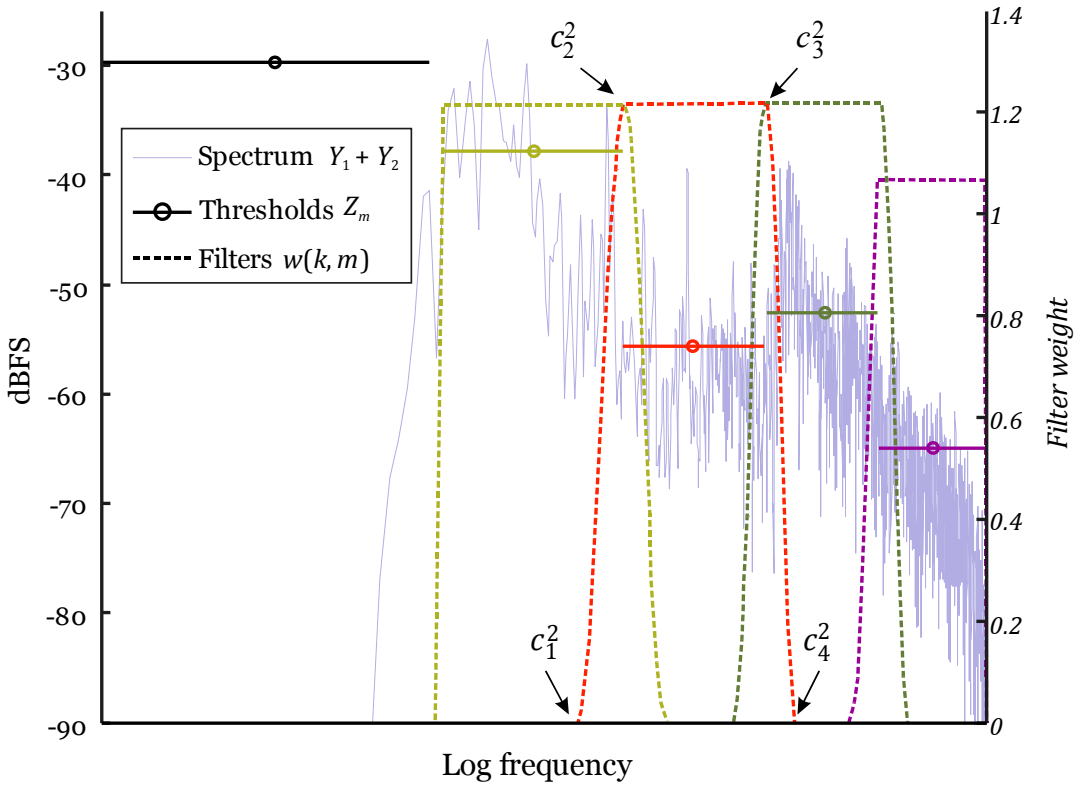

Figure 7.2: Amplitude discrimination of H-SEMANTICS on a windowed frame of audio

The AD process applied subsequently to each band, functions as a binary mask, allowing only the bins with magnitudes higher than their respective sub-band thresholds to pass. Similarly to 6.5, the resulting spectrum $\hat{S}(k, l)$ contains the averaged bins from the two channels as in (7.4):

$$\hat{S}(k,l) = \begin{cases} \frac{1}{2}\sum_{i=1}^{2} Y_i(k,l) & \text{if} \begin{cases} Y_1(k,l) > Z_m \\ Y_2(k,l) > Z_m \\ k \in [c_2^m, c_3^m] \end{cases} m \in \{1,2,\dots \text{M}\} \\ \varepsilon, & \text{otherwise,} \end{cases} \quad (7.4)$$

where ε is the machine epsilon. According to Section 6.5, satisfactory performance is achieved when the number of sub-bands is fixed in the range 3 to 5. An example of the way that H-SEMANTICS calculates the thresholds for AD on a frame of audio is presented in Figure 7.2, for M = 4 and α = 0.25. The spectrum is shown in dBFS only for presentation purposes, although Y_i are magnitude spectra in the proposed approach.

7.4 F_0 restoration

As mentioned in Section 7.2 of this chapter, the observed mixture is subjected to high-pass filtering (with the cut-off frequency f_c = 200 Hz), which helps with the scaling during the NIC subtraction. However, this value of f_c also removes the f_0 of the singing part for some frames, when the singing pitch is lower than a G_3 (≈ 196 Hz). As expected, this occurs mainly in songs by male singers.

In this section, a method is described that attempts to restore the fundamental frequency in frames where pitch estimation shows that f_0 has been filtered out. At this stage, the pitch estimation method from the *MIRToolbox* [287, 288] is used, with the parameters set such that only the most significant pitch of the frame is extracted. The effectiveness of the pitch estimation is significantly facilitated by the processes that the signal has been previously subjected to, namely the NIC subtraction and AD. The approach described below replaces the bin that contains the estimated f_0 with the one from the short-term spectrum of VIC, i.e. $FVIC(k, l)$.

At first, $\Phi(l)$ is defined as the pitch estimated in every frame l, and this is assigned to the corresponding bin in the spectrum of the frame. This restoration process can be expressed as follows:

$$\hat{S}(k,l) = FVIC(k,l) \quad if\ k = (i+1)\Phi(l) < f_c,\ \ i \in \{0,1,2,\dots\Theta\}, \tag{7.5}$$

where Θ is the maximum number of harmonics that are restored. The parameter Θ, which is fixed throughout the process, can be set in such a way that it accounts for harmonics higher than the f_0 that might be lost during high-pass filtering. In the present study though, it suffices to restore just the f_0 (i.e. $\Theta = 0$), as this accounts for the restoration of the fundamental frequency down to an estimated G_2 ($\approx$ 98 Hz). Songs with vocal frequencies falling below this musical note are considered very rare. As a reference, only a bass singer can, at his extreme lower end, extend down to an F_2 ($\approx$ 87 Hz). The other five predominant vocal ranges (i.e. soprano, mezzo-soprano, alto, tenor, baritone) do not cover notes that are below an A_2 (= 110 Hz) [289].

7.5 Pruning music-only time-segments

As a general observation, songs contain some time-segments, where the vocal rests. This can occur either between the melodic phrases of the singer ("breaths") or when the singer pauses and the music develops instrumentally. Beyond these, there are always segments in the time domain where the singing vocal is not present and, therefore, no frequency overlapping is occurring. These music-only time segments smear the output of any frequency-based SVS system and can produce inaccuracies in subsequent applications, such as melody transcription. Thus far, only the frequency overlapping case has been discussed in the proposed methods of this thesis. However, comprehensive SVS must reject the time segments where the vocal is simply not present. In this

section, two different approaches to this problem are described. At this point, it should be stressed that the primary objective here is to enhance the measured performance of the proposed system. Therefore, having a low detection rate is not highly critical and instead, these approaches focus primary on achieving the lowest possible false alarm rate.

7.6 Energy-based pruning

A consequence of the processes detailed above is that of considerable attenuation of the energy of the music components. As a result, the energy of music-only time segments will be much lower than the vocal segments. The latter can be exploited in order to identify and eventually remove such purely music segments.

This method takes place after all the signal frames have been processed as described previously and the signal is transformed back to the time domain with the inverse short-term Fourier transform (ISTFT) using the phase of VIC for the overlap-add reconstruction process [261].

The estimated signal, i.e. $\hat{s}(t)$, is segmented into v rectangular segments of 250 ms and the root mean square (RMS) amplitude of each segment is computed. The length (i.e. 250 ms) of segments is chosen in such a way to avoid inaccuracies due to possible transient peaks in the energy (e.g. attack transients). Moreover, the use of a shorter segment length may not be beneficial. This view is supported by the observation that in the case of shorter segments (e.g. <200 ms) where the vocal rests, there usually exists a reverb "tail" from the vocal. The pruning of such segments is not necessarily welcome.

The proposed approach estimates a decision threshold that classifies each segment either as vocal-and-music or music-only (i.e. above or below the threshold respectively). The first step in this process is to create a vector $R_{\hat{s}}(v)$ containing the RMS value of each segment. An initial threshold is then set by

subtracting the standard deviation from the mean of the given vector. This step biases the decision towards having fewer false alarms, i.e. segments that contain vocal but are classified as music-only, at the expense of more miss-detections, i.e. segments that are music-only but are not pruned. This threshold is then adjusted dynamically in the following manner. The frequency spectra from the segments of VIC and from the estimated signal are cross-correlated, and the result acts as a segment-specific weight resulting in a dynamic decision boundary as can be seen in Figure 7.3.

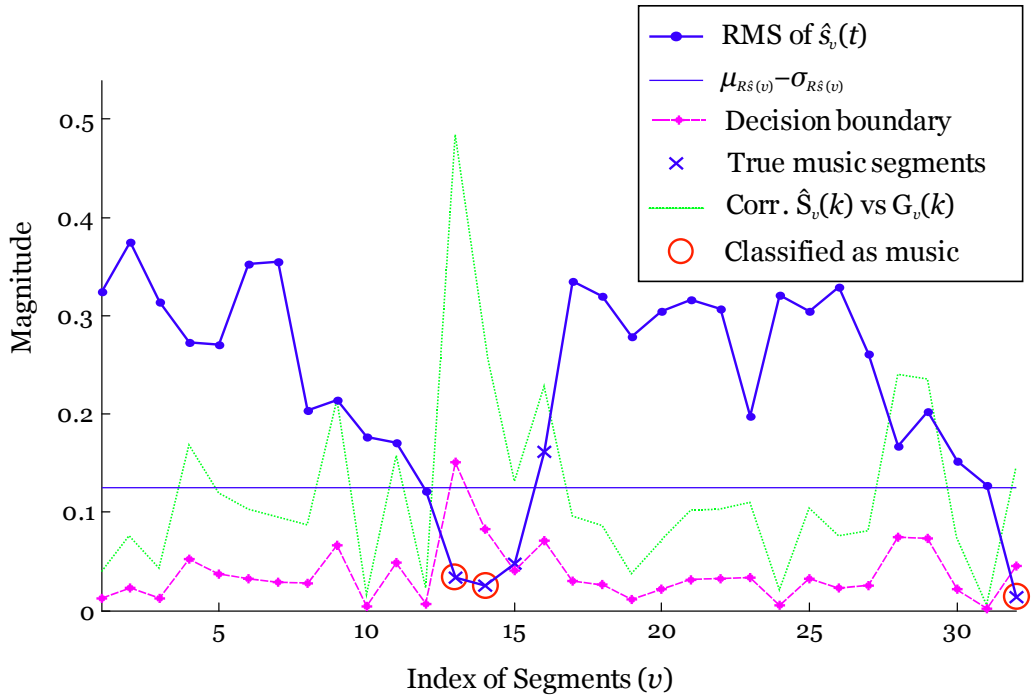

Figure 7.3: ENIC pruning for the song excerpt *Salala* [160], for γ = 0.4. The PMCC index "Correlation $\hat{S}_v(k)$ *vs* $G_v(k)$" modulates the initial threshold ($\mu R_{\hat{s}(v)} - \sigma R_{\hat{s}(v)}$) to form a dynamic decision boundary.

A parameter γ is defined such that if the correlation exceeds it, the initial threshold is amplified linearly, otherwise it is attenuated similarly. The parameter γ is set in such a way that ENIC produces the least number of false alarms at the expense of misdetections. The value that is found to meet this

requirement for the case of H-SEMANTICS is γ = 0.4.

This approach is found to reject most of the non-vocal segments, while maintaining a low false alarm rate. In order to maintain the low false alarm rate, a complementary approach is to consider a segment as purely music, only when it has a similarly labelled neighbour.

Formally, this approach begins with the calculation of RMS for each segment:

$$R_{\hat{s}(v)} = \sqrt{\frac{\sum_{t=1}^{\mathrm{T}} s_v(t)^2}{\mathrm{T}}} \quad \forall v \tag{7.6}$$

The decision for binary classification is provided through (7.7), the derivation of which is presented in Section A.4 of the appendix.

$$p(\hat{s}_v) = \frac{1}{2} + \left[\frac{1}{2\gamma}\left(\frac{R_{\hat{s}}(v)}{\gamma}\left(\left(\mu_{R_{\hat{s}}} - \sigma_{R_{\hat{s}}}\right)\left|\frac{1}{\gamma \mathrm{K}}\sum_{k=1}^{\mathrm{K}}\frac{(\hat{S}_v(k) - \mu_{\hat{S}_v})(G_v(k) - \mu_{G_v})}{\sigma_{\hat{S}_v}\sigma_{G_v}}\right|\right)\right)\right] \quad \forall v, \tag{7.7}$$

where $\hat{S}_v(k)$ and $G_v(k)$ are the moduli of the Fourier transforms of Hann-windowed time segments of $\hat{s}(t)$ and *NIC* respectively, and γ sets the level of correlation between $\hat{S}_v(k)$ and $G_v(k)$ above which the decision boundary is increased.

The threshold $p(\hat{s}_v) \in [0, 1]$ is a practical way of classifying a segment as music-only (i.e. $p(\hat{s}_v) < 0.5$) or vocal with music (i.e. $p(\hat{s}_v) > 0.5$). Following the classification process, the groups of adjacent segments that are identified as music-only are convolved with an inverted Tukey window (a = 0.75) [290]. This window is used after the IFFT process and provides a smooth transition between the vocal and the music-only sections.

