UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTMENT OF COMPUTER SCIENCE

PH.D. IN COMPUTER SCIENCE

SIGNAL TRANSFORMATIONS FOR IMPROVING INFORMATION REPRESENTATION, FEATURE EXTRACTION AND SOURCE SEPARATION

Tutor: Prof. Goffredo Haus
Coordinator: Prof. Paolo Boldi

Ph.D. Thesis by:
Presti Giorgio
Matr. Nr. R10490

XXIX Ph.D. cycle
A.A. 2015-2016

Musica est exercitium arithmeticae occultum
nescientis se numerare animi

Contents

Chapter 1

Introduction

1.1 Abstract

This thesis is about new methods of signal representation in time-frequency domain, so that required information is rendered as explicit dimensions in a new space. In particular two transformations are presented: Bivariate Mixture Space and Spectro-Temporal Structure-Field. The former transform aims at highlighting latent components of a bivariate signal based on the behaviour of each frequency base (e.g. for source separation purposes), whereas the latter aims at folding neighbourhood information of each point of a $\mathbb{R}^2$ function into a vector, so as to describe some topological properties of the function. In the audio signal processing domain, the Bivariate Mixture Space can be interpreted as a way to investigate the stereophonic space for source separation and Music Information Retrieval tasks, whereas the Spectro-Temporal Structure-Field can be used to inspect spectro-temporal dimension (segregate pitched vs. percussive sounds or track pitch modulations). These transformations are investigated and tested against state-of-the-art techniques in fields such as source separation, information retrieval and data visualization.

In the field of sound and music computing, these techniques aim at improving the frequency domain representation of signals such that the exploration of the spectrum can be achieved also in alternative spaces like the *stereophonic panorama* or a virtual *percussive* vs. *pitched* dimension.

1.2 Keywords

Sound and Music Computing; Signal Processing; Image Processing; Bivariate Signals; Stereophonic Audio; Upmix; Linear Pattern Recognition; Pitched/Glissando/Percussive Sounds; Fourier Transform; Radon Transform; Hough Transform.

1.3 Motivation

From the author's perspective, music can be the perfect playground for acquiring knowledge about the world, since by definition it arouses the observers, and encourages them to dive more into the exploration of that complex intangible system, related to almost all fields of science. In fact, as Carl Stumpf pointed out between 1883 and 1890 [1], music cannot be explained as a sum of simple systems: its intrinsic nonlinearity forces researchers to find complex models to investigate signals which in other disciplines are generally handled from a more elementary perspective.

This point of view led to major scientific leaps many times in history: for example both Ehrenfels and Wertheimer (considered as the fathers of Gestalt psychology) started their revolutionary work from the observation of musical phenomena [2, 3]. For what concerns the goals of this work, it can be said that music complexity can be an interesting benchmark to test many signal processing techniques.

Furthermore, music is something that we can process instinctively, so *music* intended as a human ability can also suggest new models of information processing. Of course the *Human Information Processing* paradigm is now considered as obsolete in the context of (music) psychology [4], nevertheless it suggested that there are processes going on inside human mind that can be borrowed by those scientists interested in signal processing. However, it is this writer's opinion that the only reasonable path that the white rabbit of psychology can suggest is the one that leads to neuroscience, so it is by looking at the biology of perception that new ways to inspect signals can be found.

This does not necessarily mean that the processing techniques of our brain should be copied, instead it would be more interesting to learn which information is highlighted at each step, no matter how it is computed. This is important because it helps in building models that can be meaningful in every intermediate state, in opposition to models that copy our mechanics but whose intermediate states are difficult to interpret, such as deep neural-networks.

Finally, goals of this research can be put in a more explicit form. This work is attempting to satisfy the need of a more meaningful spectral representation (i.e. *Fourier Transform* and *Short Term Fourier Transform*) by answering two questions:

$Q1$: *How can the Fourier Transform be improved for considering the relationship that underlies bivariate signals?*

$Q2$: *How can the Short-Term Fourier Transform be improved for considering the relationship between neighbour frames?*

These questions will be answered in the context of Sound and Music Computing, with expectations of serendipity, especially in fields like harmonic analysis, multivariate analysis and image processing. For example, they can lead to the definition of

new acoustic features together with alternative implementations of well-known source separation strategies such as [57, 61, 52]. In particular ***Q*1** is related to the analysis of the stereophonic space, while ***Q*2** enables the ability to explore new *percussive* vs. *pitched* spaces.

At the time of writing, the only peer-reviewed publication related to this thesis (especially to ***Q*1**) is [62], which addresses the basic mathematical machinery on which the Bivariate Mixture Space is built, together with applications in data visualization.

1.4 Structure

In Chapter 2 a brief overview of the basic concepts of different disciplines are recalled within the proposed perspective, then in Chapter 3 a quick review of some Sound and Music Computing techniques is given to put this work in some context. In Chapter 4 a technique for representing bivariate signals is exposed, namely the *Bivariate Mixture Space*, as an answer to ***Q*1**. In Chapter 5 *Spectro-Temporal Structure Field* is exposed as an extensive answer to ***Q*2**. Finally, in Chapter 6, both techniques are put together and tested in the context of source separation applications. Conclusions and future works are discussed in Chapter 7. To help reading this thesis, a table of the acronyms can be found in Appendix A, while Appendix B contains raw data from the tests presented in Chapter 6.

Finally, a Matlab implementation of the proposed methods can be found as a GitHub repository at the following URL: `https://github.com/Kuig/LIM-Toolbox`

Chapter 2

Background

2.1 Signal Processing

José Moura, former president of the IEEE Signal Processing Society, defined signal processing as

> "[...] an enabling technology that encompasses the fundamental theory, applications, algorithms, and implementations of processing or transferring information contained in many different physical, symbolic, or abstract formats broadly designated as signals." [5]

Clearly, many disciplines are involved with this definition, and it would be very hard to draw a general overview of the subject within this work, thus just the basic concepts relative to fundamentals of the current context will be discussed. In particular the milestone work of Jean Baptiste Joseph Fourier [6] and the interesting work of Johann Radon [7] will be examined.

2.1.1 Fourier Series and Transform

Through the Fourier series (FS) and its extension as an integral transformation, a signal can be decomposed into a set of orthogonal trigonometric bases. This is realized by taking the scalar product of the bases by the input: the only non-zero-sum outcomes are those where the base matches a component of the signal, in which case the result is the power of that base contained in the input. More precisely, FS can be defined as follows.

Let $[-T/2, T/2]$ be an interval of length T of the signal $x(t)$ to be decomposed, and consider the definite integral:

$$\int_{-T/2}^{T/2} |x(t)|\, dt \tag{2.1.1}$$

FS exists if and only if Eq. 2.1.1 is finite and the portion of $x(t)$ within the interval contains a finite number of type-1 discontinuities[1] and a finite number of maxima and minima. If these conditions are satisfied, the signal can be defined as the FS:

$$x(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(2\pi kt/T) + b_k \sin(2\pi kt/T) \tag{2.1.2}$$

where $\frac{k}{T}$ is the frequency f of base k, and the "=" sign means that inside the interval FS converges to $x(t)$, while outside the interval it converges to a periodic repetition of the $x(t)$ portion within the interval.

The main reason why Fourier developed this technique is to simplify the study of heat transfer, but his technique lately gave birth to the field of *harmonic analysis* (from the Greek word "harmonikos", meaning "skilled in music"), a branch of mathematics concerned with the representation of functions or signals as the superposition of basic waves. In practice, thanks to FS, information is taken from the function itself to the series coefficients, that can be calculated as follow:

$$a_0 = \frac{1}{T} \int_{-T/2}^{T/2} x(t)\, dt \tag{2.1.3a}$$

$$a_k = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos(2\pi kt/T)\, dt \tag{2.1.3b}$$

$$b_k = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin(2\pi kt/T)\, dt \tag{2.1.3c}$$

It is common to see Eq. 2.1.2 written in a more meaningful form, which uses only cosines as bases, and highlights the magnitude M_k and phase Φ_k spectra:

$$x(t) = a_0 + \sum_{k=1}^{\infty} M_k \cos(2\pi kt/T - \Phi_k) \tag{2.1.4}$$

[1] A discontinuity of type 1 occurs in x_0 when $\lim_{x \to x_0^+} \neq \lim_{x \to x_0^-}$

Moreover, considering Euler's form of sine and cosine, Eq. 2.1.2 can also be written as in Eq. 2.1.5. This is called the *bilateral form* of FS since k range is now extended also to $-\infty$:

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{i2\pi kt/T} \tag{2.1.5}$$

where negative frequencies ($k < 0$) can be interpreted as a different geometric representation of Eq. 2.1.4, (in real signals $c_{-k} = c_k$). In this form, complex coefficients c_k can be calculated as:

$$c_k = \frac{a_k - jb_k}{2} = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-j2\pi kt/T}\, dt \tag{2.1.6}$$

The relationship among the coefficients of all forms can be made explicit:

$$M_k = |c_k| = \frac{\sqrt{(a_k)^2 + (b_k)^2}}{2} \tag{2.1.7a}$$

$$\Phi_k = -\angle c_k = \arctan(-\frac{b_k}{a_k}) \tag{2.1.7b}$$

Given that in Eq. 2.1.5 and 2.1.6 $\frac{1}{T}$ defines bases frequency quantization Δf, letting $T \to \infty$ takes the discrete series to a new, more general, continuous form, where Δf becomes df; $k\Delta f$ becomes f; and the summation of Eq. 2.1.5 becomes an integral (parameter c_k is replaced with its definition):

$$x(t) = \int_{-\infty}^{\infty} e^{j2\pi ft}\, df \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft}\, dt \tag{2.1.8}$$

In particular, the second integral of Eq. 2.1.8 is called Fourier transform $\mathcal{F}$ and takes a signal from time domain t to frequency domain f (a general form of Eq. 2.1.6):

$$\mathcal{F}\{x(t)\} = X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft}\, dt \tag{2.1.9}$$

While the first integral of Eq. 2.1.8 can be considered as the reverse operation (a general form of Eq. 2.1.5):

$$\mathcal{F}^{-1}\{X(f)\} = x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft}\, df \tag{2.1.10}$$

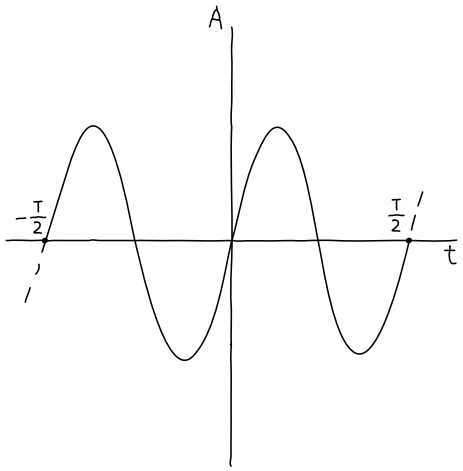


Figure 2.1: A signal in the interval $\pm\frac{T}{2}$ with a periodicity of $\frac{2}{T}$

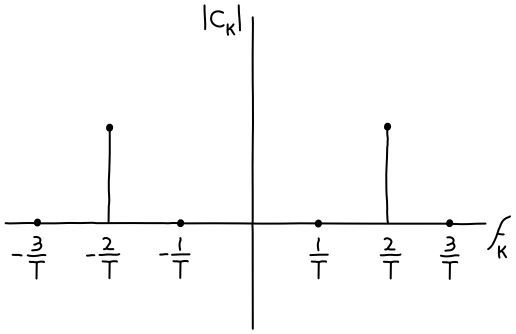


Figure 2.2: FS of the signal in Fig. 2.1, now defined in the space of its periodicities

It is clear that the only difference between $\mathcal{F}$ and its inversion is the sign of the exponent, which in case of real signals is negligible since it can be seen as a reflection of the frequency axis.

The transformation of signal in Fig. 2.1 is shown in Fig. 2.2.