7.7 MFCC-based pruning

Although the method described in 7.6 is effective inside the proposed system, there are some practical limitations. First, all of the signal must pass through the system in order to calculate $\mu_{R\hat{s}(v)}$. This is not advantageous in terms of memory requirements and reduces operational speed. Furthermore, depending on the sampling rate as well as the window size, the signal will have to be transformed back and forth from the frequency to the time domain in order to meet the specified length condition (i.e. 250 ms). In other words, if the number of FFT points (i.e. K) ≠ S_R / 4, $\hat{s}(t)$ needs to be subjected to STFT, again reducing efficiency.

This section therefore describes an alternative approach to pruning music-only segments. This is based on one of the most popular parametric representations of audio, the mel-frequency cepstral coefficients (MFCC). The fundamental assumption of this method is that the MFCCs of the frames of the estimated signal, i.e. $\hat{S}(k)$, and NIC, i.e. $G(k)$, are more similar when the $\hat{S}(k)$ contains only music. In the opposite case, the MFCCs of $\hat{S}(k)$ show greater similarity to the MFCCs of the VIC frame. Therefore, the decision in the proposed method is based on the ratio of similarity of the features of the estimated signal against the VIC and NIC respectively.

Briefly, the proposed MFCC-based music pruning works as follows. The MFCCs are extracted for $\hat{S}(k)$, $G(k)$ and $FVIC(k)$. Next, the Euclidean distances for the pairs of [$\hat{S}(k)$, $FVIC(k)$] and [$\hat{S}(k)$, $G(k)$] are measured. The decision boundary is obtained by processing the ratio of the distances of the first and second pair with a simple moving average filter (lag = 19) in order to simultaneously deal with any transients as discussed in 7.6. The edges of the filter output are calculated as follows. The first point is not averaged, and the points from 2 to 10 are averaged over 3, 5, 7, … 19 points of the input signal. The reverse process is followed for the ending edge of the filter output [291], while the lag (i.e. τ) of 19 was chosen because it is consistent with the 250 ms segments that

were used in ENIC (i.e. nineteen 4096-length frames at 87.5% overlap for approximately 278 ms for S_R = 44.1 kHz and 256 ms when S_R = 48 kHz).

According to the literature [292], the performance of the MFCC depends upon the number of coefficients used, the number of filters in the filter-bank, the type of filters, the spacing of the filters, and the equation that is used for warping the Fourier transform bins. The filter-bank used comprises 40 unit-weight overlapping triangular filters equally spaced in the whole mel-frequency spectrum. An important aspect of the approach is that it uses 13 coefficients, including the usually discarded zeroth coefficient. The zeroth coefficient can be perceived as the average energy of all the filtered bands. As seen in 7.6 the energy of the signal is vital for each classification, and so the inclusion of the zeroth bin is deemed necessary. However, the major deviation in MICED with respect to many conventional methods [293-295] is that the whole of the processing is performed "on the fly", and there is no modelling, codebook, or training, e.g. [296]. In fact, this is one of the main reasons that a direct comparison of the proposed segregation method with existing ones is not applicable.

Formally, the Euclidean distance ratio for each frame is calculated as in (7.8):

$$d(l) = \sqrt{\frac{\sum_{n=0}^{12}\left(M_{\hat{S}(l)}(n) - M_{FVIC(l)}(n)\right)^2}{\sum_{n=0}^{12}\left(M_{\hat{S}(l)}(n) - M_{G(l)}(n)\right)^2}}, \tag{7.8}$$

where M are the MFCCs for the l^{th} frame.

The result is used as the input to the moving average filter, which is calculated subsequently as follows.

$$MA(l) = \begin{cases} \sum_{j=1}^{2l-1} \frac{d(j)}{2l-1}, & \text{if } \; l \leq \frac{\tau-1}{2} \\ \sum_{j=\frac{1-\tau}{2}}^{\frac{\tau-1}{2}} \frac{d(l+j)}{\tau}, & \text{if } \; \frac{\tau-1}{2} < l \leq \mathrm{L} - \frac{\tau-1}{2} \\ \sum_{j=l-\mathrm{L}}^{\mathrm{L}-l} \frac{d(l+j)}{2(\mathrm{L}-l)+1}, & \text{if } \; l > \mathrm{L} - \frac{\tau-1}{2} \end{cases} \quad \forall\, l, \qquad (7.9)$$

where L is the total number of l, and τ is the odd number of lag in samples. Segments that are above 0.5 are classified as music-only. Figure 7.4 shows the results of this approach on the same song that was used in Figure 7.3.

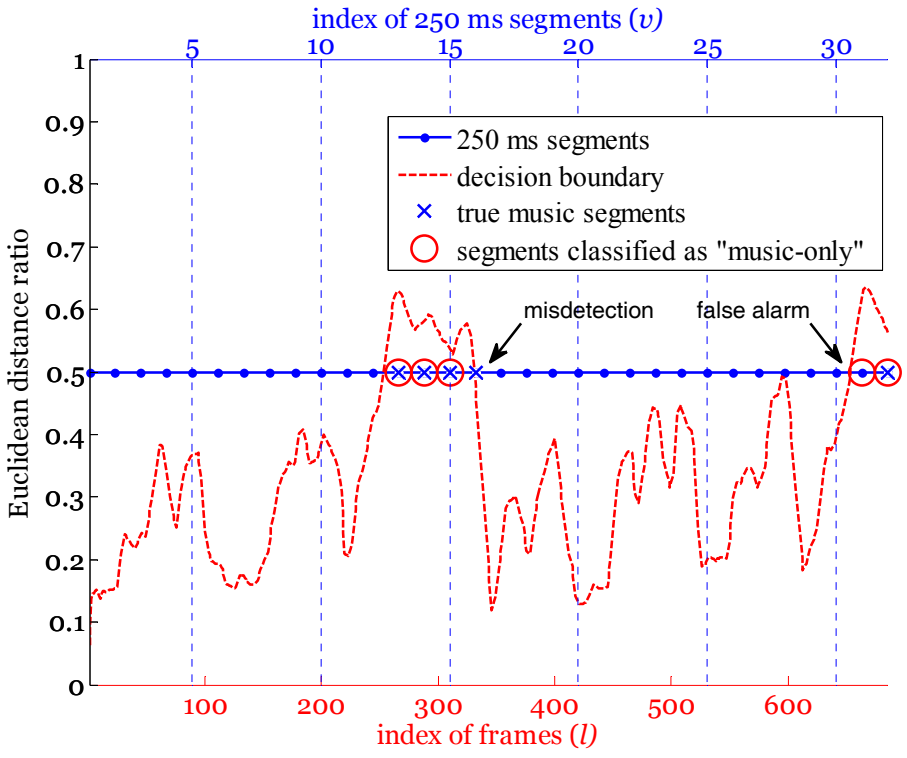

Figure 7.4: MICED pruning for the song excerpt *Salala* [160]. Although the method operates on 250 ms segments, the "resolution" of the decision boundary is much higher, as it is calculated based on a moving average of overlapping 93 ms frames (l).

Not surprisingly, it can be seen that this approach produces similar results with the one proposed in 7.6, albeit with a few false alarms. Particularly, ENIC (Figure 7.3) detects correctly 13, 14, and 32 without false alarms. For the same excerpt, MICED detects correctly segments 13, 14, 15, and 32, while 31 is a false alarm. This is attributed mainly to the fact that, in contrast to ENIC, MICED is calculated for every frame, without the need for information on the subsequent parts of the signal. Therefore, it allows for the whole system to be implemented in real time.

7.8 Experimental investigations

The experimental investigation is presented here in two stages. The first stage concerns the time-domain segregation, while the second stage discusses the overall evaluation of the system.

	Original		with ENIC		with MICED	
	M + V	music	Detected	F. alarms	Detected	F. alarms
Salala	27	5	3	0	4	1
Nude	53	2	2	1	2	1
Kunlarim	31	0	0	2	0	2
Help Me	29	6	2	0	3	0
Only	31	2	2	0	0	0
Resistencia	24	0	0	1	0	0
Americano	25	5	3	0	3	2
Monkey	13	5	5	1	3	1
Roads	51	4	1	1	1	1
Tanto (1)	51	0	2	0	0	2
Tanto (2)	88	3	1	3	2	4
Don't Know	27	1	0	1	0	2

Table 7.1: Performance of ENIC and MICED on the time-domain separation. "M + V" and "music" shows the hand-labelled vocal-and-music and music-only segments (250 ms).

The results that concern specifically the time-domain separation are presented in Table 7.1. In this table, the 250 ms segments from the original mixtures are manually labelled as either music and voice (i.e. M + V) or music-only (shown as "music" in the table). The segments that are successfully detected by the proposed methods as music-only are presented for ENIC and MICED respectively. The column labelled as 'F. alarms' shows the false alarms for each system, i.e. segments that are actually "M + V" but are incorrectly labelled as music-only.

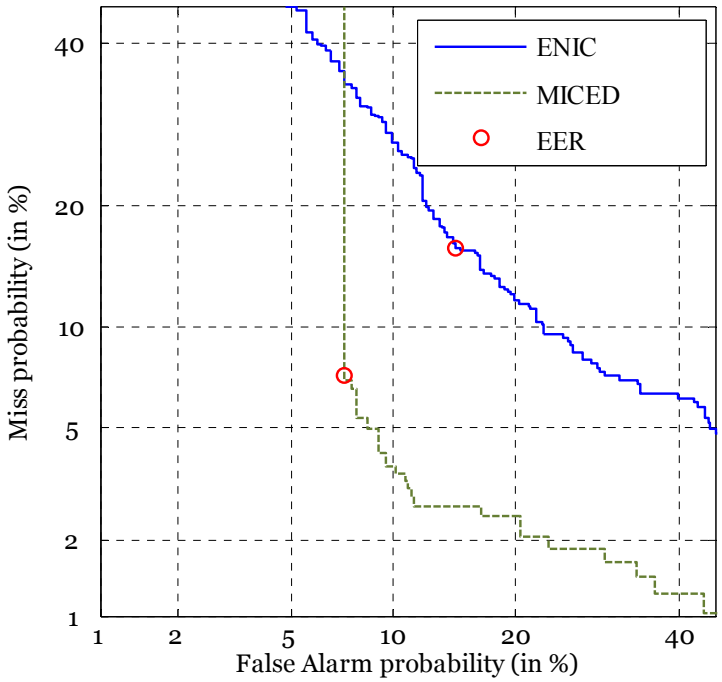

Figure 7.5: DET plots and equal error rate (EER) for ENIC and MICED

Also, as the time-domain segregation is essentially a detection task, a detection error trade-off (DET) plot [297] is offered in Figure 7.5 (using *DETware 2.1* [298]). In the figure, MICED exhibits a lower equal error rate (EER) than ENIC. However, as discussed in this chapter, the most important goal of the proposed pruning method is a low false alarm rate. This is important, especially because

every false alarm resulting from the pruning process has a negative impact on the overall separation effectiveness.

In order to have a direct comparison of H-SEMANTICS with the previous systems that are proposed in this thesis (chapters 5 and 6), the same experimental setup and metrics are used for assessing the separation effectiveness. Furthermore, the performance of the system is evaluated using the same database. For this purpose, the *bss_eval* results (in dB) for song excepts using a Hann window of 4096 samples, overlap of 87.5%, M = 3, and γ = 0.4 (for ENIC) are shown in Figure 7.6 for SIR. As with previous experiments, the full table of the numeric results is provided in Section A.3 of the appendix.

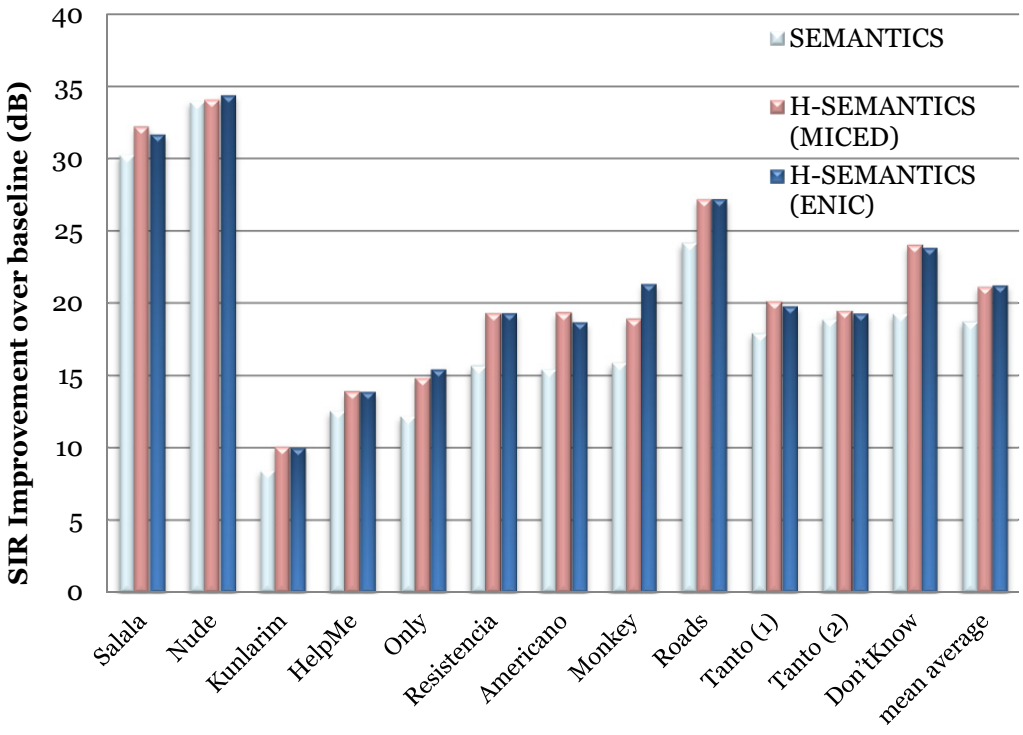

Figure 7.6: SIR performance of SEMANTICS and H-SEMANTICS

The improvement of H-SEMANTICS with ENIC over the previous system, i.e. SEMANTICS, spans from 0.41 dB to 5.44 dB for SIR (with μ = 2.52 dB).

For SAR (Figure 7.7), the improvement range is 0.04 dB to 1.82 dB (with μ = 0.89 dB). Although the SAR does not show significant improvement, it clearly demonstrates the advantage of H-SEMANTICS in terms of increasing separation without introducing additional artefacts. This is mainly attributed to the restoration of f_0: an addition to the proposed system that also enhances the audible result.

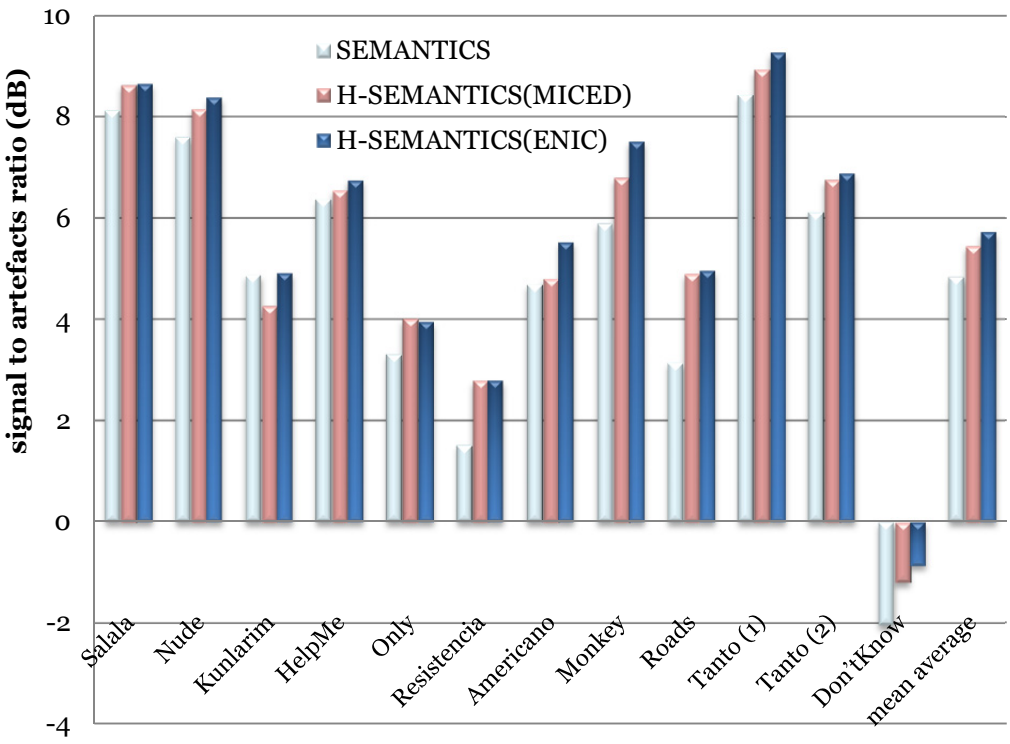

Figure 7.7: SAR performance of SEMANTICS and H-SEMANTICS

As discussed in earlier chapters, the SDR is a collective separation metric and is directly correlated with the perception of acoustic separation [266]. As seen in Figure 7.8, this measure also shows improvement, ranging from 0.54 dB to 2.18 dB. Furthermore, the overall effectiveness of separation when compared to the baseline is also highly encouraging (μ = 8.59 dB). Overall, H-SEMANTICS outperforms its predecessor in all three aspects and shows a significant SIR

improvement over the baseline. It is also noteworthy that the greatest improvements are associated with songs for which the SIR obtained with SEMANTICS are relatively low (<20 dB). In other words, the newly introduced system shows a more consistent capability in achieving separation, even in the cases where the previous system performed relatively poorly.

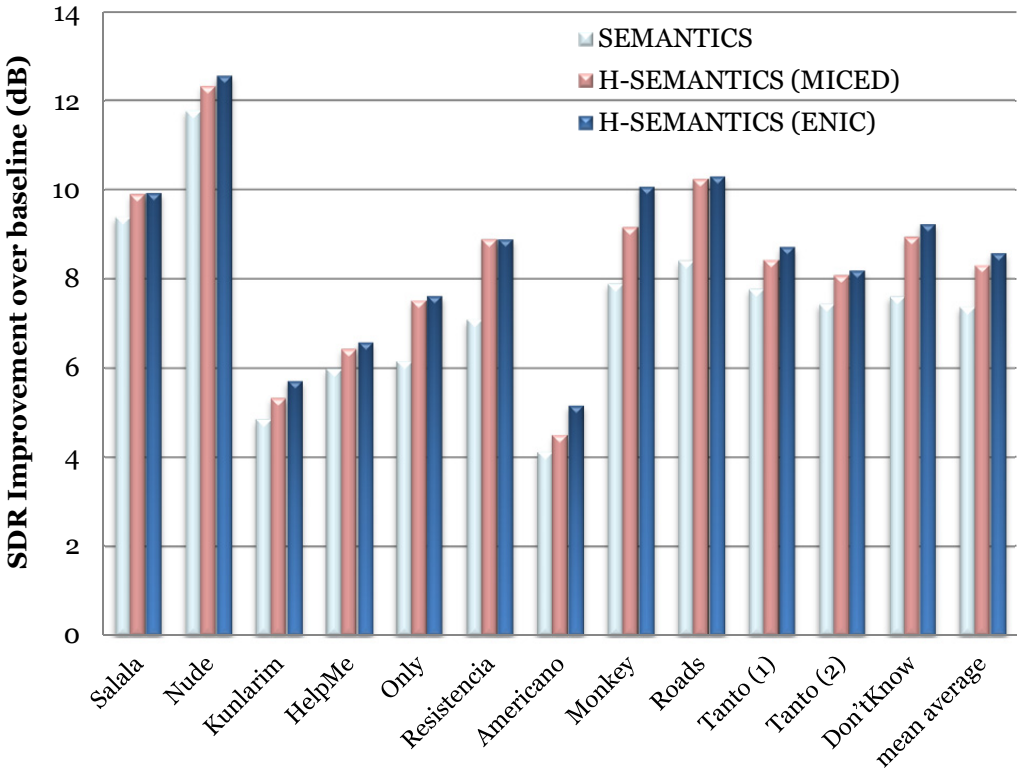

Figure 7.8: SDR performance of SEMANTICS and H-SEMANTICS

It is also shown in the figure that the ENIC pruning method is better than MICED (although sometimes marginally). As presented earlier, although MICED has lower EER, it cannot ensure a <7.5% false alarm probability, independently of the miss probability. This is why ENIC outperforms MICED in the *bss_eval* performance metrics.

7.9 Chapter summary

A novel approach to unsupervised SVS, termed *H-SEMANTICS*, has been proposed and investigated. The approach entails modifications and extensions to the previously introduced system, including the introduction of a different method of threshold estimation for the amplitude discrimination process based on overlapping tapered cosine filters for equally spaced mel sub-bands. In addition, restoration of the fundamental frequency (when required) is incorporated, and a different method for the estimation of phase during ISTFT is deployed. More importantly, two time-domain music-only pruning methods are proposed. The first is based on the RMS of the estimated signal and its correlation with the NIC in the time domain, while the second makes use of the distance of the MFCC representation of each of the two independent components to the estimated signal. It is shown that, while the first one (termed ENIC) provides better performance, the MFCC-based system (termed MICED) is more efficient computationally. The experimental results are based on *bss_eval*, and clearly demonstrate that H-SEMANTICS is improving the performance of its predecessor by an average of 1.2 dB SDR and 2.5 dB SIR for the case of ENIC and by 1 dB SDR and 2.4 dB SIR for the case of MICED. It should also be noted that the previous system has already been shown to offer superior performance over the ADRess-based source separation method SEMANICS as described in Chapter 5.

8
SUMMARY, CONCLUSIONS, AND FUTURE WORK

The aim of the research presented in this thesis has been to develop an effective method in order to separate the singing voice from stereophonically produced commercial recordings. For this purpose, an extensive literature review has been carried out and the shortcomings of previous work in the field have been identified. The novel aspect of the work has been the creation of three systems whose effectiveness has been confirmed through experimental evaluations. The overall findings of this endeavour are summarised in this chapter. The chapter also includes suggestions with regards to the continuation of the project.

8.1 Summary and conclusions

The singing voice is the most prominent content of music tracks that can be described as songs. The separation from its music accompaniment is considered highly desirable. One of the main reasons is that important applications of MIR such as song identification, singer identification, melody extraction, lyrics recognition, and lyrics alignment require the vocal element alone and hence the effective separation of this from the accompanying music. This is supported by studies that exhibit severe limitations when information

retrieval is attempted without prior source separation as discussed in Section 1.2.

MIR can be broadly separated in two categories, namely the symbolic, and the audio information retrieval (AIR). The former presents disadvantages as it is based on cues that are not always available, especially in the contemporary era. In contrast, AIR draws information from the audio signal *per se* and thus is more promising towards a generic solution for MIR. Furthermore, AIR is significantly helped by successful singing voice separation (SVS).

The complexity of the challenge of SVS has been identified as two-fold: separation in the frequency domain (separation of overlapping music and voice frequencies), and segregation in the time domain (discrimination between vocal and music time segments). The latter has received considerable attention from the research community, while the former is still a niche field. The challenge in this area is found at the inherent diversity of music, as well as the strong harmonic correlation that overlapping music and singing voice exhibit.

The literature review in this thesis has covered the mechanics of the singing voice and highlighted its differences from speech. These include certain formants that are more prominent in singing, as well as the duration ratio of vowels to consonants which is significantly lower in speech.

The quest towards a generic solution for SVS has drawn the interest of the industry and the research community alike. In fact, several commercial approaches exist that are related to SVS. However, they are usually dependent on supervised processes, and therefore are not appropriate for large databases. It is worth noting that such methods are motivated by disciplines across a plethora of fields, such as neuroscience, cognitive psychology, and psychoacoustics.

In this thesis, three different categories of approaches have been analysed in this thesis. The first one, described in Chapter 2, is named computational

auditory scene analysis (CASA) and is largely based on the modelling of the human auditory system. In fact, a substantial part of the theoretical background of CASA derives from the seminal work of Bregman on ASA. One significant idiom of CASA is that because of its vague definition, it lacks methodology boundaries. Therefore, the thesis has taken the approach of a unified description with emphasis on features that are quite common amongst different approaches. These are the cochleagram, the correlogram, and the time-frequency binary and soft masks.

The second category that has been analysed in Chapter 3 is the blind source separation or BSS. The family of BSS algorithms is largely based on statistical properties of the target signal and its model is associated with the cocktail party problem. In contrast to CASA, which targets mainly single channel mixtures, BSS requires multiple observations of the mixtures. In this context, the thesis described the anatomy of the principal component analysis (PCA) that is based on the second order bivariate cumulant, namely the covariance. The chapter continued with the younger 'sibling' of PCA, the independent component analysis (ICA). ICA encompasses PCA in its pre-processing stage, and draws its efficacy from the fourth moment, the kurtosis. Since the empirical estimation of kurtosis poses a challenge not only in ICA but generally in statistics, the chapter also described methods of how this problem is tackled in the algorithm Fast ICA (FICA) which has been of considerable importance in this study.

Finally, the third category (Chapter 4), detailed in this study, makes assumptions that are present mainly in the context of stereophonic commercial recordings. A key facet of this approach is the inter-channel phase consistency of the sources. In other words, this category makes use of the inter-channel (or inter-aural) intensity difference (IID). The algorithm that has been analysed from this category is one that shares significant theoretical background with the rest of IID methods. This algorithm is termed azimuth discrimination and re-synthesis (ADRess).

Having identified the downsides of all these attempts and analysed the methods of contemporary music production, a new method, termed SEMANICS has been proposed in Chapter 5. The method is based on the modification of ADRess and the introduction of two novel techniques, termed amplitude discrimination (AD), and non-vocal independent component (NIC) subtraction.

The motivation for the use of AD has been that, with rare exceptions, a song's lyrics are intended to be intelligible and so it is necessary that the singing voice is the dominant sound source in the final mixture. Therefore AD identified localised sub-band thresholds. These thresholds act as a "brick walls", and allow the inclusion of the individual frequency bins in the output only when their magnitudes exceed them.

The NIC subtraction makes use of FICA in order to extract two components from the original mixtures, one that contains most of the singing voice (i.e. VIC), and another that consists of a mixture of the remaining sources and a reduced share of the vocal (NIC). Despite of the inherent permutation problem of the yielded components in ICA, the output with less vocal (i.e. the NIC) can be successfully pinpointed due to its weak correlation with the original mixture, and subsequently subtracted from the estimated signal after AD.