To see the evolution of spectrum over time, a signal can be split into shorter segments of equal length transformed separately. This process is called short-term Fourier transform ($STFT$) and can be described as an extension of $\mathcal{F}$, where the signal is masked with a windowing function $w(t)$ before taking $\mathcal{F}$:

$$X(f,\tau) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j2\pi ft}\,dt \tag{2.1.11}$$

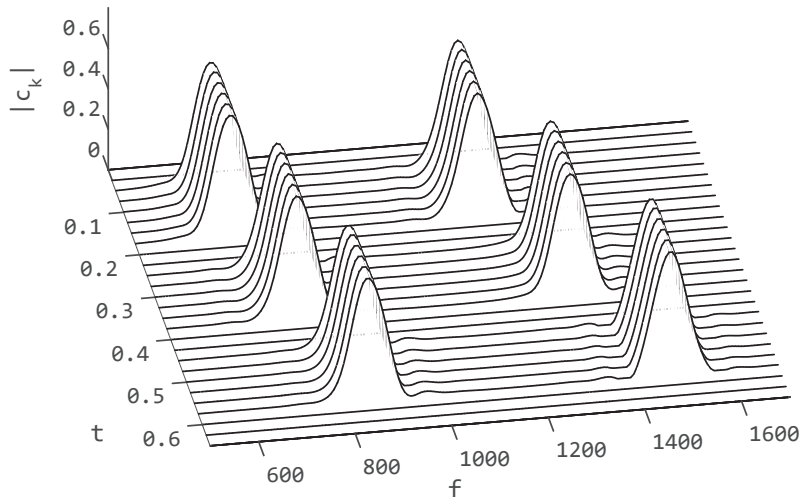


Figure 2.3: Short term Fourier transform of a triplet of DTMF tones, each one composed of two harmonics, lasting about 0.125 seconds.

where the new parameter τ represent the time offset of window w. In particular w should be non-zero only inside a finite interval around the origin (usually truncated Gaussian or Hann windows are used). As it happens in FS with $[-T/2, T/2]$ interval, smaller windows provide higher time resolution (w is well localized in time) but lower frequency resolution (less bases fit into the segment). An example of $STFT$ is shown in Fig. 2.3, where the finite interval visibly lowers the frequency resolution. What should be a single line is spread into a more smooth and wide curve because - by a property of $\mathcal{F}$ - the multiplication of the window w in the time domain, corresponds to a convolution in the spectral domain.

As a final remark, let's say that $\mathcal{F}$ can also be performed on n-dimensional signals, such as images or volumes, not to be confused with the multivariate condition implied by $\boldsymbol{Q1}$: existing methods for decomposing two spectra such as [8] are not tailored on the exact desiderata of our context, even if they somehow share the same mathematical principles.

2.1.2 Radon Transform

The Radon transform is the integral transform which takes a function defined on the plane to a function defined on the (two-dimensional) space of lines in the plane, whose value at a particular point is equal to the line integral of the function over that line. The Radon transform is widely applicable to tomography, the creation of

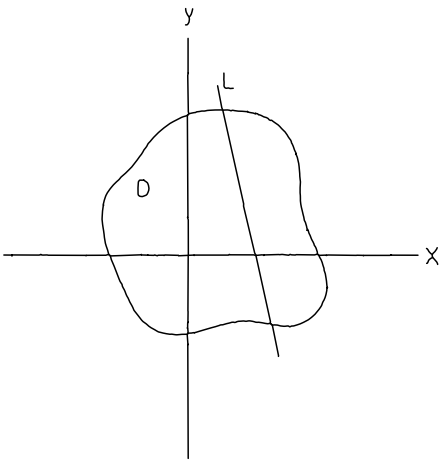


Figure 2.4: Line L through a domain D (adapted from [9])

an image from the projection data associated with cross-sectional scans of an object. The following definitions are based on the work of [9, 10, 11]. [2]

Let (x, y) be coordinates of points in the plane, and consider some function f defined on a domain D of $\mathbb{R}^2$. If L is any line in the plane (see Fig. 2.4), then the mapping defined by the projection (i.e. line integral) of f along all possible lines L is the Radon transform of f:

$$\mathcal{R}f = \check{f} = \int_L f(x, y)\, ds \tag{2.1.12}$$

where ds is an increment of length along L. To define the transformation more precisely, it will be useful to set up some coordinates and be more rigorous about the integration along all lines L. Consider Fig. 2.5, where the equation of L is given in the normal form by:

$$p = x \cos\phi + y \sin\phi \tag{2.1.13}$$

The line integral in Eq. 2.1.12 depends on p and ϕ. This can be indicated explicitly by writing:

$$\check{f}(p, \phi) = \int_L f(x, y)\, ds \tag{2.1.14}$$

If $\check{f}(p, \phi)$ is known for all p and ϕ, then $\check{f}(p, \phi)$ is the two-dimensional Radon transform of $f(x, y)$.

[2]The Radon transform is closely related to the Hough transform [12], and they are equivalent under certain circumstances [13]. In this context only the Radon transform will be examined, since it has a well-founded mathematical basis and, for the goals of this thesis, is more intuitive as well.

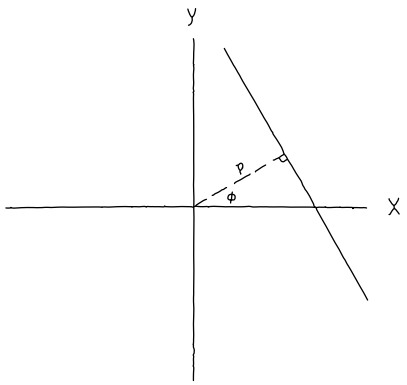


Figure 2.5: Coordinates to describe line in Fig. 2.4 (adapted from [9])

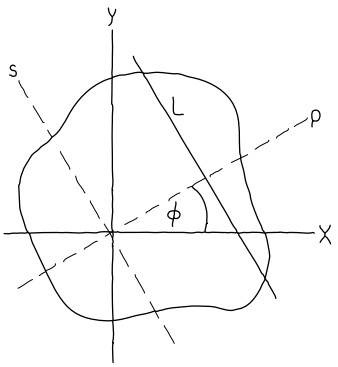


Figure 2.6: The line in fig. 2.4 relative to original and rotated coordinates (adapted from [9])

Now suppose a new coordinate system is introduced with axes rotated by the angle ϕ. If the axes are labelled by p and s as in Fig. 2.6, then:

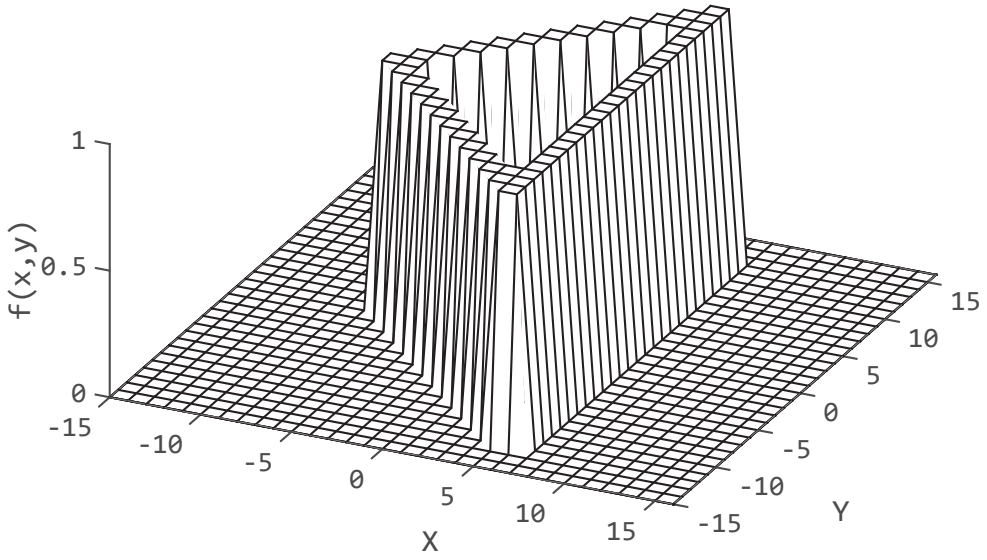
$$x = p\cos\phi - s\sin\phi \tag{2.1.15a}$$

$$y = p\sin\phi + s\cos\phi \tag{2.1.15b}$$

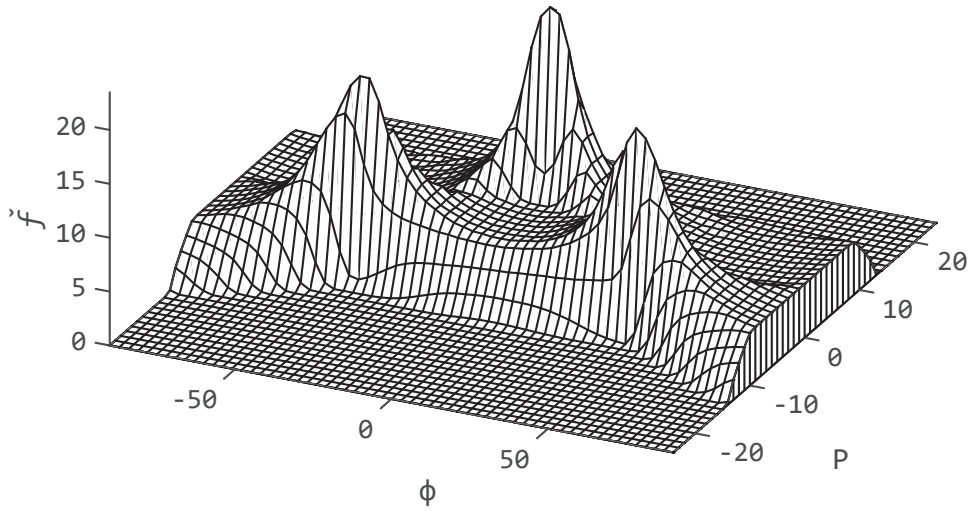
a more explicit form of the transform can be written as (of course, the limits of Eq. 2.1.16 can be finite if f vanishes outside D):

$$\check{f}(p,\phi) = \int_{-\infty}^{\infty} f(p\cos\phi - s\sin\phi,\ p\sin\phi + s\cos\phi)\,ds \tag{2.1.16}$$

In Fig. 2.7 an example of the Radon transform of an image is shown.



(a) A function in $\mathbb{R}^2$



(b) Radon transform

Figure 2.7: A function with its Radon transform. The three segments of the triangle in Fig. 2.7a becomes peaks in Fig. 2.7b, whose coordinates represents the parameters of the original segments.

The Radon transform may be inverted in many ways, the most used one exploits the connection with the n-dimensional Fourier transform $\mathcal{F}_n$. Let f be a function in $\mathbb{R}^n$, then [11, 14] demonstrated that the following relation with the Radon transform holds:

$$\mathcal{F}_n f = \mathcal{F}_1 \check{f} \tag{2.1.17}$$

where $\mathcal{F}_1$ is the one-dimensional Fourier transform along the radial direction of the Radon transform. By this definition it is easy to see the inversion as:

$$f = \mathcal{F}_n^{-1} \mathcal{F}_1 \check{f} \tag{2.1.18}$$

Nevertheless, in this context it is interesting to see another method, which just approximates a *blurred version* of f. It is called *backprojection* and works as follow.