For the purpose of objective evaluation of the separation, the *bss_eval* system has been used. Its metrics highlight three different aspects of the separation, namely, the source to interference ratio (SIR), the source to artefacts ratio (SAR), and a collective measure which is the signal to distortion ratio (SDR).

Due to lack of a widely available and standardised dataset fulfilling the requirements of the *bss_eval* metrics for the case of SVS, a database comprising songs available in multi-track format has been created as part of the study. The database includes song excerpts of varied genres and separation difficulty. The experimental results conducted with this database have shown that SEMANICS outperformed ADRess in SIR and SDR categories of assessment.

In the next chapter (Chapter 6) several disadvantages of SEMANICS have been

identified. These involve mainly the dependence of its performance on the user-set parameters, a shortcoming which is attributed primarily to ADRess. Certain modifications have been proposed in order to remove ADRess from the system, eliminating—thus—the aforementioned shortcomings. The proposed improved system has been termed SEMANTICS. The modifications include the more effective use of both original mixtures, the use of high-pass filtering (to improve the NIC scaling), the division of sub-bands according to the mel scale, and the introduction of a binary mask before the reconstruction of the time-signal. Based on sets of experimental evaluations, SEMANTICS has been found to perform better separation than SEMANICS, but—most importantly—it is found to offer performance consistency.

The H-SEMANTICS system, that was proposed in Chapter 7, includes the introduction of two novel time-domain procedures for music pruning and the integration of each of them with frequency-domain voice isolation, which is based on the enhancement of the previously established procedures. The modifications consist of an improved method for the estimation of thresholds and the restoration of the fundamental frequency that is lost due to high-pass filtering.

The first time-domain segregation method in H-SEMANTICS is named ENIC. This is based on energy of the signal and the NIC correlation. The second method, termed MICED, performs classification by comparing the distance between MFCCs of adjacent time segments and the two ICA outputs (i.e. the VIC and the NIC). In addition, MICED (as discussed in Chapter 7) is deemed appropriate for the implementation of a real-time SVS system.

The performance of the complete system based on each of the above music-pruning methods has been analysed and measured using a set of experimental investigations. The outcomes have clearly illustrated that the effectiveness in singing voice separation is considerably improved through the proposed approaches.

8.2 Suggestions for future work

The work presented in the context of this thesis does not provide a comprehensive solution to the challenge of unsupervised SVS. This section discusses some suggestions, which can extend the approaches that have been presented in this study. It should be noted that a number of these suggestions stem from research that has been performed during the programme of the study but were not mentioned in the main body of the thesis, as they did not lead to concrete results.

One of the main challenges across all systems discussed here is the estimation of the phase for the reconstruction of the time signal. In fact, this seems to be a common obstacle in most methods that perform spectral processing [299, 300]. Indeed, the room for improvement in performance is in the order of 4 dB. In other words, if the phase of the *a capella* is used in the reconstruction of the signal, the separation effectiveness can be increased by approximately 4 dB in SDR and SAR.

A further improvement that is proposed here is the automated selection of the ideal number of sub-bands. It seems that the ideal number of equal mel-spaced sub-bands for AD ranges from 3 to 5, as discussed in Chapters 6 and 7. This is attributed to a common technique that is used in the mixing process, namely multiband compression. This technique performs dynamic range compression on individual sub-bands that are selected manually by the mixing engineers. As the parameters of compression are different for each sub-band, this results in very distinguishable sub-band "cross-points" that could be automatically identified. This is envisaged to significantly improve the separation of the proposed systems in this thesis.

Since an important aspect of SVS is to facilitate applications of MIR, it would be beneficial to be able to measure its performance with respect to the level of aid that it provides to such applications. For example, existing methods that

perform singer identification or singer melody extraction [301] could be measured in terms of effectiveness with and without SVS. This would also provide a more usable metric than *bss_eval* and would give valuable feedback in order to improve specific aspects of the proposed systems. In fact, a glance of this concept is given in Chapter 7 with regards to the f_0. It has been found that the estimation of f_0 by the *MIRToolbox* is significantly better *after* H-SEMANTICS has separated the singing voice. Extending the aforementioned concept, the successful f_0 estimation after separation could provide important melodic cues for the purpose of note-based source separation [302].

Another area that could be promising is that of sinusoidal or pitch-tracking [303]using a standard algorithm like the McAuley and Quartieri [304]. Indeed, the vocal sinusoids were visually distinguishable in spectrograms produced after the separation with H-SEMANTICS. This has been observed during the experimental investigations.

Although the target mixture of this project has been the studio-produced stereophonic recordings, SEMANTICS and H-SEMANTICS have shown separation efficiency in live recordings as well. However, due to the nature of live recordings, the *a capella* and instrumental tracks are very difficult to acquire without cross-contamination. Therefore, the metrics of *bss_eval* cannot produce accurate results. An area of future research could be the development of a robust metric that measures the separation efficiency when SVS addresses live recordings. The interesting challenge in this case is that live recordings have sources that usually exhibit delay between observations due to non-coincidental[27] microphone techniques. Taking into consideration the aforementioned, a large standardised database should be developed that could help researchers to compare their methods with a common reference point. Finally, future work is envisaged to include further enhancement of MICED in order to minimise the false alarm rate with the use of probabilistic distances.

[27] Non-coincidental are the microphone techniques that produce amplitude, as well as timing differences between observations/channels.

RELATED PUBLICATIONS

S. Sofianos, A. M. Ariyaeeinia, R. Polfreman, and R. Sotudeh "H-SEMANTICS: A hybrid approach to singing voice separation," in *Journal of the Audio Engineering Society (JAES)*, vol. 60, pp. 831-841, 2012.

S. Sofianos, A. M. Ariyaeeinia, and R. Polfreman, "Singing voice separation based on non-vocal independent component subtraction and amplitude discrimination," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010, pp. 221-224. (ISBN: 978-3-200-01940-9)

S. Sofianos, A. M. Ariyaeeinia, and R. Polfreman, "Towards effective singing voice extraction from stereophonic recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, 2010, pp. 233-236. (ISBN: 978-1-4244-4296-6)

A
APPENDIX

A.1. Dependence of mixtures and heuristic approaches

Consider two random vectors s_1 and s_2 with uniform distributions, unit variance, and zero mean. As can be seen from Figure A.1 (a), their joint density is a "square". This derives from the theory that the joint density of two independent variables is the product of their marginal densities [305].

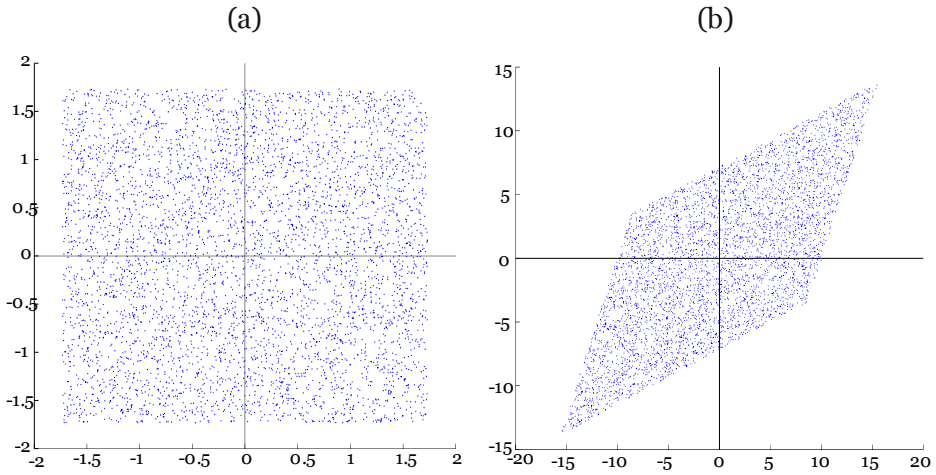

Figure A.1: Joint distribution of (a) two sources (i.e. s_1 and s_2) with uniform distributions ($\mu \approx 0, \sigma \approx 1$) and (b) their mixtures (i.e. x_1 and x_2)

If these two vectors/sources are mixed using the mixing matrix from Figure

3.2, this will result in two new mixtures x_1 and x_2 that—in a stereo recording—would represent the left and the right channels accordingly. Their joint distribution can be seen in Figure A.1 (b). Note that *the mixtures are not statistically independent*, because if the value of x_1 is known it is easy to predict the x_2. Without doubt, if x_1 reaches its minimum or maximum range limit, this completely determines the value of X_2. Furthermore, it can be seen that in Figure A.1 (b) that the *edges* of the parallelogram are in the direction of the mixing matrix that is used. However, such heuristic methods would only work for uniformly distributed sources (which is a highly unrealistic scenario), and it would be of high complexity (i.e. computational cost). The sought-after method for estimating the mixing matrix should be fast and reliable.

A.2. Othogonalisation approaches for FICA

This method of deflationary orhogonalisation is related to the Gram-Schmidt process [191] and in the case of FastICA it means that the IC's are estimated one at a time. In practice, N independent components (or to be accurate $\mathbf{w}_N$ vectors) are estimated. Then, the one-component algorithm (Subsection 3.4.6) runs for $\mathbf{w}_{i+1}$ and after every iteration step the "projections" $\mathbf{w}_{i+1}^T \mathbf{w}_j \mathbf{w}_j$, $j = 1, \dots p$ are subtracted from $\mathbf{w}_{i+1}$. Subsequently, $\mathbf{w}_{i+1}$ is re-normalised:

$$Let\ \mathbf{w}_{i+1} = \mathbf{w}_{i+1} - \sum_{j=1}^{i} \mathbf{w}_{i+1}^T \mathbf{w}_j \mathbf{w}_j, \text{and} \tag{A.1}$$

$$let\ \mathbf{w}_{i+1} = \mathbf{w}_{i+1} - \sqrt{\mathbf{w}_{i+1}^T - \mathbf{w}_{i+1}} \tag{A.2}$$

1. Choose the number of ICs to estimate (say N which is usually equal to number of mixtures). $i \leftarrow 1$.
2. Pick a random value for $\mathbf{w}_i$.
3. Do an iteration for one IC
4. Orthogonalise as in (A.1)-(A.2)
5. Let $\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\|$
6. If not converged, go back to step 3
7. $i \leftarrow i+1$

Table A.1: Estimation for more than one IC using deflationary orthogonalisation

Although the deflationary method produces satisfactory results [162], the disadvantage that estimation errors of the first vector **b** are carried on to subsequent estimations through orthogonalisation. Hence a way of tackling this problem is the symmetric orthogonalisation approach:

1. Choose the number of ICs to estimate (say N which is usually equal to number of mixtures).
2. Pick a random value for $\mathbf{w}_i$
3. Do an iteration for one IC *in parallel.*
4. Do a symmetric orthogonalisation of the matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_N]$ as in (A.3)
5. If not converged, go back to step 3

Table A.2: Estimation for more than one IC using symmetric orthogonalisation

This method is described as the case where no vectors are "privileged" over

the others and it can have desirable results in certain applications [306]. The vectors **b** are not estimated one by one but rather in parallel.

The classic method for symmetric orthogonalisation involves matrix square roots [307]:

$$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W} \tag{A.3}$$

In addition to tackling the first component estimation problem that is encountered in the deflationary method, the symmetric approach can also efficiently exploit the computational power of multi-core systems.

A.3. Collective *bss_eval* results

	Baseline	ADRess	SEMANICS	SEMANTICS	H-S MICED	H-S ENIC
Salala	-1.31	0.14	8.19	8.09	8.60	8.63
Nude	-4.21	-1.17	6.24	7.58	8.13	8.36
Kunlarim	-3.72	-3.44	-1.57	1.14	1.62	2.00
Help Me	-0.87	-0.20	2.19	5.12	5.57	5.72
Only	-4.81	-4.41	0.12	1.35	2.72	2.82
Resistencia	-6.75	-6.62	-1.13	0.35	2.15	2.15
Americano	0.11	1.12	4.85	4.24	4.62	5.28
Monkey	-2.95	-2.68	1.66	4.95	6.22	7.13
Roads	-5.45	-5.52	2.22	2.97	4.80	4.86
Tanto (1)	0.16	1.04	6.00	7.94	8.60	8.89
Tanto (2)	-1.73	-0.08	4.60	5.72	6.36	6.47
Don't Know	-10.44	-10.53	-6.08	-2.83	-1.49	-1.20

Table A.3: Absolute SDR (dB) for all systems rounded to two decimals

	Baseline	ADRess	SEMANICS	SEMANTICS	H-S MICED	H-S ENIC
Salala	-1.31	1.15	23.35	28.98	30.92	30.39
Nude	-4.17	0.37	24.58	29.81	29.95	30.25
Kunlarim	-3.72	-3.22	2.84	4.74	6.40	6.30
Help Me	-0.56	0.28	8.47	12.01	13.37	13.34
Only	-4.81	-4.22	6.45	7.40	10.05	10.64
Resistencia	-6.75	-6.44	7.04	8.95	12.59	12.59
Americano	0.11	1.82	14.73	15.55	19.51	18.82
Monkey	-2.92	-2.65	11.4	13.00	16.03	18.44
Roads	-5.43	-5.08	15.64	18.77	21.78	21.80
Tanto (1)	0.16	2.46	11.8	18.09	20.32	19.98
Tanto (2)	-1.71	1.56	14.38	17.20	17.78	17.61
Don't Know	-10.2	-10.28	2.98	9.10	13.85	13.67

Table A.4: Absolute SIR (dB) for all systems rounded to two decimals

	Baseline	ADRess	SEMANICS	SEMANTICS	H-S MICED	H-S ENIC
Salala	33.43	9.44	8.34	8.14	8.63	8.66
Nude	23.02	6.92	6.32	7.61	8.16	8.39
Kunlarim	30.17	14.54	2.19	4.88	4.27	4.92
Help Me	14.02	12.45	3.94	6.38	6.56	6.75
Only	38.06	14.77	2.15	3.31	4.02	3.96
Resistencia	39.95	14.66	0.36	1.52	2.80	2.80
Americano	35.70	11.57	5.46	4.69	4.81	5.53
Monkey	24.43	23.72	2.45	5.90	6.80	7.52
Roads	24.00	10.88	2.53	3.15	4.91	4.97
Tanto (1)	78.07	8.54	7.58	8.44	8.94	9.28
Tanto (2)	26.44	7.26	5.24	6.12	6.76	6.89
Don't Know	12.86	12.78	-3.74	-2.04	-1.19	-0.87

Table A.5: Absolute SAR (dB) for all systems rounded to two decimals

	Tanto			Roads		
System	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Ozerov, Favotte	3.6	5.5	8.1	-3.0	0.0	-0.8
Ozerov	5.1	6.9	9.8	2.5	4.7	5.3
Durrieu (1)	7.8	18.9	8.2	N/A	N/A	N/A
Durrieu (2)	6.9	17.2	7.0	N/A	N/A	N/A
Cobos	6.4	19.5	6.5	2.5	15.1	1.5
Vinyes Raso	4.9	26.7	5.3	3.0	10.4	2.1
SEMANICS	6.0	11.8	7.6	2.2	15.6	2.5
SEMANTICS	7.9	18.1	8.4	3.0	18.8	3.2
H-S MICED	8.6	20.3	8.9	4.8	21.8	4.9
H-S ENIC	8.9	20.0	9.3	4.9	21.8	5.0

Table A.6: Absolute SDR, SIR, and SAR (dB) rounded to one decimal for all unsupervised systems competing in SiSEC [282] compared with the systems in this thesis.

A.4. Formulae derivation

Equation (7.2):

The standard tapered cosine (i.e. Tukey) window is defined by [290, 308]:

$$w(k) = \begin{cases} \frac{1}{2}\left[1+\cos\left(\pi\left(\frac{2k}{\mathrm{aN}}-1\right)\right)\right] & when\ 0 \le k \le \frac{\mathrm{aN}}{2} \\ 1 & when\ \frac{\mathrm{aN}}{2} \le k \le \mathrm{N}\left(1-\frac{\mathrm{a}}{2}\right) \\ \frac{1}{2}\left[1+\cos\left(\pi\left(\frac{2k}{\mathrm{aN}}-\frac{2}{\mathrm{a}}+1\right)\right)\right] & when\ \mathrm{N}\left(1-\frac{\mathrm{a}}{2}\right) \le k \le \mathrm{N}, \end{cases} \quad \text{(A.4)}$$

where N is the length of the window, and 'a' defines the ratio of the tapered length to the length of the flat section.

In order to modify the above for defined corners, 'a' is set to 1, and N becomes $2(c_2^m - c_1^m)$ for the first tapered section of the window (i.e. the ascending part), and $2(c_4^m - c_3^m)$ for the descending tapered section as shown in Figure 7.2 (i.e. c_3^m to c_4^m). Accordingly, k is offset by c_1^m in the first case and by $c_3^m + (c_4^m - c_3^m)$ in the second case. Thus, equation (A.4) for m = 1, 2, ... M sub-bands becomes:

$$w(k,m) = \begin{cases} \frac{1}{2}\left[1+\cos\left(\pi\left(\frac{k-c_1^m}{c_2^m-c_1^m}-1\right)\right)\right] & when\ c_1^m \le k < c_2^m \\ 1 & when\ \ c_2^m \le k \le c_3^m \\ \frac{1}{2}\left[1+\cos\left(\pi\left(\frac{k-2c_3^m+c_4^m}{c_4^m-c_3^m}-1\right)\right)\right] & when\ \ c_3^m < k \le c_4^m. \end{cases} \quad \text{(A.5)}$$

The window should also be of unit weight since it is needed to calculate the mean of the section c_1^m to c_4^m. In order to calculate the weight let:

$$L_1 = c_2^m - c_1^m - 1, \quad \text{(A.6)}$$

$$L_2 = c_3^m - c_2^m, \text{and} \quad \text{(A.7)}$$

$$L_3 = c_4^m - c_3^m - 1. \quad \text{(A.8)}$$

Since L_1 and L_3 are the tapered sections, by definition their sums are N / 2 where N is their respective length. Thus:

$$x\frac{L_1}{2} + xL_2 + x\frac{L_3}{2} = L_1 + L_2 + L_3, \quad \text{(A.9)}$$

where *x* is a weight. Solving this equation gives:

$$x = \frac{2(L_1 + L_2 + L_3)}{(L_1 + L_2 + L_3) + L_2} = \frac{2(c_4^m - c_1^m)}{c_4^m - c_1^m + c_3^m - c_2^m}. \tag{A.10}$$

Applying this weight to each part of (A.5) gives equation (7.2).

Equation (7.3):

Linear- to mel-scale frequency conversion [286]:

$$g(x) = 2595\, log_{10}(1 + \frac{x}{700}). \tag{A.11}$$

Mel- to linear-frequency conversion [286]:

$$f(x) = 700\left(10^{\frac{x}{2595}} - 1\right). \tag{A.12}$$

Each equal mel sub-band has a width of:

$$\Delta = \frac{g(\mathrm{S_R}/2)}{\mathrm{M}}, \tag{A.13}$$

where M is the number of sub-bands and S_R is the sampling rate.

Sub-band corners in the mel domain:

$$Q_p^m = \Delta\big(m - H(p)\big) \quad \forall m, p, \tag{A.14}$$

where $H(p)$ = [1+α, 1, 0, –α], while α is the percentage of overlap, m is the sub-band index, and p is the corner index.

Wrapping mels to FFT bins:

$$c_p^m = f\left(Q_p^m\right)\frac{\mathrm{K}}{\mathrm{S_R}} = c_p^m = 700\left(10\ \frac{\Delta\left(m - H(p)\right)}{2595} - 1\right)\frac{\mathrm{K}}{\mathrm{S_R}}. \tag{A.15}$$

where K is the FFT size. Simplify:

$$10\ \frac{\Delta\left(m - H(p)\right)}{2595} = \tag{A.16}$$

$$10\ \frac{2595\ log_{10}(1 + \frac{\mathrm{S_R}}{1400})\left(m - H(p)\right)}{2595\mathrm{M}} = \left(1 + \frac{\mathrm{S_R}}{1400}\right)^{\frac{m-H(p)}{\mathrm{M}}}.$$

Therefore:

$$c_p^m = \frac{700\mathrm{K}\left(\left(1 + \frac{\mathrm{S_R}}{1400}\right)^{\frac{m-H(p)}{\mathrm{M}}} - 1\right)}{\mathrm{S_R}}. \tag{A.17}$$

Equation (7.7):

The PMCC between $\hat{S}_v(k)$ and $G_v(k)$ is given by:

$$\rho\left(\hat{S}_v(k), G_v(k)\right) = \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}} \frac{(\hat{S}_v(k) - \mu_{\hat{S}_v})(G_v(k) - \mu_{G_v})}{\sigma_{\hat{S}_v}\sigma_{G_v}}. \tag{A.18}$$

If the modulus of (A.18) is multiplied by 1 / γ it will give an amplifying gain for $\rho > \gamma$ and an attenuating gain for $\rho < \gamma$:

$$\rho\left(\hat{S}_v(k), G_v(k)\right) = \left|\frac{1}{\gamma\mathrm{K}}\sum_{k=1}^{\mathrm{K}} \frac{(\hat{S}_v(k) - \mu_{\hat{S}_v})(G_v(k) - \mu_{G_v})}{\sigma_{\hat{S}_v}\sigma_{G_v}}\right|, \tag{A.19}$$

where $\rho\big(\hat{S}_v(k), G_v(k)\big) \in \left[0, \frac{1}{\gamma}\right]$, $\gamma > 0$.

The initial threshold is calculated by subtracting the standard deviation from the mean of the vector $R_{\hat{s}}(v)$. Subsequently, this is modulated by (A.19) as follows:

$$Th(\hat{s}_v) = \big(\mu_{R_{\hat{s}}} - \sigma_{R_{\hat{s}}}\big)\hat{\rho}\big(\hat{S}_v(k), G_v(k)\big). \tag{A.20}$$

This gives a threshold such that $\hat{s}_v$ is classified as vocal when $R_{\hat{s}}(v) > Th(\hat{s}_v)$, and music-only when $R_{\hat{s}}(v) < Th(\hat{s}_v)$. By definition, $R_{\hat{s}}(v)$ is in the range [0, 1]. In order to have the same bounds as (A.19), $R_{\hat{s}}(v)$ is divided by γ. Therefore the decision:

$$\check{p}(\hat{s}_v) = \frac{R_{\hat{s}}(v)}{\gamma} - \big(\mu_{R_{\hat{s}}} - \sigma_{R_{\hat{s}}}\big)\hat{\rho}\left(\hat{S}_v(k), G_v(k)\right), \tag{A.21}$$

where $\breve{p}(\hat{s}_v) \in [-\gamma, \gamma]$ indicates vocal segment when positive, and music-only segment otherwise.

For an intuitive result, bounds are normalised to [0, 1] for a music-only classification when $p(\hat{s}_v) < 0.5$. Therefore (A.21) becomes (7.7).

A.5. CD Contents

The CD included with this thesis contains audio examples from the three novel methods described in this thesis, namely SEMANICS, SEMANTICS, and H-SEMANTICS. The examples are in in the form of .wav files and they comprise the original mixtures, the *a capella*, and the estimated output from their respective systems. A short description of the examples is also included in the CD in the form of a text file.

REFERENCES

[1] A. S. Bregman, *Auditory scene analysis : the perceptual organization of sound*, 1st ed. Cambridge, MA, USA ; London, UK: MIT Press, 1990.

[2] D. Byrd and T. Crawford, "Problems of music information retrieval in the real world," *Information Processing & Management,* vol. 38, pp. 249-272, 2002.

[3] M. Markaki, A. Holzapfel, and Y. Stylianou, "Singing Voice Detection using Modulation Frequency Features," in *Proceedings of the ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, Brisbane, Australia, 2008, pp. 7-10.

[4] Y. Li and D. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 1475-1487, 2007.

[5] M. Rocamora and P. Herrera-Boyer, "Comparing audio descriptors for singing voice detection in music audio files," presented at the 11th Brazilian Symposium on Computer Music, Sao Paulo, Brazil, 2007.

[6] B. S. Brook, *Thematic catalogues in music, an annotated bibliography : including printed, manuscript, and in-preparation catalogues; related literature and reviews; an essay on the definitions, history, functions, historiography, and future of the thematic catalogue*. Hillsdale, NY, USA: Pendragon Press, 1972.

[7] H. Barlow and S. Morgenstern, *A dictionary of musical themes*. London, UK: Williams & Norgate, 1949.

[8] L. R. v. Köchel, *Chronologisch-thematisches Verzeichnis sämmtlicher Tonwerke Wolfgang Amade Mozart's*. Leipzig, Germany: Breitkopf & Härtel, 1862.

[9] W. Schmieder, A. Dürr, and Y. Kobayashi, *Bach-Werke-Verzeichnis*. Leipzig, Germany: Breitkopf & Härtel, 1998.

[10] M. Kassler, "Toward Musical Information Retrieval," *Perspectives of New Music,* vol. 4, pp. 59-67, Spring - Summer 1966.

[11] F. Yazhong, Z. Yueting, and P. Yunhe, "Popular Song Retrieval Based on Singing Matching," in *Proceedings of the IEEE Pacific Rim Conference on Multimedia (PCM): Advances in Multimedia Information Processing*, Hsinchu, Taiwan, 2002, pp. 639-646.

[12] Y. E. Kim and B. Whitman, "Singer Identification in Popular Music Recordings Using Voice Coding Features," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[13] T. Hosoya, M. Suzuki, A. Ito, and S. Makino, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[14] S. Z. K. Khine, T. L. Nwe, and L. Haizhou, "Singing voice detection in pop songs using co-training algorithm," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 1629-1632.

[15] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 169-172.

[16] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 375-378.

[17] J.-J. Aucouturier and F. Pachet, "Improving Timbre Similarity: How high's the sky?," *Journal on Negative Results in Speech and Audio Sciences,* vol. 1, 2004.

[18] J.-J. Aucouturier, F. Pachet, P. Roy, and A. Beurive, "Signal + Context=Better Classification," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[19] C. Anderton, *MIDI for musicians*. New York City, NY, USA: Amsco Publications, 1986.

[20] J. Foote, "An overview of audio information retrieval," *Multimedia Systems,* vol. 7, pp. 2-10, 1999.

[21] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-Based Music Information Retrieval: Current Directions and Future Challenges," *Proceedings of the IEEE,* vol. 96, pp. 668-696, 2008.