Let $\xi = (\cos\phi, \sin\phi)$ be a unit vector. Consider an arbitrary function $\psi(t, \xi)$ where $t = \xi \cdot \mathbf{x} = x\cos\phi + y\sin\phi$. The backprojection operator $\mathcal{B}$ is defined by [15]:

$$\mathcal{B}\psi = \int_0^\pi \psi(x\cos\phi + y\sin\phi,\, \xi) d\phi \tag{2.1.19}$$

Since ξ is completely determined by specifying ϕ, and since $\mathcal{B}\psi$ is a function of (x, y), it may be useful to write:

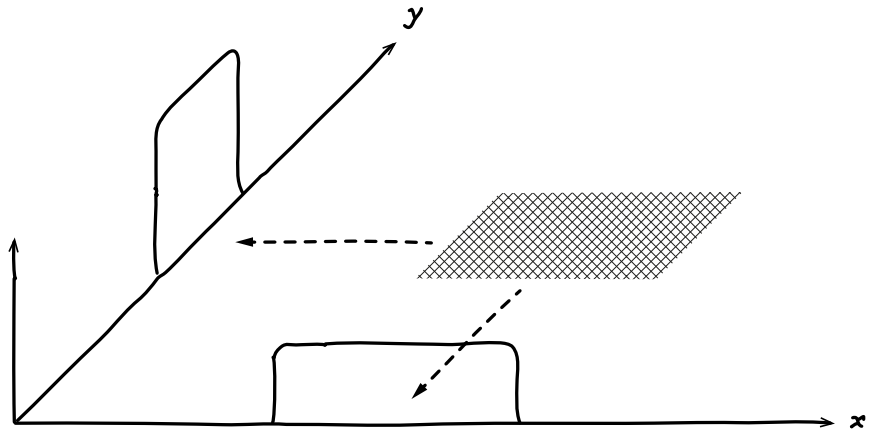
$$[\mathcal{B}\psi](x, y) = \int_0^\pi \psi(x\cos\phi + y\sin\phi,\, \xi) d\phi \tag{2.1.20}$$

Or, in terms of polar coordinates (r, θ), where $x = r\cos\theta$ and $y = r\sin\theta$:

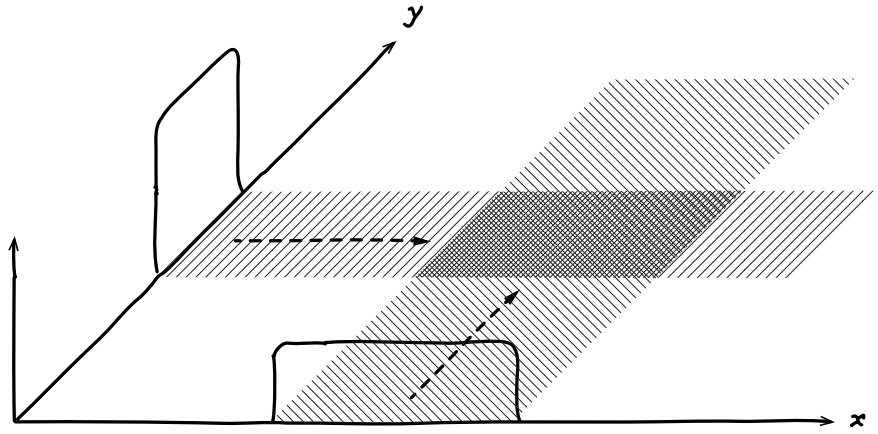
$$[\mathcal{B}\psi](r, \theta) = \int_0^\pi \psi[r\cos(\theta - \phi),\, \phi] d\phi \tag{2.1.21}$$

The backprojection operator for only two projections is illustrated in Fig. 2.8 and the operation for fixed ϕ is illustrated in Fig. 2.9. Observe that if $\psi(p, \phi)$ is identified with the projection function $\check{f}(p, \phi) = \mathcal{R}f(x, y)$, then for fixed angle ϕ, the incremental contribution to $\mathcal{B}\psi$ at the point (x, y) is just the value of $\check{f}(p, \phi)$ multiplied by $d\phi$ when p is computed from $p = x\cos\phi + y\sin\phi$. Of course, that value may be found by integrating f along the line that passes through (x, y) and is at a distance $p = x\cos\phi + y\sin\phi$ from the origin (this will be exploited in Chapter 5). If this ϕ-backprojection is known for all angles ϕ, then the complete backprojection is obtained by integration over ϕ as indicated by Eq. 2.1.20.

The fact that a single point in the Radon space tells something about the geometrical relation of a set of points of the input function, is an interesting hint of how ***Q2*** can be answered. Precisely, backprojection will be exploited in Chapter 5 as a way to back-propagate Radon information onto the input function.



(a) Two profiles of a rectangular object (i.e. the Radon transform for two values of ϕ).



(b) Backprojection of profiles and superposition to form an approximation to original object

Figure 2.8: The backprojection operation (adapted from [9])

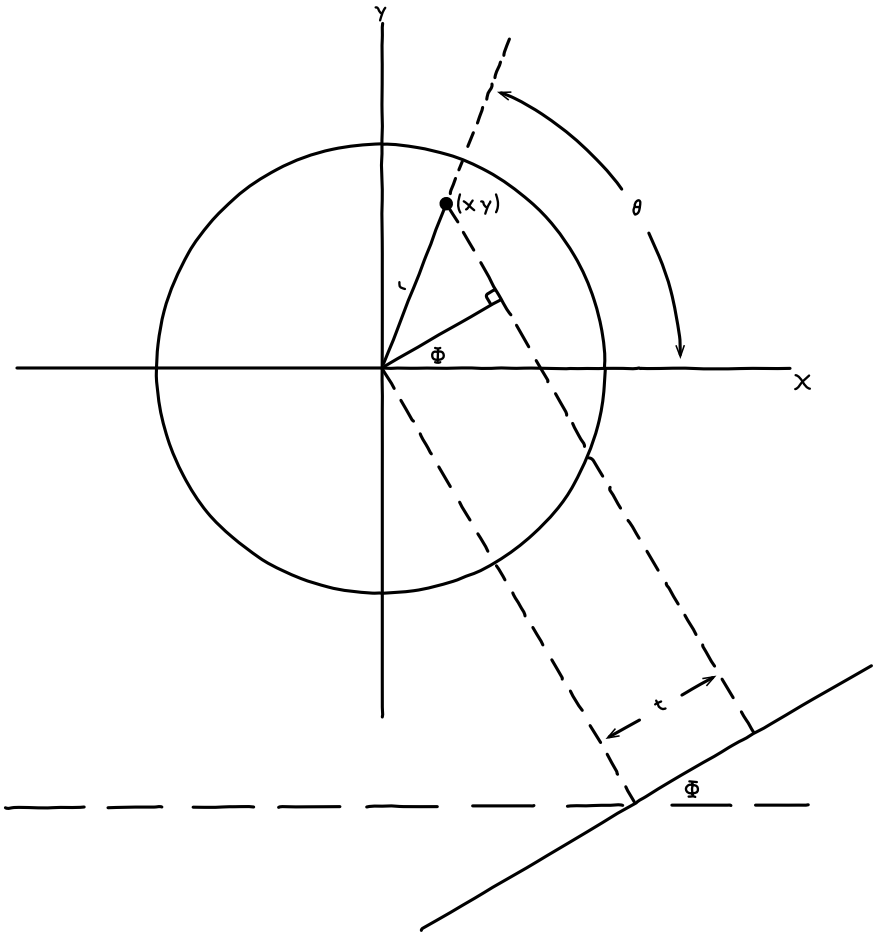


Figure 2.9: Geometry for obtaining the ϕ-backprojection. For a fixed angle Φ, the incremental contribution $d(\mathcal{B}\psi)$ to $\mathcal{B}\psi$ at the point (x, y) or, equivalently, (r, θ) is given by $\psi(t, \Phi)d\phi$. The full contribution to $\mathcal{B}\psi$ at (x, y) is found by integrating over ϕ as indicated in Eq. 2.1.20. Note that $t = x\cos\Phi + y\sin\Phi = rcos(\theta - \Phi)$. (adapted from [9])

2.2 Multivariate Signal Processing

Multivariate signal processing is a set of techniques aimed at handling or modelling multivariate signals. This branch of signal processing endows some of the knowledge coming from multivariate statistics, a term which include all statistics where there are more than two variables analysed simultaneously. Multivariate signal processing is frequently used for data decomposition and reduction; clustering and classification; investigation of signals dependencies; signal modelling and prediction.

Before describing very briefly some of the techniques used in sound and music computing, a proper definition of a multivariate signal is called for.

A multivariate signal $\mathbf{x}(n)$ of dimension M consists of M signals $x_i(n)$

$$\mathbf{x}(n) = \{x_1(n);\, x_2(n);\, \ldots\, ;\, x_M(n) \,|\, 0 \leq n < N\} \tag{2.2.1}$$

where N is the signal length, and each $x_i(n)$ is called a *mixture* of S latent sources $s_j(n)$ weighted by a scalar value a_{ij} such as

$$x_i(n) = a_{i1}s_1(n) + a_{i2}s_2(n) + \cdots + a_{iS}s_S(n) \tag{2.2.2}$$

Introducing vector $\mathbf{s}(n)$ to collect all sources $s_j(n)$ and vector $\mathbf{A}$ to collect all weights a_{ij}, a more compact definition can be given

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) \tag{2.2.3}$$

Finally, When $M = 2$, $\mathbf{x}(n)$ is called a *bivariate* signal.

Among the goals of multivariate signal processing, the most relevant in the context of this work, are those relative to the possible approximations of sources $\mathbf{s}(n)$ and weights $\mathbf{A}$ given certain constrains. In some cases, this is referred as a *source separation* problem.

2.2.1 Principal Component Analysis

Defined between 1901 and 1933 by Karl Pearson [16] and Harold Hotelling [17], sometimes referred as the most simple linear decomposition method, the aim of principal component analysis (PCA) is to rearrange $\mathbf{x}(n)$ as M linearly uncorrelated signals called principal components (PC). A graphical example is shown in Fig. 2.10.

PCs are calculated as orthogonal transformations of the original space (a linear combination of input variables which transformation vectors are determined by data itself), defined in such a way that the first PC has the largest possible variance (i.e. accounts for as much of the variability in the data as possible), and each succeeding

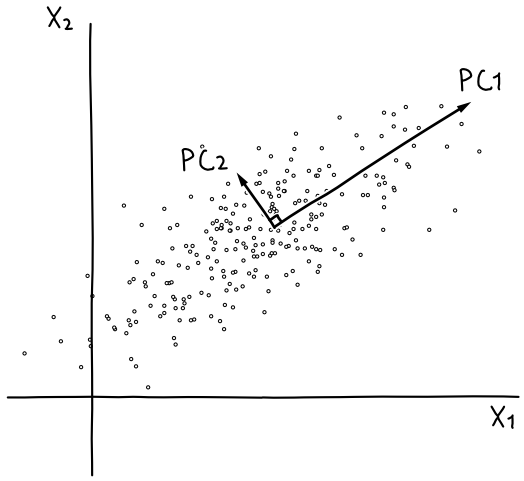


Figure 2.10: Signal samples according to original space $x_i(n)$ and to new space defined by orthogonal Principal Components

PC has the highest variance under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set (that is, with a very low absolute Pearson correlation).

Discarding PCs with the lowest variance allows the reconstruction of a simplified approximation of the input, this is generally exploited in lossy data compression and exploratory data analysis (i.e. dimensionality reduction).

2.2.2 Independent Component Analysis

The general framework that allowed the development of independent components analysis (ICA) was prepared by Jeanny Herault and Christian Jutten in 1986 [18], but it was only in 1994 that Pierre Comon better defined this technique [19].

The objective of ICA is to find a decomposition of a multivariate signal $\mathbf{w}(n)$ of size M in terms of M independent non-Gaussian source signals

$$\mathbf{w}(n) = \mathbf{A}\mathbf{s}(n) \tag{2.2.4}$$

In order to decide if two signals are truly independent it is not sufficient to consider their correlation, because (with the exception of Gaussian distributed variables) two signals which are uncorrelated may still be dependent. To achieve separation of independent source components, a stronger measure of independence is needed. The two broadest definitions of independence used in ICA are thus *Minimization of mutual information* and *Maximization of non-Gaussianity*. The former family of algorithms

uses measures like the Kullback-Leibler Divergence and maximum entropy, while the latter uses kurtosis and negentropy.