[22] H. v. Helmholtz and A. J. Ellis, *On the sensations of tone as a physiological basis for the theory of music*. London, UK: Longmans, Green, and Co., 1863.

[23] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *The Journal of the Acoustical Society of America,* vol. 25, pp. 975–979, 1953.

[24] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic Music Transcription and Audio Source Separation," *Cybernetics and Systems,* vol. 33, pp. 603-627, 2002.

[25] J. McDermott, "The cocktail party problem," *Current Biology,* vol. 19, pp. 1024-1027, 2009.

[26] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters,* vol. 22, pp. 533-544, 2001.

[27] H. Ezzaidi and M. Bahoura, "Voice singer detection in polyphonic music," in *Proceedings of the IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, Medina, Tunisia, 2009, pp. 884-887.

[28] A. S. Malegaonkar, A. M. Ariyaeeinia, and P. Sivakumaran, "Efficient Speaker Change Detection Using Adapted Gaussian Mixture Models," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 15, pp. 1859-1869, 2007.

[29] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing,* vol. 55, pp. 627-641, 2003.

[30] P. Comon, "Blind identification and Source Separation in 2 x 3 Under-determined mixtures," *IEEE Transactions on Signal Processing,* vol. 52, pp. 11-22, January 2004.

[31] K. K. Govindarajan, *A neural network model of auditory scene analysis and source segregation*. Boston, MA: Boston University, Center for Adaptive Systems and Dept. of Cognitive and Neural Systems, 1994.

[32] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, 1st ed. Hoboken, NJ, USA ; Canada: John Wiley & Sons, 2006.

[33] S. O. Haykin and Z. Chen, "The Cocktail Party Problem," *Neural Computation,* vol. 17, pp. 1875-1902, 2005.

[34] R. Polfreman, "A task analysis of music composition and its application to the development of Modalyser," *Organised Sound,* vol. 4, pp. 31-43, 1999.

[35] T. Macho, "Weltenlärm, Schweigen, Stille,," *Österreichische Musikzeitschrift,* vol. 7/8, pp. 440-445, 1994.

[36] J. Cage, "4'33''," ed, 1952.

[37] T. W. Adorno, *Philosophy of modern music*. London, UK: Sheed & Ward, 1987.

[38] M. Goto. (2008, 19th March). *RWC Music Database*. Available: http://staff.aist.go.jp/m.goto/RWC-MDB/

[39] M. Goto and H. Hashiguchi, "RWC Music Database: Popular, Classical, and Jazz Music Databases," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[40] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, "A Benchmark Dataset for Audio Classification and Clustering," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[41] (2004). *http://ismir2004.ismir.net/melody_contest/results.html*.

[42] S. Vembu and S. Baumann, "Separation of Vocals from Polyphonic Audio Recordings," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[43] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, pp. 1462-1469, July 2006.

[44] G. Erten and F. M. Salam, "Voice extraction by on-line signal separation and recovery," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing,* vol. 46, pp. 915-922, July 1999.

[45] Y. Li and D. Wang, "Singing Voice Separation from Monaural Recordings," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Victoria, BC, Canada, 2006.

[46] P. Herrera-Boyer, "Setting up an Audio Database for Music Information Retrieval Benchmarking," in *The MIR/MDL Evaluation Project White Paper Collection*, 2002.

[47] E. M. Voorhees, "Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC," in *The MIR/MDL Evaluation Project White Paper Collection*, 2003.

[48] (2003). *The MIR/MDL Evaluation Project White Paper Collection*. Available: http://www.music-ir.org/evaluation/wp.html

[49] J. Sundberg, *The acoustics of the singing voice* vol. 236 no. 3. New York City, NY, USA: Scientific American, 1977.

[50] C. Darwin, *On the origin of species: By means of natural selection, or The preservation of favoured races in the struggle for life*, Reissue ed. USA: Bantam Classics, 1999.

[51] D. E. Callan, V. Tsytsarev, T. Hanakawa, A. M. Callan, M. Katsuhara, H. Fukuyama, and R. Turner, "Song and speech: Brain regions involved with perception and covert production," *NeuroImage,* vol. 31, pp. 1327-1342, 2006.

[52] J. W. van den Berg, *Physiology on physics of voice production*: Acta Physiol Pharmacol Neerlandica, 1956.

[53] I. R. Titze, *The Myoelastic Aerodynamic Theory of Phonation*, 1st ed.: National Center for Voice and Speech, 2006.

[54] U. Michels and G. Vogel, *dtv-Atlas Musik*: dtv, 2005.

[55] Y. E. Kim, "Singing Voice Analysis/Synthesis," Ph.D., School of Architecture and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA, 2003.

[56] H. Chanson, *Applied Hydrodynamics: An Introduction to Ideal and Real Fluid Flows*. Leiden, Netherlands: CRC Press, 2009.

[57] D. E. Hall, *Musical Acoustics*, 2nd ed. Belmont, CA, USA: Brooks/Cole Publishing, 1991.

[58] A. Laukkanen and T. Leino, *Ihmeellinen ihmisääni (The Wonderful Human Voice)*. Tampere, Finland: Gaudeamus, 1999.

[59] U. G. Goldstein, "An articulatory model for the vocal tracts of growing children," Ph.D., Massachusetts Institute of Technology, Cambridge, MA, USA, 1980.

[60] E. Joliveau, J. Smith, and J. Wolfe, "Tuning of vocal tract resonance by sopranos," *Nature,* vol. 427, p. 116, 8th January 2004.

[61] J. Sundberg, *The science of the singing voice*. DeKalb, IL, USA: Northern Illinois University Press, 1987.

[62] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands, 1960.

[63] John, *[1:1] Genesis*: New Revised Standard Version.

[64] H. Spencer. (1857) The Origin and Function of Music. *Fraser*.

[65] M. Greene and J. Conway, *Learning to Talk: A Study in Sound of Infant Speech Development*. New York City, NY, USA: Folkways Records, 1963.

[66] W. Enard, M. Przeworski, S. E. Fisher, C. S. L. Lai, V. Wiebe, T. Kitano, A. P. Monaco, and S. Paabo, "Molecular evolution of FOXP2, a gene involved in speech and language," *Nature,* vol. 418, pp. 869-872, 2002.

[67] O. Jespersen, *Language; its nature, development and origin*. London, UK: Allen & Unwin, 1922.

[68] A. Loscos, "Spectral Processing of the Singing Voice," Ph.D., Information and Communication Technologies, University Pompeu Fabra, Barcelona, Spain, 2007.

[69] K. Nef, *Einführung in die Musikgeschichte (Introduction to Music History)*. Basel, Switzerland: Kober, 1920.

[70] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, "Voice-selective areas in human auditory cortex," *Nature,* pp. 309-311, 2000.

[71] "Voice Frequency," in *McGraw-Hill Dictionary of Scientific & Technical Terms*, 6th ed: The McGraw-Hill Companies, Inc, 2003.

[72] *Equal-loudness contour,* ISO Standard 226:2003.

[73] R. B. Dannenberg, "Music Representation Issues, Techniques, and Systems," *Computer Music Journal,* vol. 17, pp. 20-30, Autumn 1993.

[74] Y.-R. Chien, H.-M. Wang, and S.-K. Jeng, "An Acoustic-Phonetic Approach to Vocal Melody Extraction," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Miami, FL, USA, 2011.

[75] V. Rao and P. Rao, "Vocal Melody Extraction in the Presence of Pitched Accompaniment in Polyphonic Music," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, pp. 2145-2154, November 2010.

[76] V. Rao, S. Ramakrishnan, and P. Rao, "Singing Voice Detection in Polyphonic Music using Predominant Pitch," in *Proceedings of Interspeech*, Brighton, UK, 2009.

[77] Z. Xun and T. Francesca, *Karaoke: A Global Phenomenon*, 1st ed. London, UK: Reaktion Books, 2007.

[78] P. Iyer. (1999, August 23-30) Daisuke Inoue. *Time 100*.

[79] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2003, pp. 55-58.

[80] M. Bell. (2002, April) How can I isolate the vocals from a stereo mix? *Sound on Sound*.

[81] H. Robjohns. (2004, September) Can I remove the vocals from a track using phase? *Sound on Sound*.

[82] S. A. Gelfand, *Hearing : an introduction to psychological and physiological acoustics*, 4th ed. New York City, NY, USA: Marcel Dekker, 2004.

[83] (2011, 01/11/2011). *Adobe Audition http://www.adobe.com/products/audition.html*.

[84] (2008, 25/03/2008). *http://www.celemony.com/cms/index.php?id=358*.

[85] (2008, 26/03/2008). *http://www.antarestech.com/products/auto-tune5.shtml*. Available: http://www.antarestech.com/products/auto-tune5.shtml

[86] A. Friberg, E. Schoonderwaldt, and P. N. Juslin, "CUEX: An algorithm for automatic extraction of expressive tone parameters in music performance from acoustic signals," *Acta acustica united with acustica,* vol. 93, p. 10, 2007.

[87] S. Pant, V. Rao, and P. Rao, "A melody detection user interface for polyphonic music," in *Proceedings of the National Conference on Communications (NCC)*, Madras, India, 2010, pp. 1-5.

[88] F. Holm, "NAMM 2009," *Computer Music Journal,* vol. 33, pp. 61-64, 2009.

[89] M. Vinyes, J. Bonada, and A. Loscos, "Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking," presented at the 120th Audio Engineering Society (AES) Convention, Paris, France, 2006.

[90] (2010, 21/08/2010). *AudioScanner http://mtg.upf.edu/static/mass/software/index.htm*.

[91] A. Wang, "The Shazam Music Recognition Service," *Communications of the Association for Computing Machinery (CASM),* vol. 49, pp. 44-48, 2006.

[92] A. Wang, "An industrial-strength audio search algorithm," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, MD, USA, 2003, pp. 713-718.

[93] (2011, 12/2011). *Hit'n'Mix DJ Mashup Software*. Available: http://www.hitnmix.com/

[94] A. Cox. (2011) Hit 'n' Mix review. *PCPlus* [Review].

[95] A. Narayanan and D. Wang, "Robust speech recognition from binary masks," *The Journal of the Acoustical Society of America,* vol. 128, pp. EL217-EL222, 2010.

[96] B. H. Juang, "Speech recognition in adverse environments," *Computer Speech & Language,* vol. 5, pp. 275-294, 1991.

[97] J. Sundberg, *The Science of Musical Sounds*: Academic Press, 1992.

[98] R. Miller, *On the art of singing*. New York City, NY, USA: Oxford University Press, 1996.

[99] L. Feng, A. B. Nielsen, and L. K. Hansen, "Vocal Segment Classification in Popular Music," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Philadelphia, PA, USA, 2008.

[100] P. Rao, "Musical Information Extraction from the Singing Voice," in *Proceedings of the IET National Conference on Signal and Image Processing Applications*, Pune, Maharashtra, India, 2009.

[101] D. Wang, "Feature-based speech segregation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds., ed New York City, NY, USA: IEEE Press, 2006, pp. 81-114.

[102] Y. E. Kim, "Structured Encoding of the Singing Voice using Prior Knowledge of the Musical Score," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 1999.

[103] J. M. Fisher, *Grosse Stimmen*. Stuttgart, Germany: Metzler Music, 1993.

[104] R. Celletti, *A History of Bel Canto*. USA: Oxford University Press, 1997.

[105] T. Saitou, M. Unoki, and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication,* vol. 46, pp. 405-417, 2005.

[106] A. Lamarche, S. Trenström, and P. Pabon, "The Singer's Voice Range Profile: Female Professional Opera Soloists," *Journal of Voice,* vol. 24, pp. 410-426, 2010.

[107] A. Meribeth, *Dynamics of the Singing Voice*. Berlin, Germany: Springer, 2009.

[108] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," ed, 1993.

[109] S. C. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters," *IEEE Transactions on Speech and Audio Processing,* vol. 13, pp. 92-104, 2005.

[110] J. Naylor and J. Porter, "An effective speech separation system which requires no a priori information," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, ON, Canada, 1991, pp. 937-940.

[111] R. R. Fay, "Auditory Scene Analysis," *Bioacoustics,* vol. 17, pp. 106-109, 2008.

[112] W. L. Gulick, *Hearing : physiology and psychophysics*. New York City, NY, USA ; London, UK: Oxford University Press, 1971.

[113] S. A. Shamma and C. Micheyl, "Behind the scenes of auditory perception," *Current Opinion in Neurobiology,* vol. 20, pp. 361-366, 2010.

[114] B. Roberts, B. R. Glasberg, and C. J. Moore, "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," *The Journal of the Acoustical Society of America,* vol. 112, 2002.

[115] W. DeLiang, "Primitive Auditory Segregation Based On Oscillatory Correlation," *Cognitive Science,* vol. 20, pp. 409-456, 1996.

[116] D. Huron, "Auditory Scene Analysis: The Perceptual Organization of Sound," *Psychology of Music,* vol. 19, pp. 77-82, 1991.

[117] K. Koffka, *Principles of Gestalt psychology*. London, UK: Routledge & Kegan Paul, 1955.

[118] W. Köhler, *Gestalt psychology*. New York City, NY, USA: Liveright, 1929.

[119] C. Bey and S. McAdams, "Schema-based processing in auditory scene analysis," *Attention, Perception, & Psychophysics,* vol. 64, pp. 844-854, 2002.