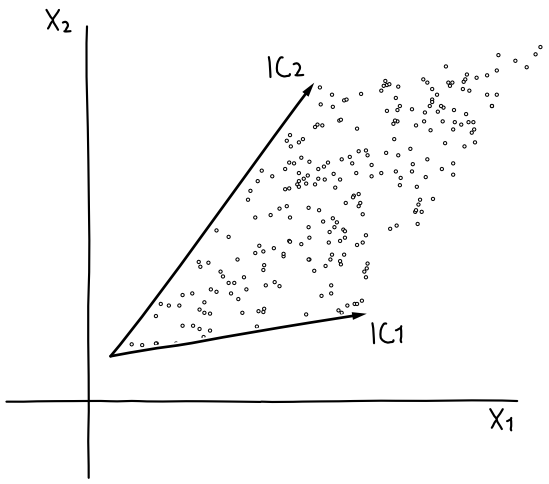


Figure 2.11: Non-Gaussian signal samples according to original space $x_i(n)$ and to new space defined by independent components

Even if the found $\mathbf{s}(n)$ are actually the original sources mixed in $\mathbf{w}(n)$, ICA is not able to guess the correct scale and polarity of sources. Nevertheless ICA guarantees that found sources are non-Gaussian and independent. As can be seen in Fig. 2.11, if PCA can be informally described as a roto-translation of original space, ICA is composed by a richer set of linear transformations.

2.2.3 Non-Negative Matrix Factorisation

Known by chemists as *self modeling curve resolution* since 1971 [20], proper non-negative matrix factorizations techniques was used by Paatero *et al.* in 1994-1995 under the name *positive matrix factorization* [21, 22] and became more widely known as non-negative matrix factorization (NMF) after Lee and Seung investigated the properties of the algorithm and published some simple and useful algorithms in 1999-2001 [23, 24].

NMF is a set of techniques based on linear algebra where a matrix $\mathbf{V}$ is factorized into two matrices $\mathbf{W}$ and $\mathbf{H}$. All of the matrices have no negative elements: in applications such as processing of $STFT$-based spectrograms this property is inherent to the data being considered. Since the problem admits no general exact solutions, it is commonly approximated numerically with different approaches.

Matrix $\mathbf{W}$ is called a *features* matrix, while $\mathbf{H}$ is called a *coefficients* matrix. As Fig. 2.12 shows, if target matrix $\mathbf{V}$ is a spectrogram, features can be seen as spectral

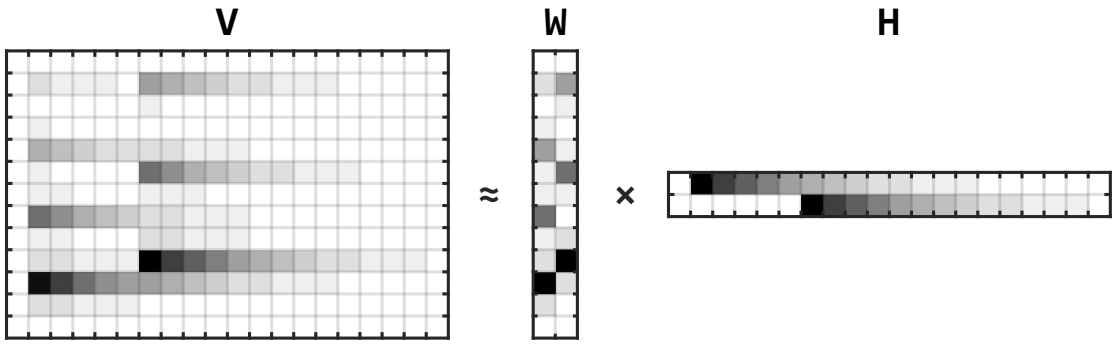


Figure 2.12: Matrix $\mathbf{V}$ can be factorized as the product of $\mathbf{W} \times \mathbf{H}$

templates (or bases), while coefficients are temporal activation patterns.

Different types of NMF arise from using different cost functions for measuring the divergence between $\mathbf{V}$ and $\mathbf{WH}$. Each cost function leads to a different algorithm, usually minimizing the divergence by using iterative update rules.

2.3 Sound and Music Computing

Sound and Music Computing (SMC) is a research field that studies the whole sound and music communication chain from a multidisciplinary point of view. It aims at understanding, modelling and generating sound and music through computational approaches by combining scientific, technological and artistic methodologies.

This definition of SMC somehow depends on the one of signal processing given in Section 2.1; what actually distinguishes SMC from other signal processing sub-fields is the family of signals examined: as explained in Chapter 1 the music phenomena has many peculiarities.

Even if harmonic analysis is born with the intent of simplifying the description of functions through their decomposition into frequency spectra, music signals retain their complexity even in the frequency domain. Moreover, in the context of SMC, scientists are frequently interested in perceptual properties of sound. This led to the proliferation of many acoustic features, i.e. many ways to summarize the information present into the signal.

Acoustic features may refer to low-level signal properties as well as high-level ones, but are always extracted from a signal represented in *time- frequency- cepstral- lag-* and, in some cases, *Radon*-space (for an exhaustive review on this topic see [25]). An enhancement of the frequency domain representation through the introduction of two new spaces should provide new tools for the feature extraction task.

To define these spaces and answer $\boldsymbol{Q1}$ and $\boldsymbol{Q2}$, it is worth spending some time

discussing a number of basic SMC concepts, in particular, those related to how multichannel audio signals are handled and how some timbral properties reflect onto the spectrum.

2.3.1 Stereophonic Representation

In a music production and distribution context, low level digital audio signals are generally represented as Pulse Code Modulation (PCM) streams, where stereophonic information is encoded as a couple of channels, that may represent left (L) and right (R) speaker signals. Alternatively, it is also common to represent the same information as the sum M and difference S of the channels, that is just a 45 degrees rotation of the original space:

$$M = \frac{L + R}{\sqrt{2}} \tag{2.3.1a}$$

$$S = \frac{L - R}{\sqrt{2}} \tag{2.3.1b}$$

This method is called mid-side (MS) and offers the ability to process what is perceived *in front* of the listener separately from what can be perceived *laterally*, in opposition to LR technique where the distinction is made for signals that come from the *left* or from the *right* direction. Fig. 2.13 gives a qualitative example of how M, S, L and R channels can be placed in the azimuthal listening plane.

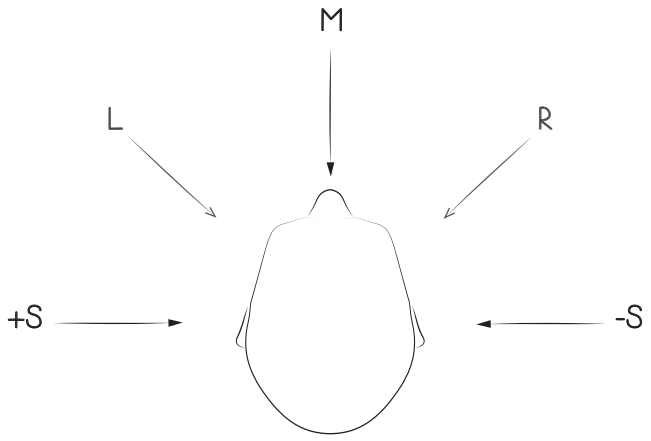


Figure 2.13: Exemplification of MS v.s. LR encoding in the stereo image

The most common way to distribute a sound between L and R is to exploit intensity level differences (see Section 2.4.1): sound technicians use the *panoramic potentiometer* (*pan*, for short) to feed different amounts of a monophonic signal to

distinct speakers, giving the listener the illusion of a sound where azimuthal position is located somewhere between the speakers (the *ghost source* effect) [26, 27]. Eq. 2.3.2 shows an implementation of pan

$$\begin{aligned} L &= a_L \cdot source \\ R &= a_R \cdot source \end{aligned} \tag{2.3.2}$$

Here $a_L^2 + a_R^2 = 1$, and the relationship between $a_{L,R}$ and the perceived position of the ghost source is explained by the *stereophonic law of sines* (for further readings see [28]).

At this point the couple of signals composing the stereophonic mixture can be thought as a bivariate signal, since many other sources may have been processed in the same way and mixed together. Eventually, to increase realism, convolutive phenomena are added (e.g. reverberation and delays between L and R). This introduces a non-zero *interaural time difference* (and thus *interaural phase difference*) for each frequency component, increasing the overall complexity of the mixture.

The listener perceives the final mixture as a dense panorama of sounds coming from all directions between the two speakers. This virtual panorama is called the *stereophonic image* or *stereophonic field*.

Finally, the stereo image becomes a comprehensive example of a complex mixture of signals if it is considered that non-linearities are usually added at the end of mixing phase in a process called *mastering*.

Some strategies can be found in the literature to explore these kind of mixtures, such as those exposed in Sections 3.1 and 3.2. In Chapter 4 a novel generalization of these techniques will be provided while answering ***Q*1**.

2.3.2 Pitched vs. Percussive Sounds

Pitch is a perceptual property of sound that allows the ordering on a frequency-related scale [29], or more commonly, pitch is the quality that makes it possible to judge sounds as "higher" or "lower" in the sense associated with musical melodies [30]. It is implicit in this definition that not all pitched sounds must have a harmonic structure (i.e. this makes the distinction between definite and indefinite pitch), nor must they be perfectly periodic. The most important thing seems to be that the energy must be concentrated around a finite number of peaks and this behaviour must be clear and stable enough to be distinguishable from noise [31]. By these properties, a steady pitch seen in the time-frequency domain of the $STFT$ (where time is displayed horizontally and frequency vertically) should appear as a concentration of energy in horizontal structures.

On the other hand, percussive sounds are defined as short bursts of sound, usually but not necessarily followed by a pitched component. The very beginning phase of

this class of sounds is generally referred as transient: an abrupt change in sound pressure made of a broad band of frequency components (the shorter it is, the more it resembles a Dirac impulse), which can be represented in the time-frequency domain as energy being organised in vertical structures.

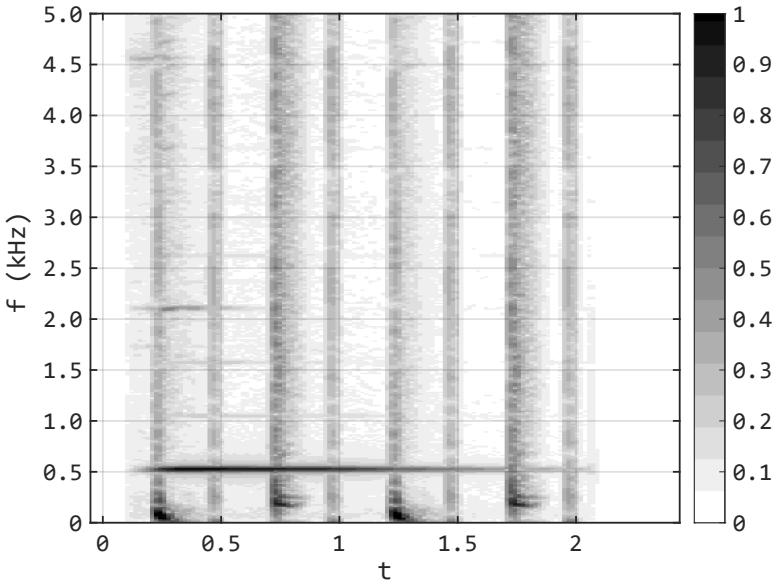


Figure 2.14: STFT of drum samples mixed with a vibraphone tone (harmonic sounds can also be characterized by frequency modulations, as shown in the example of Fig. 3.6a)

The horizontal v.s. vertical arrangement of energy of these two sound classes is clearly visible in Fig. 2.14, where a drum track is mixed with a vibraphone tone. Also patterns with different angles and shapes are possible if pitched sounds are performing a glissando or a vibrato, nevertheless, after a certain modulation speed, they will be perceived respectively as percussions [3] or new steady sounds (as in FM synthesis).

This kind of information is very interesting in the context of SMC, as suggested by the works cited in Section 3.3, and answering to ***Q2*** should enrich the $STFT$ enough to tell, for each point of the $STFT$, which class it belongs to. This is just the more immediate use of the transformation proposed in Chapter 5, which in principle can be used for many other goals, such as an alternative framework for works such as those described in Section 3.4 and 3.5.

[3] Think of an electronic kick-drum sample, usually realized by a very fast descending tone

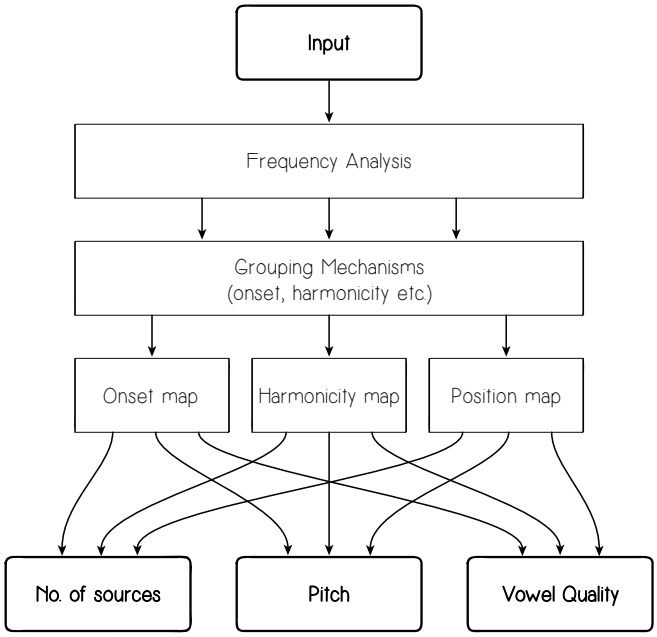


Figure 2.15: The Auditory Scene Analysis framework

2.3.3 Computational Auditory Scene Analysis

Computational Auditory Scene Analysis (CASA) [32, 33] is a *machine listening* framework that aims at the separation of sound mixtures in the same way humans do.

It starts from the assumption that perfect separation is not possible, thus the proper goal is defined as *the organization of available information into useful structures*. This is realized thanks to the *Auditory Scene Analysis* proposed by Bregman in the 1990's [34, 35]. The process starts by breaking mixture into small elements, those are then grouped into sources using perceptually motivated heuristics, then sources are enriched with aggregate attributes. This architecture is shown in Fig. 2.15.

Frequency analysis can be performed simulating the cochlea with Gammatone filterbanks [36, 37] (sometime referred as *cochleagrams*), while grouping rules are inherited by Gestalt psychology such as: common onset and modulation; harmonicity; spatial location [38].

Even if not directly related to this work, the paradigm of CASA is worth mentioning for being both a good example of biological-inspired computing and a possible recipient of the proposed techniques.

2.3.4 Source Separation Evaluation Tools

Most of the audio source separation techniques rely on technology such as ICA, NMF and CASA, as well as other case-specific techniques. When evaluating the results of those algorithms it is important to find objective measures that may be used to compare separation results. Particularly relevant in SMC are those based on the work of Vincent *et al.* [39], that is the decomposition of the recovered sources into a set of components, respectively the target, the artefacts and the interference signals. General purpose tools like BSS Eval toolbox [40] work on a matrix decomposition of the signal, while audio-specific tools like PEASS toolkit [41] uses a perceptual approach, thus providing results more similar to the subjective evaluation of users. Since the latter tool is going to be used in this work, a brief description of the PEASS output is called for.

PEASS returns set of 4 scores: overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS), and artefact-related perceptual score (APS), all in the range $0 - 100$. These scores are computed by calculating the energy ratio between the input signal (coming from a source separation algorithm) and its decomposition in different components such as the proper target signal, the artefacts, and the interference signal coming from other sources. This decomposition is realized thanks to a comparison with original source signals, performed in a cochleagram-like space. To achieve perceptual relevance, a weighting of the most salient components followed by a non-linear mapping of the scores is performed.

Ideally OPS should provide an overall idea of the goodness of the separation strategy, TPS should tell how much of the source has been recovered, IPS tells how much of other sources ended in the extracted one, and APS should measure the *artefacts* introduced by the separation algorithm (described by the authors as the *musical noise*). Higher scores should correspond to better source separation techniques.

Unfortunately, even if these metrics may be useful to quickly compare different strategies, they received some criticism among the SMC community, since in some cases the results obtained with PEASS correlate extremely poorly with subjective evaluation [42]. Nevertheless, for the purpose of this work, they still provide a good starting point for the comparison with other methods.

Another interesting aspect that is worth to mention regards the datasets that may be found to validate source separation techniques. Many datasets have been collected in the last years, mainly grouped by the kind of task they are thought to validate. For example in the Signal Separation Evaluation Campaign (SiSEC) community, the following tasks are addressed:

- Underdetermined speech and music mixtures
- Professionally produced music recordings

- Mixtures of speech and real-world background noise
- Asynchronous recordings of speech mixtures
- Biomedical signals

For the reasons expressed in Section 1.3, professionally produced music recordings will be examined in this context. For this task SiSEC relays on the DSD100 dataset [43], a database of 100 professionally produced songs provided with original tracks.

Unfortunately the evaluation proposed in Chapter 6 requires that, for every song, each parameter of the tested algorithms must be tuned manually. For this reason a smaller dataset will be used, namely the MASS dataset [44]. This dataset is composed of 12 excerpts of 6 songs, 4 of which with alternative mixes, for a total of 16 test cases.

Unlike the DSD100 dataset, MASS comes with no State-of-the-Art scores, but this does not represent a big issue, since the goal of the tests is not to evaluate the overall source separation capabilities of the proposed approaches, but just to compare them to similar techniques.

2.4 Neurophysiology of Perception

As it has been said in Chapter 1, a desideratum of this work is to represent information in the most meaningful way while answering to $\boldsymbol{Q1}$ and $\boldsymbol{Q2}$. To achieve this goal it is useful to look at the neurophysiology of perception to learn which *data types* our brain (even if not consciously) is used to.

Of course a whole book can be filled with information about this topic (actually *many* books), so just the most relevant part of the brain in respect to this context will be examined, that are those parts responsible for the analysis of $\mathbb{R}^2$ functions (the *primary visual cortex* and the *inferior colliculus*) and those dedicated to bivariate signal processing (the *superior olivary complex*).

2.4.1 Neural Pathway of Hearing

The cochlea is an astonishing organ, which behaves like an *acoustic prism* and (together with the cochlear nuclei) is capable of translating mechanical acoustic information into a complex data stream comparable to a frequency domain representation of sound. Frequency information is represented in many nuclei as tonotopically organized groups of neurons, that is, tones close to each other in terms of frequency are represented in topologically neighbouring regions in the brain.

The neural pathway of hearing (shown in Fig. 2.16) that brings information from the cochlea to the auditory cortex begins with the two cochlear nuclei performing monaural non-linear spectral processing, then information from left and right ears is

combined in the superior olivary complex (SOC). From this point of the neural path, all sites receive almost the same information from the two ears. The inferior colliculus (ICC) is involved in spectro-temporal processes and in the integration and routing of multi-modal sensory perception, while the medial geniculate body relays auditory information to the cortex and influences the direction and maintenance of attention [45]. In this context, some of the SOC and ICC roles are considered.

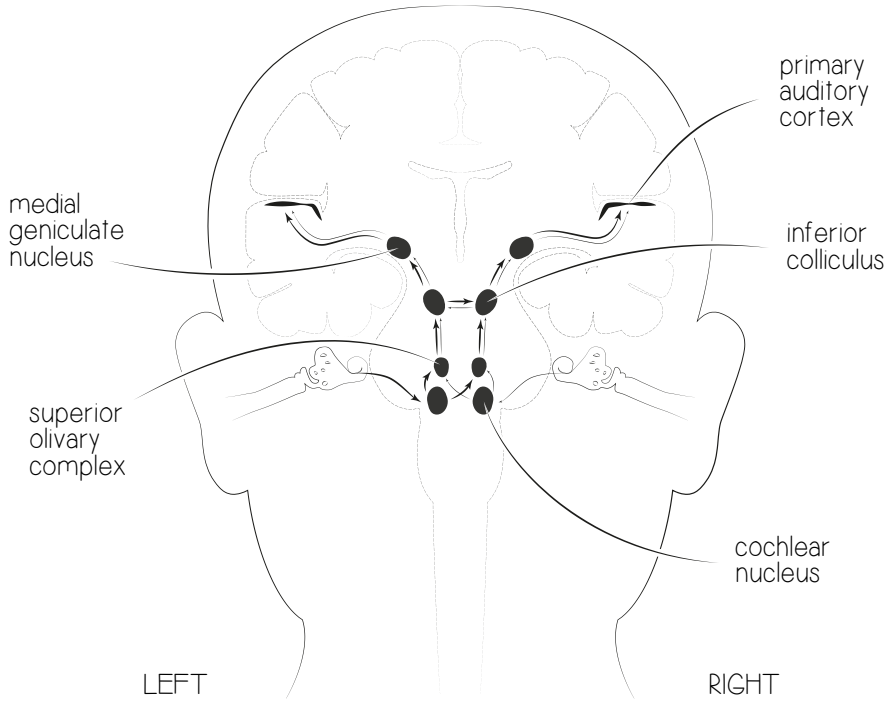


Figure 2.16: Hearing neural pathway (simplified)

SOC: The medial superior olive (MSO) is a specialized nucleus of SOC that is believed to measure the time difference of arrival of sounds between the ears (the inter-aural time difference or ITD and inter-aural phase difference or IPD). The ITD is a major cue for determining the azimuth of sounds. The lateral superior olive (LSO) is believed to be involved in measuring the difference in sound intensity between the ears (the inter-aural level difference or ILD). The ILD is a second major cue in determining the azimuth of high-frequency sounds.

This function of the SOC suggests that acoustic information is not merely represented in the brain as independent streams from the two ears, instead a compacted spectral information representation with satellite information seems to be projected to ICC and other nuclei [46].

Since this kind of frequency domain data-integration seems to answer ***Q1***, it has been taken into account in the approach proposed in Chapter 4; at first applied to audio signals, then extended to any bivariate signal. Information is taken into a new space where magnitude and phase differences (corresponding to ILD and IPD) are made explicit along with frequency domain data.

ICC: According to [47, 48, 49, 50], the ICC reacts to spectro-temporal patterns in a way that, together with the primary auditory cortex (A1), it realizes auditory receptive fields sensible to frequency variations over time, like glissando and vibrato.

These receptive fields are realized thanks to a virtual representation of spectrum over time, as recently modelled by [51, 52]. These models are similar to those of the visual cortex described in Section 2.4.2, where the $STFT$ is processed as an image (e.g. Fig. 2.14). In other words, some neurons are tuned to fire when certain pitch movement are detected, rather than some others fires when pitch is steady.

Since the visual cortex is a better studied area, it will be taken as a model rather than ICC.