[120] M. Hartmann W, "Pitch Perception and the Segregation and Integration of Auditory Entities," *Auditory Function, Neurobiological Bases of Hearing,* pp. 623-645, 1988.

[121] G. J. Brown and M. Cooke, "Temporal synchronization in a neural oscillator model of primitive auditory stream segregation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, F. R. David and G. O. Hiroshi, Eds., ed: L. Erlbaum Associates Inc., 1998, pp. 87-103.

[122] M. C. Teich, "Fractal neuronal firing patterns," in *Single neuron computation*, ed Waltham, MA, USA: Academic Press Professional Inc., 1992, pp. 589-625.

[123] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D., Massachusetts Institute of Technology, Cambridge, MA, USA, 1996.

[124] G. J. Brown, "Computational auditory scene analysis: A representational approach," *The Journal of the Acoustical Society of America,* vol. 94, p. 2454, 1993.

[125] Z. Chen, "An odyssey of the cocktail party problem," Adaptive Systems Lab, McMaster University, Hamilton, ON, Canada, 2003.

[126] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication,* vol. 35, pp. 141-177, 2001.

[127] D. K. Mellinger, "Event Formation and Separation in Musical Sound," Ph.D., Department of Computer Science, Stanford University, Stanford, CA, 1991.

[128] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, USA, 2002, pp. 553-556.

[129] G. J. Brown and M. Cooke, "Perceptual grouping of musical sounds: A computational model," *Journal of New Music Research,* vol. 23, pp. 107-132, 1994.

[130] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D., Stanford University Stanford, CA, USA, 1985.

[131] R. F. Lyon, "A computational model of binaural localization and separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Boston, MA, USA, 1983, pp. 1148-1151.

[132] M. Bodden, "Modelling human sound-source localization and the cocktail party effect," *Acta Acoustica,* vol. 1, pp. 43-55, 1993.

[133] C. Liu, B. C. Wheeler, J. W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *The Journal of the Acoustical Society of America,* vol. 108, pp. 1888-1905, 2000.

[134] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America,* vol. 114, pp. 2236-2252, 2003.

[135] M. Cooke, *Modelling auditory processing and organisation*: Cambridge University Press, 1993.

[136] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, pp. 278-290, 2008.

[137] D. P. W. Ellis and D. Rosenthal, "Mid-level representations for Computational Auditory Scene Analysis," in *International Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada, 1995.

[138] D. P. W. Ellis, "A perceptual representation of audio," MSc, Electrical Engineering and Computer Science (EECS), Massachussets Institute of Technology (MIT), USA, 1992.

[139] M. Slaney, "Lyon's Cochlear Model," Apple Computer, Inc, 1988.

[140] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Paris, France, 1982, pp. 1982-1285.

[141] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics,* vol. 16, pp. 55-76, 1988.

[142] P. I. M. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," in *Proceedings of the Symposium on Hearing Theory*, Eindhoven, The Netherlands, 1972, pp. 58-69.

[143] S. Grossberg, K. K. Govindarajan, L. L. Wyse, and M. A. Cohen, "ARTSTREAM: a neural network model of auditory scene analysis and source segregation," *Neural Networks,* vol. 17, pp. 511-536, 2004.

[144] M. A. Cohen, S. Grossberg, and L. L. Wyse, "A spectral network model of pitch perception," *The Journal of the Acoustical Society of America,* vol. 498, pp. 862-879, 1995.

[145] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research,* vol. 47, pp. 103-138, 1990.

[146] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Communication,* 2007.

[147] F. van der Meero and W. Bakker, "Cross correlogram spectral matching: Application to surface mineralogical mapping by using AVIRIS data from Cuprite, Nevada," *Remote Sensing of Environment,* vol. 61, pp. 371-382, 1997.

[148] M. Slaney and R. F. Lyon, "A Perceptual Pitch Detector," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, NM, USA, 1990.

[149] K. D. Martin, "A Blackboard System for Automatic Transcription of Simple Polyphonic Music," M.I.T Media Laboratory Perceptual Computing Section, 1996.

[150] K. D. Martin, "Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing," Massachusetts Institute of Technology (MIT) Media Laboratory Perceptual Computing Section, Cambridge, MA, USA 399, 1996.

[151] T. Tolonen and M. Karjalainen, "A Computationally Efficient Multipitch Analysis Model," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 8, pp. 708-716, November 2000.

[152] O. Jensen, J. Kaiser, and J.-P. Lachaux, "Human gamma-frequency oscillations associated with attention and memory," *Trends in Neurosciences,* vol. 30, pp. 317-324, 2007.

[153] L. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology,* vol. 41, pp. 35-39, 1948.

[154] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication,* vol. 34, pp. 267-285, 2001.

[155] S. Yang and W. DeLiang, "Robust Speaker Recognition Using Binary Time-Frequency Masks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 645-648.

[156] A. M. Reddy and B. Raj, "Soft Mask Estimation for Single Channel Speaker Separation," in *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Jeju, Korea, 2004.

[157] W. DeLiang, "Time–Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design," *Trends in Amplification,* vol. 12, pp. 332-353, December 2008.

[158] S. G. Karadoğan, J. Larsen, M. S. Pedersen, and B. J. Bünsow, "Robust isolated speech recognition using binary masks," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, 2010, pp. 1988-1992.

[159] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication,* vol. 48, pp. 1486-1501, 2006.

[160] A. Kidjo and P. Gabriel, "Salala," in *Djin Djin*, ed: EMI France, 2007.

[161] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Adelaide, South Australia, Australia, 1994, pp. 77-80.

[162] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *AIP Conference Proceedings 151 on Neural Networks for Computing*, Snowbird, UT, USA, 1986.

[163] P. Comon and C. Jutten, *Handbook of Blind Source Separation* Burlington, MA, USA: Academic Press, 2010.

[164] P. Dinh-Tuan, "Fast algorithms for mutual information based independent component analysis," *IEEE Transactions on Signal Processing,* vol. 52, pp. 2690-2700, 2004.

[165] R. Mutihac and M. M. Van Hulle, "Comparison of Principal Component Analysis and Independent Component Analysis for Blind Source Separation," *Romanian Reports in Physics,* vol. 56, 2004.

[166] A. Cichocki and L. Moszczynski, "New learning algorithm for blind separation of sources," *IET Electronics Letters,* vol. 28, pp. 1986-1987, 1992.

[167] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. New York City, NY, USA: Chichester : Wiley, 2001.

[168] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE,* vol. 86, pp. 2009-2025, 1998.

[169] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined Instantaneous Audio Source Separation via Local Gaussian Modeling," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brazil, 2009, pp. 775-782.

[170] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete bss for convolutive mixtures based on hierarchical clustering," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Granada, Spain, 2004.

[171] T.-W. Lee, A. J. Bell, and R. H. Lambert, "Blind Separation of Delayed and Convolved Sources," *Advances in Neural Information Processing Systems (NIPS),* vol. 9, pp. 758-764, 1997.

[172] A. Nesbit, E. Vincent, and M. D. Plumbley, "Extension of sparse, adaptive signal decompositions to semi-blind audio source separation," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brazil, 2009, pp. 605-612.

[173] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind Source Separation and Independent Component Analysis: A Review," *Neural Information Processing - Letters and Reviews,* vol. 6, 2005.

[174] D. Maino, A. Farusi, C. Baccigalupi, F. Perrotta, A. J. Banday, L. Bedini, C. Burigana, G. De Zotti, K. M. Gorski, and E. Salerno, "All-sky astrophysical component separation with fast independent component analysis (FASTICA)," *Monthly Notices of the Royal Astronomical Society,* vol. 334, pp. 53-68, Jul 2002.

[175] C. J. James and O. Gibson, "ICA with a reference: extracting desired electromagnetic brain signals," *IEE Seminar Digests,* vol. 2002, p. 4, 2002.

[176] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine,* vol. 2, pp. 559–572, 1901.

[177] T. W. Anderson, *An introduction to multivariate statistical analysis*. J. Wiley: New York City, NY, USA, 1958.

[178] I. T. Jolliffe, *Principal component analysis*. New York City, NY, USA: Springer-Verlag, 1986.

[179] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd ed ed. Boston ; London: Academic Press, 1990.

[180] S. O. Haykin, *Neural Networks: A Comprehensive Foundation*. ON, Canada: Prentice Hall, 1999.

[181] J. Ylipaavalniemi, "Variability of Independent Components in functional Magnetic Resonance Imaging " MSc, Department of Computer Science and Engineering, Helsinki University of Technology, Espoo, Finland, 2005.

[182] E. Oja and J. Karhunen, "Signal Separation by Nonlinear Hebbian Learning," in *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, Perth, Australia, 1995, pp. 83-87.

[183] J. Karhunen and J. Joutsensalo, "Generalizations of principal component analysis, optimization problems, and neural networks," *Neural Networks,* vol. 8, pp. 549-562, 1995.

[184] P. Comon, "Independent Component Analysis, A new concept?," *Signal Processing,* vol. 36, pp. 287-314, 1994.

[185] M.-M. Deza and E. Deza, *Dictionary of Distances*: Elsevier, 2006.

[186] A. Belouchrani and A. Cichocki, "Robust whitening procedure in blind source separation context," *IET Electronics Letters,* vol. 36, pp. 2050-2051, 2000.

[187] G. Strang, *Introduction to linear algebra*, 4th ed. Wellesley, MA, USA: Wellesley-Cambridge Press, 2009.

[188] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks: Theory and Applications*: Wiley, 1996.

[189] E. Oja, *Subspace Methods of Pattern Recognition*. England: Research Studies Press ; Wiley, 1983.

[190] M. Richardson, "Principal Component Analysis," Mathematical Modelling and Scientific Computing, University of Oxford, Oxford, UK, 2009.

[191] T. Lloyd and D. Bau, *Numerical Linear Algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1997.

[192] C. Jutten and J. Herault, "Blind separation of sources, Part 1: an adaptive algorithm based on neuromimetic architecture," *Signal Processing,* vol. 24, pp. 1-10, 1991.

[193] P. S. Laplace, *Théorie analytique des probabilités*. Paris, France, 1812.

[194] R. K. Olsson and L. K. Hansen, "Blind Separation of More Sources than Sensors in Convolutive Mixtures," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 657-660.

[195] L. Te-Won, M. S. Lewicki, M. Girolami, and T. J. A. S. T. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters,* vol. 6, pp. 87-90, 1999.

[196] M. E. Davies and C. J. James, "Source separation using single channel ICA," *Signal Processing,* vol. 87, pp. 1819-1832, 2007.

[197] M. G. Kendall, A. Stuart, and J. K. Ord, *The advanced theory of statistics*, 4th ed. London, UK: Charles Griffin, 1977.

[198] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proceedings of the 4th International Workshop on independant components analysis and blind signal separation*, Aussois, France, 1999, pp. 365-371.

[199] T.-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, "A unifying information-theoretic framework for independent component analysis," *Computers & Mathematics with Applications,* vol. 39, pp. 1-21, 2000.

[200] A. Hyvärinen, P. Hoyer, and M. Inki, "Topographic Independent Component Analysis," *Neural Computation,* vol. 13, pp. 1527-1558, 2001.

[201] D. Russell and T. Rossing, "Testing the Nonlinearity of Piano Hammers Using Residual Shock Spectra," *Acta Acoustica,* vol. 84, 1998.

[202] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing,* vol. 8, pp. 320-327, May 2000.

[203] L. Parra and C. Spence, "On-line convolutive blind source separation of nonstationary sources," *Journal of VLSI Signal Processing,* vol. 26, pp. 39-46, 2000.

[204] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks,* vol. 13, pp. 411-430, 2000.

[205] S. Comani, H. Preissl, D. Mantini, Q. Campbell, G. Alleva, and H. Eswaran, "Comparison of algorithms for fetal signal reconstruction: Projector Operator vs. Independent Component Analysis," *International Congress Series,* vol. 1300, 2007.

[206] A. Hyvärinen, "Gaussian moments for noisy independent component analysis," *IEEE Signal Processing Letters,* vol. 6, pp. 145-147, 1999.

[207] A. Hyvärinen, "Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood," *Neurocomputing,* vol. 22, pp. 49-67, 1998.

[208] C. Jutten, M. Babaie-Zadeh, and S. Hosseini, "Three easy ways for separating nonlinear mixtures?," *Signal Processing,* vol. 84, pp. 217-229, 2004.

[209] R. M. Parry and I. Essa, "Blind Source Separation using Repetitive Structure," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Madrid, Spain, 2005, pp. 143-148.

[210] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "Solution of the Shannon's problem on the monotonicity of entropy," *Journal of the American Mathematical Society,* vol. 17, 2004.

[211] J. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing Stereo Music with Score-Informed Source Separation," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Victoria, BC, Canada, 2006.

[212] A. J. Bell and T. J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation,* vol. 7, pp. 1129-1159, 1995.

[213] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proceedings F Radar and Signal Processing,* vol. 140, pp. 362-370, 1993.

[214] A. Hyvärinen, "A family of fixed-point algorithms for independent component analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Bavaria, Germany, 1997, pp. 3917-3920 vol.5.

[215] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions : with formulas, graphs and mathematical tables*, 9th ed. New York City, NY, USA: Dover Publications, 1973.

[216] P. J. Huber, "Projection Pursuit," *The Annals of Statistics,* vol. 13, pp. 435-475, 1985.