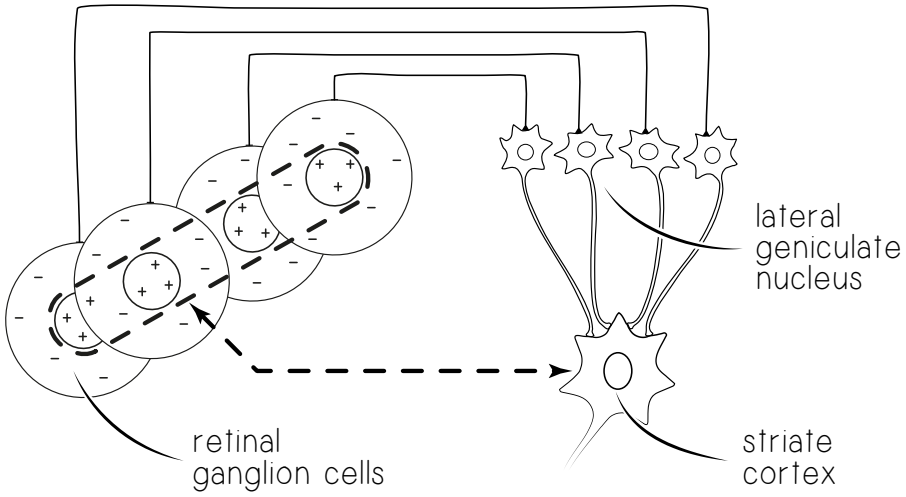
2.4.2 Primary Visual Cortex

The primary visual cortex (V1) is the best-studied visual area in the brain. It is the simplest, earliest cortical visual area. Among other tasks, it is excellent in pattern recognition. Individual V1 neurons have strong tuning to a small set of stimuli: the neuronal responses can discriminate small changes in visual orientations, spatial frequencies and colours. They are organized in cortical columns, each one firing when a certain feature is matched such as lines with particular orientation or frequency. Pioneering works on this topic were made by Hubel and Weisel across the 20th century, in particular in [53, 54] they shed some light on how different neurons of V1 can be sensible to different patterns, thanks to the integration of information over simple- and complex-cells.

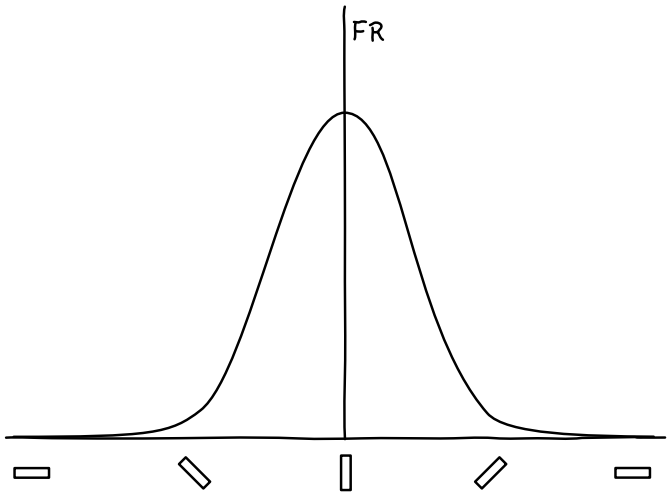
Fig. 2.17a shows the organization of simple receptive fields: A large number of lateral geniculate cells (of which four are illustrated) have receptive fields with *on* centres[4] arranged along a straight line on the retina. All of these project upon a single cortical cell (which synapses are supposed to be excitatory). The receptive field of the cortical cell will then have an elongated *on* centre as indicated by the dashed lines in the receptive-field diagram to the left of the figure.

All of these neurons are arranged in V1 as cortical columns, each one sensible to a certain angle of a certain receptive field. For example, in Fig. 2.17b it is possible to observe the firing activity of a column of the striate cortex sensible to vertical lines.

[4]Retinal receptive fields with *on* centre are composed by ganglion cells that fire when the stimulus is composed by a light in the centre of the receptor surrounded by a darker area



(a) Neuron of the striate cortex sensible to a certain pattern thanks to the connection with aligned ganglion cells



(b) Neurons fire rate as function of angle of stimulus

Figure 2.17: Receptive Fields organisation and activity

Thanks to these structures visual stimuli are paired with structural informations about the content of the perceived image.

Even if these features are usually simulated with Gabor filters [55, 56], in this work a different approach is proposed, which is more useful in answering ***Q2***. The approach described in Chapter 5 is based on the similarities with the Radon transform explained in Section 2.1.2: the array of receptive fields aligned at a certain angle ϕ, merged into a single neuron, can be considered as the line integral over L shown in eq. 2.1.14.

Finally, the similarities between auditory and visual receptive fields suggests that this kind of techniques may be useful for both sound and image processing.

Chapter 3

State-of-the-Art Techniques

Since in next chapters some more punctual comparison with existing techniques will be done, a brief discussion over cited works is called for.

3.1 A Bivariate Signal Decomposition Model

In 2009, Lilly and Olhede [8] defined a way to model a non-stationary oscillatory bivariate signal as an ellipse evolving in time.

In their work the model has been tested on the signal coming from the path of an oceanographic floater. As shown in Fig. 3.1, the path has been decomposed into a time series of ellipses plus a residual. This basically corresponds to the separation of the high frequency periodic component from a low frequency random walk.

Then, three features are extracted from the ellipses (represented in Fig. 3.2): *Root-mean-square amplitude*, *eccentricity* and *orientation*.

This model works very well under the assumption that the path can be approximated by a frequency and amplitude modulated periodic signal, which is not convenient in a general case (let alone complex audio mixtures), nevertheless, it is worth mentioning since the assumption holds when considering single frequency bases, in which case its rationale can be a possible interpretation of the method proposed in Chapter 4, where Bivariate Spectrum metrics will be introduced. Those metrics will be able to grasp the same information as the *Root-mean-square amplitude*, *eccentricity* and *orientation* metrics of this context: by answering ***Q1***, frequency, amplitude and phase modulations of the x and y components of the ellipses can be studied in a single framework.

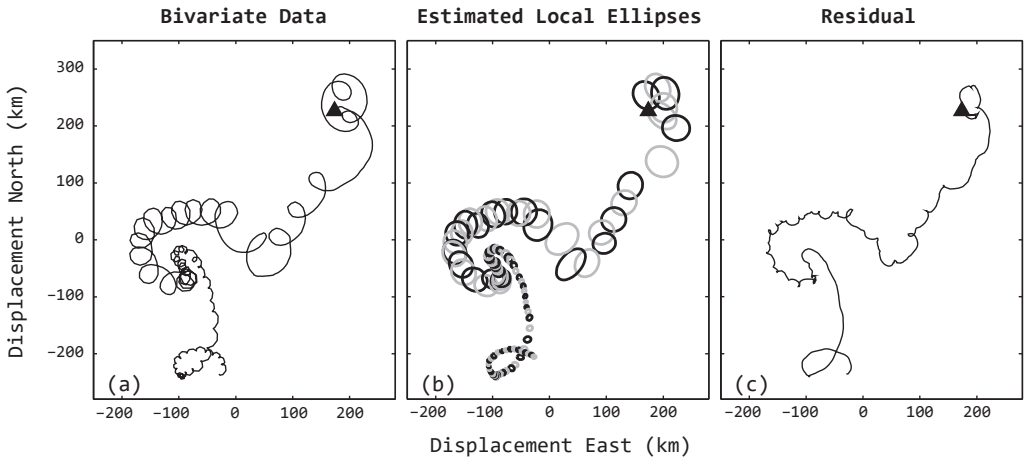


Figure 3.1: The trajectory of a freely drifting oceanographic float. The original data (a) is decomposed into a modulated bivariate oscillation (b) plus a residual (c). (adapted from [8])

3.2 Stereophonic Field Analysis

In 2004, Barry *et al.* developed a sound source separation algorithm called *Azimuth Discrimination and Resynthesis* (ADRess) [57], that performs the task of separation based purely on azimuth discrimination within the stereo field.

They assumed the use of pan as a means to achieve image localisation within stereophonic recordings. As such, only ILD (more precisely *inter-channel* level difference, or ICLD) should exists between left and right channels for a single source (see Eq. 2.3.2).

By using gain scaling and phase cancellation techniques they expose frequency dependent nulls across the azimuth domain, then source separation and resynthesis is performed by taking inverse Fourier transformation for only those frequencies which azimuth position is within a certain range H. Frequency magnitude is retrieved by taking its peak in the azimuth domain, while phase is kept as that in the input signal.

The tolerance H has been introduced because overlapping frequency components in the input mixtures introduce errors in the calculation of their azimuthal position.

Two key aspects are relevant in this context. First, ADRess does not take into account explicitly the phase of frequency bins. Second, they define an *azimuth domain* which is worth some attention.

Taking a single time window of the whole signal as an example, let $L(f)$ and $R(f)$ be the discrete Fourier transform of left and right channel respectively. Then

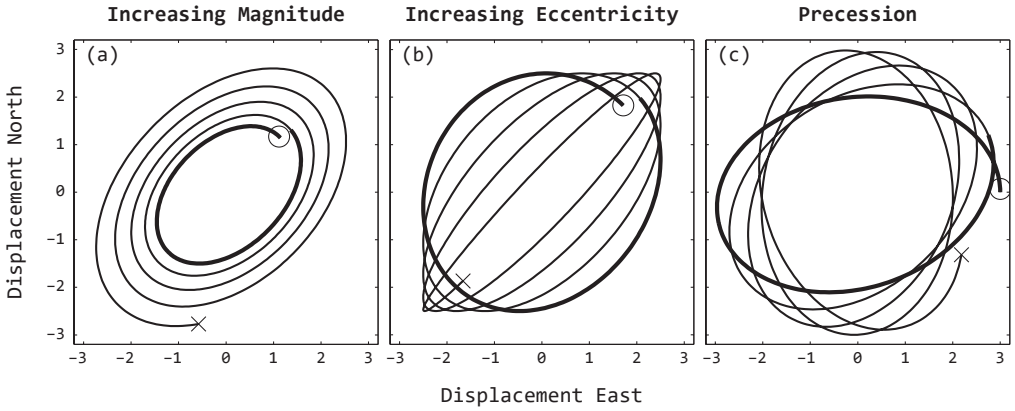


Figure 3.2: An ellipse with uniformly increasing relative amplitude (a), uniformly increasing relative eccentricity (b), and uniformly precessing (c). The bold portion of the line in all three panels shows an initial single orbit. (adapted from [8])

frequency-azimuth planes are defined as

$$\begin{aligned} Az_R(f,g) &= |L(f) - g \cdot R(f)| \\ Az_L(f,g) &= |R(f) - g \cdot L(f)| \end{aligned} \tag{3.2.1}$$

where $0 \leq g \leq 1$ is the azimuthal dimension. An example of how the mixture behaves in this space is shown in Fig. 3.3.

It must be said that the term *azimuth* is used loosely, authors are not dealing with angles of incidence, instead, the azimuth they speak of is purely a function of the intensity ratio created by the pan tool. Nevertheless, the most important concept introduced by [57] is that of an *underlying stereophonic space* that exists between the mixtures. This concept will be generalized considering phase in Chapter 4.

In 2006, Briand *et al.* [58] proposed a parametric representation of multichannel audio based on PCA. They unified many models under the binaural cue coding method (i.e. the use of ICLD and ICPD to infer panning and diffusion of each base of the Fourier transform).

In their model, each frequency bin of the Fourier transform of the signals $L(f)$ and $R(f)$ can be described with a magnitude value $M(f)$ related to PCA, panned with angle $\sigma(f)$ and a measure of the phase difference between the two channels $\Delta\phi(f)$.

$$M(f) = \sqrt{|L(f)^2| + |R(f)^2|} \tag{3.2.2a}$$

$$\sigma(f) = \arctan\left(\frac{|R(f)|}{|L(f)|}\right) \tag{3.2.2b}$$

$$\Delta\phi(f) = \angle L(f) - \angle R(f) \tag{3.2.2c}$$

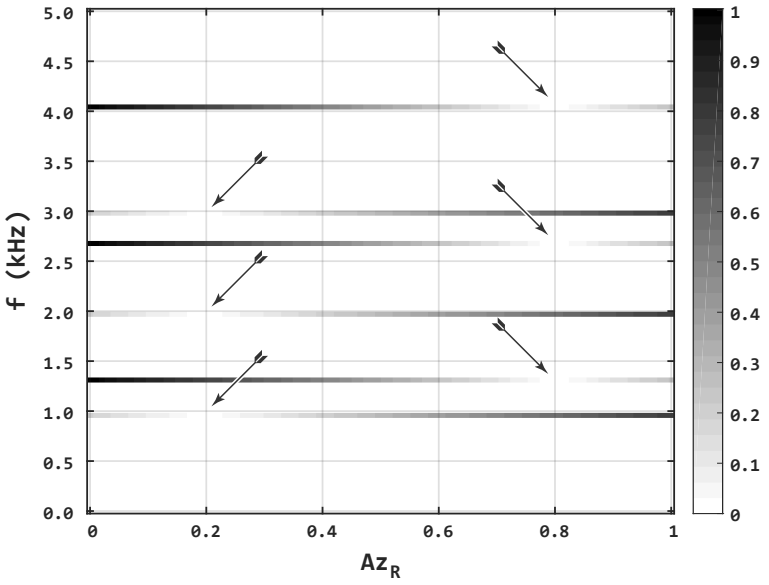


Figure 3.3: The Frequency-Azimuth spectrogram for the right channel. Two synthetic sources are visible, each comprising of 3 non-overlapping partials. The arrows indicate frequency dependent nulls caused by phase cancellation.

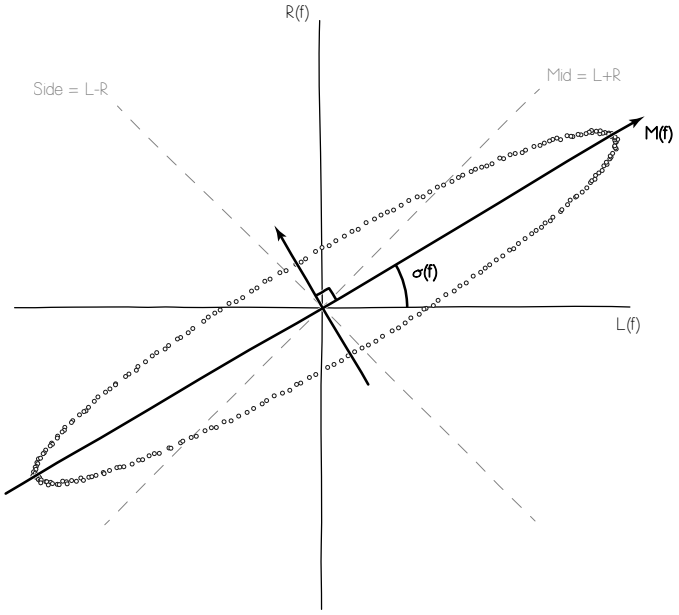


Figure 3.4: The PCA is equivalent to a rotation of the stereo signal coordinate system, and results in one principal component signal (*direct*) and a remaining orthogonal *ambient* signal

An example of these concepts translated in the time domain can be seen in Fig. 3.4.

From the aforesaid work, many others derived interesting strategies to handle stereophonic data: For example in 2007 Goodwin and Jot [59] discussed a spatial analysis-synthesis scheme which applies PCA to an $STFT$ representation of the original audio to separate it into primary and ambient components.

Here, each $STFT$ sub-band is treated as a vector in time and each channel vector is modelled as a sum of a primary component and an orthogonal ambience component.

In 2009, Vickers [60] proposes a similar technique, but with other geometrical interpretations, while working on an upmixing problem. He also considers anti-phase signals as legitimate direct sources from the back direction by using a geometric mean approach.

Finally, in 2015, Kraft and Zöler [61] drops some constrains and with an even more simple calculation based on Mid-Side decomposition in the frequency domain, can separate ambient sound from direct sound, always for upmixing purposes.

The underlying assumptions for all of these works are the following:

1. As in many pan algorithms, panning conserves signal power
2. Pan only works on ICLD, thus there is no phase distortion in the distribution of the source in L and R
3. Both time and frequency resolutions are high enough to minimize component overlapping of multiple sources
4. Left and right ambient signals are similar in magnitude, but different in phase
5. Ambience power is much less than direct signal power.

Interestingly it can be already seen how the measures of Eq. 3.2.2 can qualitatively describe also the features depicted in Fig. 3.2. In Chapter 4 these models will be merged with the idea of *stereophonic space* as a comprehensive answer to ***Q*1**.

Finally, it is worth citing the writer's work of 2014 [62] which provided a primary introduction to the present thesis, but failed in formalizing it properly. In that work just an information visualization strategy based on the very same principle proposed in Chapter 4 was proposed.

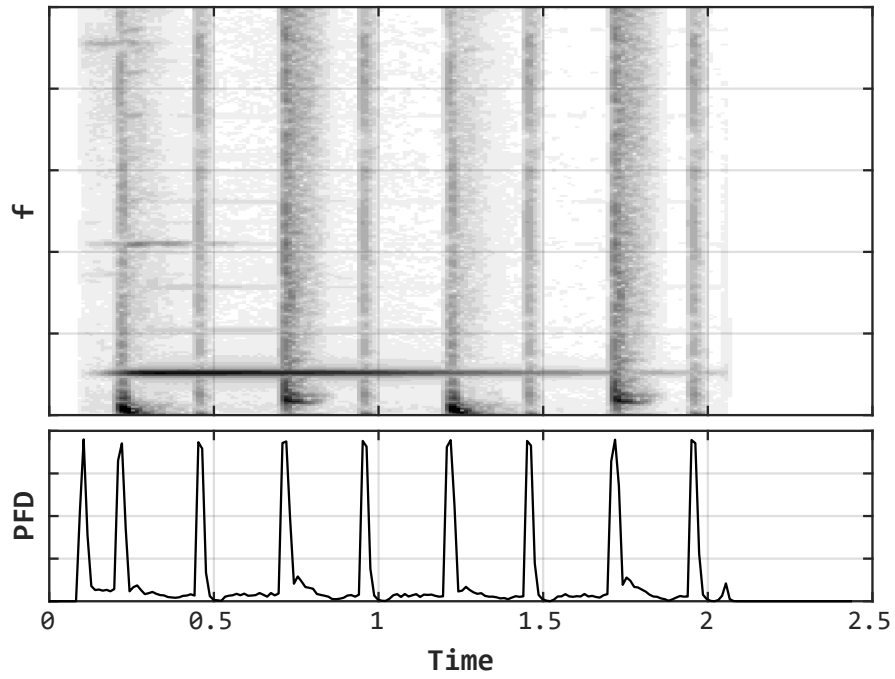


Figure 3.5: Top plot shows the $STFT$ of a drum loop mixed with a note of Vibraphone, the bottom plot shows percussive feature detection signal

3.3 Pitched and Percussive Sounds Detection

In 2005, Barry *et al.* [63] developed a drum source separation technique using percussive feature detection and spectral modulation.

At first a percussive temporal profile is derived by analysing each frame of a $STFT$ of the signal. The frame is then scaled according to this measure. That regions of the spectrogram with low percussive measures are scaled down significantly (Fig. 3.5), thus only the percussive regions remain.

The percussive feature detection (PFD) is performed for each frame by counting how many frequency bins of the log-difference of consecutive $STFT$ frames exceed a certain threshold. The output of the comparison with the threshold value can also be used as a binary spectral mask for isolating percussive sound from non-percussive sound. Section 3.4 will provide some perceptual motivation for this approach.

In 2012 Salamon and Gómez [64] introduced a system for automatically extracting the main melody of a polyphonic piece of music from its audio signal. In their work four signal processing steps are involved:

1. Sinusoid extraction: An $STFT$ analysis, enhanced by some pre- and post-processing, provides a number of spectral peaks.

2. Salience function: Peaks are given a score based on their harmonic relationship.

3. Pitch contour creation: A first selection subdivides the peaks into *best candidates* and *reserves*. Best candidates are used to create pitch contour segments, while reserves are used in case they can fill small gaps. Various statistics are gathered for each pitch contour.

4. Melody selection: Pitch contours are grouped based on gathered statistics in order to guarantee a monophonic and glitch-free melody line.

In a broad sense, this technique can be described as a way to track over time the fundamental frequency f_0 of a sparse and almost-harmonic signal in a noisy environment, relaying on prior statistical knowledge about the phenomena. Answering to ***Q2*** could introduce further information useful in steps 2 and 3, since information about neighbour frames can help predicting pitch trajectories.

The same year, they also proposed a different strategy, where f_0 tracking is enhanced based on vocal signal separation from other pitched or percussive sound [65]. The main motivation is that certain contexts, or music genres, are more difficult than others (for a review on this topic see [66]).

The assumption they made is that pitched sounds are modelled by sparseness in frequency and smoothness in time; percussive sounds are modelled by smoothness in frequency and sparseness in time; while vocal sounds are modelled by both sparseness in frequency and in time.

The proposed singing voice separation is realized with NMF, and is composed of three stages: a manual subdivision of signal in *singing* v.s. *non-singing* segments; a training phase where percussive bases and pitched bases are learnt from non-singing segments; and a separation phase where voice is extracted from singing segments as a *remainder*.

The most important thing about these works is that they all rely on how energy is organized in linear patterns in the (log)spectrum. This remark has been exploited in two different ways, as exposed in the next sections. Of course, this context is strictly related to ***Q2***.

3.4 Auditory Receptive Fields Models

In 2015 Lindeberg and Friberg [51, 52] described a theoretical and methodological framework to define computational models for auditory receptive fields (ARF). The proposal is based on a two-stage process: (i) a first layer of frequency selective temporal receptive fields, where the input signal is represented as multi-scale spectrograms, which can be specifically configured to simulate the physical resonance system in the cochlea spectrogram; (ii) a second layer of spectro-temporal receptive fields which consist of kernel-based 2D processing units in order to capture relevant auditory changes in both time and frequency dimensions.

The model is closely related to biological receptive fields (i.e. those that can be measured from neurons in ICC and A1). This work unifies in one theory a way to axiomatically derive representations like Gammatone or Gabor filters.

A set of new auditory features are proposed, being the result of the output 2D spectrogram after the kernel-based processing, using different operators like: spectro-temporal smoothings, onset and offset detections, spectral sharpenings, ways for capturing frequency variations over time and glissando estimation. An example of some of these kernels is shown in Fig. 3.6.

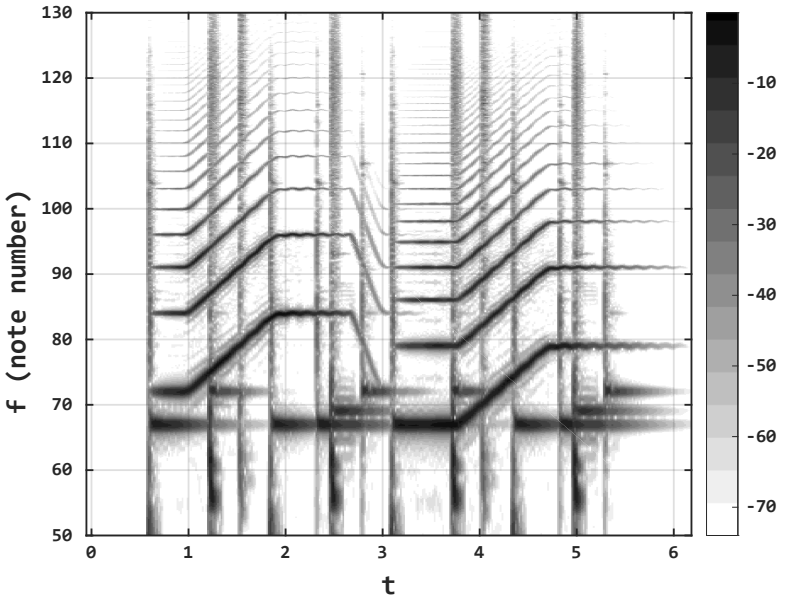
The work is interesting in this context since the information highlighted by the proposed operators can be compared with that addressed in Chapter 5, which, in turn, is aimed at answering ***Q2***. Moreover, it is interesting to see that with the correct kernels, this model can be used to isolate percussive or pitched signal in a way similar to how Gabor filters are used to find contours in image processing. Finally note how the act of differentiating along the time direction for the computation of PFD described in Section 3.3 resembles the *onset* kernel depicted in Fig. 3.6, providing a perceptual motivation to the model.