[217] B. B. Mandelbrot and R. L. Hudson, *The misbehaviour of markets : a fractal view of risk, ruin and reward*. London, UK: Profile, 2004.

[218] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. Urbana, IL, USA: University of Illinois Press, 1949.

[219] K. Sobczyk, "Information Dynamics: Premises, Challenges, and Results," *Mechanical Systems and Signal Processing,* vol. 15, pp. 475-498, 2001.

[220] A. Papoulis, *Probability, random variables, and stochastic processes*, 2nd ed. New York City, NY, USA ; London, UK: McGraw-Hill, 1984.

[221] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York City, NY, USA: Chichester : Wiley, 1991.

[222] D. Leibovici and C. Beckmann, "An introduction to Multiway Methods for Multi-Subject fMRI experiment," University of Oxford, Centre for Functional Magnetic Resonance Imaging of the Brain, Oxford, UK, 2001.

[223] M. C. Jones and R. Sibson, "What is Projection Pursuit?," *Journal of the Royal Statistical Society. Series A (General),* vol. 150, pp. 1-37, 1987.

[224] A. Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," *Advances in Neural Information Processing Systems (NIPS),* vol. 10, pp. 273-279, 1998.

[225] J. A. Lee, F. Vrins, and M. Verleysen, "A Simple ICA Algorithm for Non-Differentiable Contrasts," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, 2005, pp. 1-4.

[226] J. A. Snyman, *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*: Springer, 2005.

[227] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks,* vol. 10, 1999.

[228] R. L. Burden and J. D. Faires, "Fixed-Point Iteration," in *Numerical Analysis*, ed: PWS, 1985.

[229] V. Zarzoso and P. Comon, "How fast is FastICA," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Florence, Italy, 2006.

[230] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation,* vol. 9, pp. 1483-1492, 1997.

[231] M. D. Plumbley, S. Abdallah, J. Bello, M. Davies, J. Klingseisen, G. Monti, and M. Sandler, "ICA and related models applied to audio analysis and separation," in *Proceedings of the International Symposium on Soft Computing and Intelligent Systems for Industry (ICSC)*, Paisley, Scotland, UK, 2001.

[232] F. J. Ampel and T. Uzzle, "The History of Audio and Sound Measurement," presented at the 94th Audio Engineering Society (AES) Convention, Berlin, Germany, 1993.

[233] D. Laing. (1991) A voice without a face: popular music and the phonograph in the 1890s. *Popular Music*. 1-9.

[234] M. Chanan, *Repeated Takes: A Short History of Recording and Its Effects on Music*. Brooklyn, NY, USA: Verso Books, 1995.

[235] C. Soanes and A. Stevenson, *Oxford dictionary of English*, 2nd ed. Oxford, UK: Oxford University Press, 2005.

[236] A. G. Bell, "Experiments Relating to Binaural Audition," *American Journal of Otolaryngology,* July 1880.

[237] A. N. Scriabin, "Early Hi-Fi Wide Range and Stereo Recordings Made by Bell Telephone Laboratories in the 1930s," in *Leopold Stokowski Conducting the Philadelphia Orchestra (1931-1932)*, ed: Bell Laboratories, 1981.

[238] S. Millman, *A History of Engineering and Science in the Bell System: Communications Sciences (1925-1980)*. New York City, NY, USA: AT&T Bell Laboratories, 1984.

[239] W. B. Snow, "Basic principles of stereophonic sound," *IRE Transactions on Audio,* vol. AU-3, pp. 42-53, 1955.

[240] B. Owsinski, *The Mixing Engineer's Handbook*. Vallejo, CA, USA: MixBooks, 1999.

[241] B. Katz, *Mastering Audio: The Art and the Science*. Burlington, MA, USA: Focal Press, 2003.

[242] R. A. Berkovitz, "Four Channel Recording and Reproducing system," USA Patent, 1973.

[243] R. H. Snyder, "Sel-Sync and the "Octopus": How Came to be the First Recorder to Minimize Successive Copying in Overdubs," *Journal of the Association for Recorded Sound Collections (ARSC),* vol. 34, pp. 209-213, 2003.

[244] D. Barry and B. Lawlor, "Sound Source Separation: Azimuth Discrimination and Resynthesis," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Naples, Italy, 2004.

[245] B. Fox. (1981) Hundred years of stereo: fifty of hi-fi. *New Scientist*. 908-911.

[246] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*. Oxford, UK: Elsevier Ltd., 2007.

[247] J. M. Eargle, "Stereo/Mono Disc Compatibility: A Survey of the Problems," *Journal of the Audio Engineering Society (JAES),* vol. 17, p. 5, June 1969.

[248] G. Martin and J. Hornsby, *All you need is ears: The inside personal story of the genius who created The Beatle*. New York City, NY, USA: St. Martin's Griffin, 1994.

[249] Beatles, "Beatles '65," ed: Capitol, 1965.

[250] R. M. Aarts, E. Larsen, and D. Shobben, "Improving Perceived Bass and Reconstruction of High Frequencies for Band Limited Signals," presented at the IEEE Benelux Workshop on Model based Processing and Coding of Audio, Leuven, Belgium, 2002.

[251] E. Vickers, "The Loudness War: Background, Speculation, and Recommendations," presented at the 129th Audio Engineering Society (AES) Convention, San Francisco, CA, USA, 2010.

[252] C. Avendano and J.-M. Jot, "Frequency Domain Techniques for Stereo to Multichannel upmix," in *International Conference on Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland, 2002.

[253] M. Cobos and J. J. López, "Singing Voice Separation Combining Panning Information and Pitch Tracking," presented at the 124th Audio Engineering Society (AES) Convention, Amsterdam, The Netherlands, 2008.

[254] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main Instrument Separation from Stereophonic Audio Signals Using a Source/Filter Model," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, 2009.

[255] D. Barry, B. Lawlor, and E. Coyle, "Real-time Sound Source Separation: Azimuth Discrimination and Resynthesis," presented at the 117th Audio Engineering Society (AES) Convention, San Francisco, CA, USA, 2004.

[256] J. P. Forsyth, "Source Separation, Removal, and Resynthesis Using Azimuth-based Source Separation," Masters, Music and Performing Arts Professions, New York University, New York City, NY, USA, 2008.

[257] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *The Journal of the Acoustical Society of America,* vol. 94, pp. 98-110, 1993.

[258] R. Cooney, N. Cahill, and B. Lawlor, "An Enhanced implementation of the ADRess (Azimuth Discrimination and Resynthesis) Music Source Separation Algorithm," presented at the 121st Audio Engineering Society (AES) Convention, San Francisco, CA, USA, 2006.

[259] D. FitzGerald, D. Barry, M. Cranitch, and E. Coyle, "Automatic detection of optimal azimuth widths for sound source separation using ADRess," in *Proceedings of IET Irish Signals and Systems Conference (ISSC)*, Galway, Ireland, 2008.

[260] D. Barry, B. Lawlor, and E. Coyle, "Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm," presented at the 119th Audio Engineering Society (AES) Convention, New York City, NY, USA, 2005.

[261] A. De Goetzen, N. Bernandini, and D. Arfib, "Traditional Implementations of a Phase-Vocoder: the Tricks of the Trade," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Verona, Italy, 2000.

[262] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing,* vol. 81, pp. 2353-2362, 2001.

[263] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Separating Underdetermined Convolutive Speech Mixtures " in *Independent Component Analysis and Blind Signal Separation*. vol. 3889, ed Berlin/Heidelberg, Germany: Springer, 2006.

[264] D. S. Moore, *The basic practice of statistics*, 3rd ed. New York City, NY, USA: Freeman, 2004.

[265] Y. Gong and W. Xu, *Machine Learning for Multimedia Content Analysis*. New York City, NY, USA: Springer, 2007.

[266] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of Subjective and Objective Evaluation Methods for Audio Source Separation," *Proceedings of Meetings on Acoustics,* vol. 4, 2008.

[267] M. Cobos, J. J. Lopez, A. Gonzalez, and J. Escolano, "Stereo to Wave-Field Synthesis music up-mixing: An objective and subjective evaluation," in *Proceedings of the International Symposium on Communications, Control, and Signal Processing (ISCCSP)*, Malta, 2008, pp. 1279-1284.

[268] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A Community-Based Approach to Large-Scale Evaluation," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brazil, 2009, pp. 734-741.

[269] Bearlin, I. Calvo, and J. Rabascall, "Roads," in *New Old*, ed: BadQProductions, 2009.

[270] Tamy, "Que Pena / Tanto Faz," in *Soul Mais Bossa*, ed: Curvemusic, 2007.

[271] (2012, May). *Creative Commons Attribution-NonCommercial-ShareAlike 2.5 Generic*. Available: http://creativecommons.org/licenses/by-nc-sa/2.5/

[272] (2012, June). *Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Generic*. Available: http://creativecommons.org/licenses/by-nc-sa/3.0/

[273] "Nine Inch Nails, Only," in *With Teeth*, ed: Island, Interscope, 2005.

[274] "Los de Abajo, Resistencia," in *LDA v The Lunatics*, ed: RealWorld Records, 2005.

[275] "Suicide Sports Club, Don't Know (ft. Duke & Suspect)," in *Electric Mistress*, ed: B_Rock, 2005.

[276] B. Eno and D. Byrne, "Help Me Somebody," in *My Life in the Bush of Ghosts*, ed: Sire, 1981.

[277] P. Gabriel, "Shock the Monkey," in *Single*, ed: Geffen, 1982.

[278] S. Nazarkhan, "Kunlarim Sensiz," in *Sen*, ed: Emd International, 2007.

[279] Radiohead, "Nude," in *In Rainbows*, ed: XL Recordings, 2008.

[280] C. Renato and N. Salerno, "Tu Vuò Fà L'Americano (Arranged by ΦeelM)," ed, 2007.

[281] J. Sundberg, "A perceptual function of the singing formant," KTH Computer Science and Communication, 1972.

[282] E. Vincent, S. Araki, and P. Bofill. (24/08/2009). *Signal Evaluation Campaign: Professionally produced music recordings*. Available: http://www.irisa.fr/metiss/SiSEC08/SiSEC_professional/

[283] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, pp. 91-98, 2006.

[284] C. A. Champlin, "Hearing: An Introduction to Psychological and Physiological Acoustics (3rd edition)," *Ear and Hearing,* vol. 20, p. 439, 1999.

[285] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, DC, USA, 1979, pp. 208-211.

[286] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed.: Wiley-IEEE Press, 1999.

[287] T. Eerola and P. Toivianen, "MIR in Matlab: The MIDI Toolbox," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[288] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[289] G. Grove, S. Sadie, and J. Tyrrell, *The new Grove dictionary of music and musicians*, 2nd ed. Oxford: Grove, 2001.

[290] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE,* vol. 66, pp. 51-83, 1978.

[291] *TS 36.104 V9.7.0 Base Station (BS) radio transmission and reception (Release 9),* 3GPP Standard, 2011.

[292] Z. Fang, Z. Guoliang, and S. Zhanjiang, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology,* vol. 16, pp. 582-589, 2001.

[293] P. Gampp, "Evaluation of Robust Features for Singing Voice Detection," M.Sc., Institute of Electronic Music and Acoustics, University of Music and Dramatic Arts Graz, Graz, 2010.

[294] N. Tin Lay, S. Arun, and W. Ye, "Singing voice detection in popular music," in *Proceedings of the 12th ACM International Conference on Multimedia Retrieval (ICMR)*, New York, NY, USA, 2004.

[295] C. Wu and G. Liang, "Robust singing detection in speech/music discriminator design," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001, pp. 865-868 vol.2.

[296] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 1564-1578, July 2007.

[297] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology (EuroSpeech)*, Rhodes, Greece, 1997, pp. 1895-1898.

[298] NIST. (2011, 05/2011). *Evaluation Tools (DETware v2.1.tar.gz)*. Available: http://www.itl.nist.gov/iad/mig/tools/

[299] K. K. Wojcicki and K. K. Paliwal, "Importance of the Dynamic Range of an Analysis Windowfunction for Phase-Only and Magnitude-Only Reconstruction of Speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, 2007, pp. 729-732.

[300] D. Griffin and L. Jae, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Language Processing,* vol. 32, pp. 236-243, 1984.

[301] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing Melody Extraction in Polyphonic Music by Harmonic Tracking," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[302] K. Aczel and I. Vajk, "Note-based sound source separation of polyphonic recordings," *Infocommunications Journal,* 2009.

[303] S. Rieck, "Singing Voice Extraction from 2-Channel Polyphonic Musical Recordings," Diploma, Institute of Electronic Music and Acoustics, Graz University of Technology, Graz, Austria, 2012.

[304] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 34, pp. 744-754, 1986.

[305] D. Stirzaker, *Elementary probability*. Cambridge: Cambridge University Press, 1994.

[306] J. Karhunen, E. Oja, D. Wang, R. Vigario, and J. Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Transactions on Neural Networks,* vol. 8, pp. 486-504, May 1997.

[307] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed. Baltimore, MD, USA ; London, UK: Johns Hopkins University Press, 1996.

[308] J. W. Tukey, "An introduction to the calculations of numerical spectrum analysis," in *Spectral Analysis of Time Series*, Harris, Ed., ed New York City, NY, USA: Wiley, 1967, pp. 25-46.