3.5 Radon-Based Spectral Features

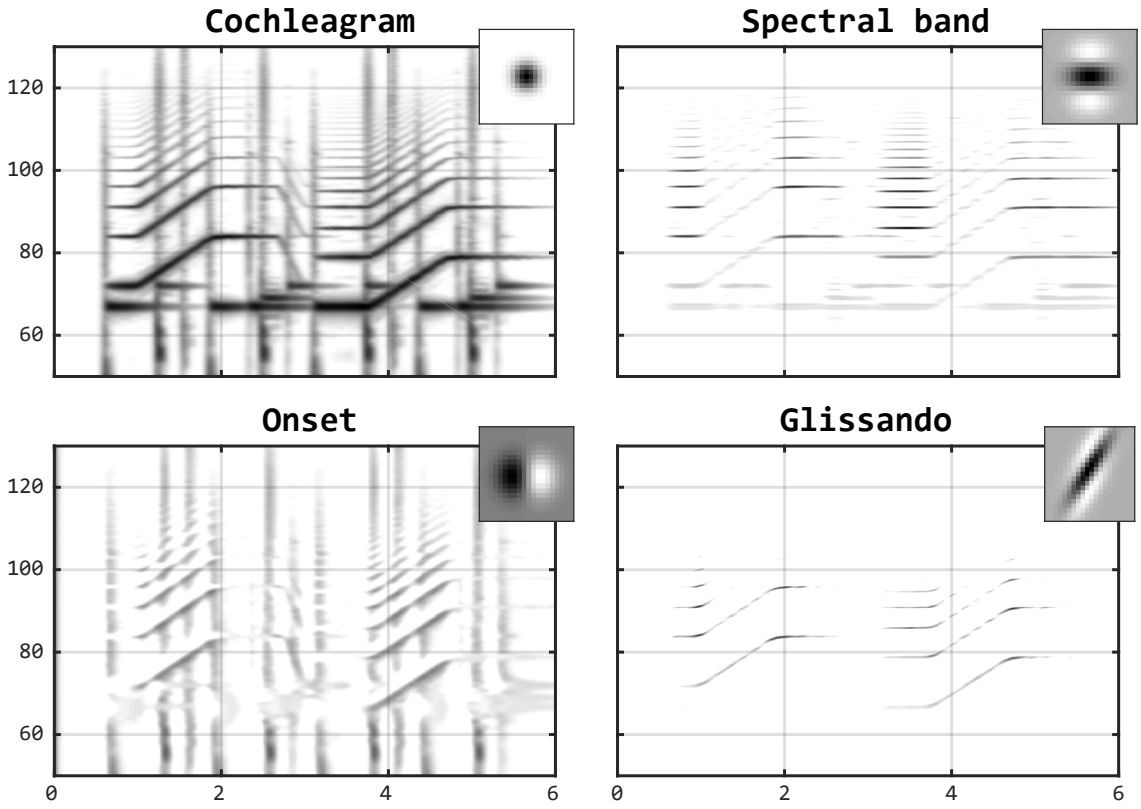
It is not rare to see image processing techniques applied to audio, and vice versa, but only a few works have been found in the literature regarding Radon transform applied to audio. After providing an example of the Radon transform applied to a $|STFT|$ in Fig. 3.7, some of these works are described.

In 2003, Özer *et al.* [67] presented a statistical method to detect the presence of hidden messages in audio signals (steganalysis). The basic idea is that, the distribution of various statistical distance measures, calculated on cover audio signals and on stego-audio signals compared to their denoised versions, are statistically different.

Among other features, the *Short-Time Fourier-Radon Transform Measure* is introduced: the mean-square distance of Radon transforms of the $STFT$ of two signals



(a) The $|STFT|$ of a loop composed of drums, xylophone and a synthesizer glissando (both frequency and amplitude are in logarithmic scale)



(b) Different kernels simulating different auditory receptive fields

Figure 3.6: ARF input and output

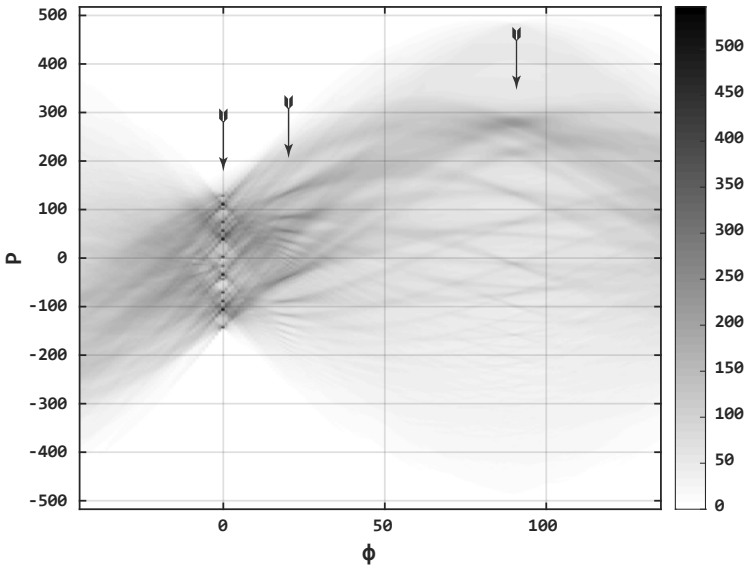


Figure 3.7: Radon transform of the $|STFT|$ in Fig. 3.6a. The three arrows highlight the peak columns caused respectively by the pitched sounds, the glissando and the percussive sounds. From the angle ϕ and the $STFT$ parameters it is possible to infer the speed of the glissando.

is defined as a new objective audio quality measure.

In particular, given the $|STFT|$ of a signal, its time projection gives the magnitude spectrum while its frequency projection yields the magnitude of the signal itself. More generally, rather than taking only the vertical and horizontal projections, they considered all the other angles, obtaining the Radon transform of the $STFT$ mass.

In 2001 Ajmera *et al.* [68] presented a new feature extraction technique for speaker recognition based on Radon transform and discrete cosine transform (DCT). The rationale of this technique lies in formulating the speaker recognition problem as the pattern recognition of images, and resolving it using machine learning tools.

In the proposed method the Radon transform is used to derive the effective acoustic features from the speech spectrogram: projections for seven orientations capture the acoustic characteristics of the spectrogram, then a DCT is applied on Radon projections to lower feature vector dimensionality.

In the same year, Myung Kim and Hoirin Kim [69] focused on the problem of classifying pornographic sounds, such as sexual scream or moan, to detect and block the objectionable multimedia contents. To represent the large temporal variations of pornographic sounds, they proposed a novel feature extraction method based on Radon transform.

They suppose that, compared with speech and music signals, the pornographic sounds show large temporal variations, fast spectral transitions among neighbouring frames, and a typical pitch at about 500 Hz. In addition, these characteristics tend to periodically repeat.

In their proposal, the Radon transform of time-frequency spectrograms is used to effectively capture the spectral characteristics of a pornographic sound.

In this context, their work outperforms strategies based on other features, such as mel-frequency cepstral coefficients. The technique is very effective for pornographic sound detection because the large variations among adjacent spectral signals produce distinct orientations in Radon domain.

All of these works exploit Radon based features, but no proposal has been found able to perform masking of signals directly (or depending on) manipulations in the Radon space. The main reason is that is very hard to describe non-linear operations in the Radon space, the effects of which can propagate back to the original space. In Chapter 5 a solution to this problem is given while answering ***Q*2**.

Chapter 4

Bivariate Mixture Space

4.1 Rationale

The basic idea behind the concept of *bivariate mixture space* (BMS) is to interpret a bivariate stationary signal in the spectral domain $\mathbf{X}(f) = \{X_1(f), X_2(f)\}$ as two orthogonal observations of an underlying continuous space $\tilde{X}(f, \alpha)$ such that

$$\begin{aligned} X_1(f) &= \tilde{X}(f, 0) \\ X_2(f) &= \tilde{X}(f, \tfrac{\pi}{2}) \end{aligned} \tag{4.1.1}$$

Moreover, their rotation $X_m(f)$ and $X_s(f)$ (see Eq. 2.3.1) should sample the space in $\alpha = \frac{\pi}{4}$ and $\alpha = \frac{3}{4}\pi$ respectively.

In other words $\tilde{X}(f, \alpha)$ can be seen as a revolving surface that interpolates $\mathbf{X}(f)$. This is done on the Fourier decomposition of the input, since the spectral domain enables the ability to work on each trigonometric base separately.

Actually no assumptions are made explicit regarding sources mixed in $\mathbf{X}(f)$, nevertheless correlated components of the signals should appear as peaks in $\left|\tilde{X}(f, \alpha)\right|$ for some α, based on how they are distributed into the mixtures, while uncorrelated components should present no peaks (i.e. they are equally distributed in $\tilde{X}(f, \alpha)$).

Any source $U_i(f)$ will spread its energy differently in $\tilde{X}(f, \alpha)$, according to how it has been mixed in $\mathbf{X}(f)$, in this case $\tilde{X}(f, \alpha)$ can be seen as an *analysis* tool. On the other hand, some knowledge on how sources are distributed, renders $\tilde{X}(f, \alpha)$ a *decomposition* tool.

Finally, studying $\tilde{X}(f, \alpha)$ it should be possible to represent $\mathbf{X}(f)$ in a more compact form, which encapsulates all the properties of the starting set of variables but in a more meaningful way.

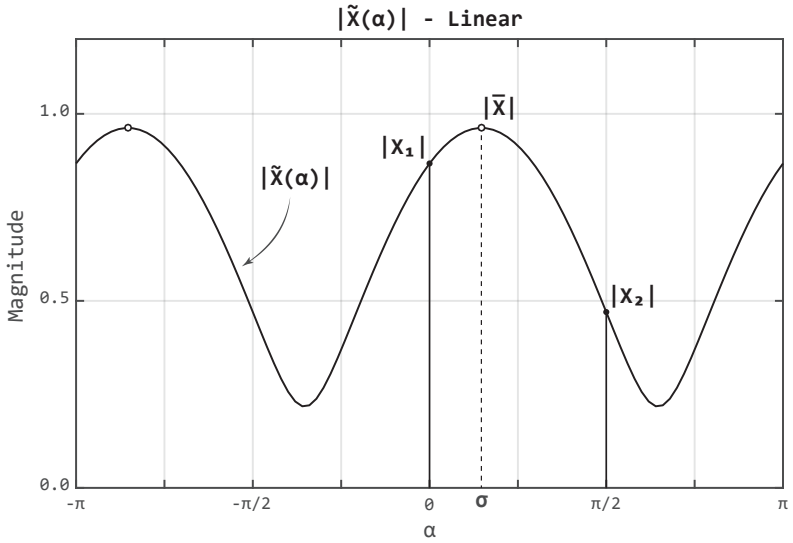


Figure 4.1: Magnitude of the BMS $\left|\tilde{X}(f,\alpha)\right|$ for a fixed f. Input X_1 and X_2 have different magnitude and phase. Principal component $\left|\overline{X}\right|$ can be seen at angle σ, while the second principal component can be picked up in $\sigma - \frac{1}{2}\pi$.

4.2 Definition

4.2.1 Bivariate Mixture Space

Let suppose a stationary latent signal $U_1(f)$ is distributed in $\mathbf{X}(f)$. The act of feeding the observations $X_1(f)$ and $X_2(f)$ with different amounts of $U_1(f)$ can be generalized by placing the source at some angle $\sigma(f)$ between the orthogonal axes (that is, the basic principle behind PCA)

$$\begin{aligned} X_1(f) &= \cos(\sigma(f)) \cdot U_1(f) \\ X_2(f) &= \sin(\sigma(f)) \cdot U_1(f) \end{aligned} \tag{4.2.1}$$

Now let define the transformation of $\mathbf{X}(f)$ space into the BMS $\tilde{X}(f,\alpha)$ that satisfies the desiderata described in Section 4.1:

$$\tilde{X}(f,\alpha) = X_1(f) \cdot cos(\alpha) + X_2(f) \cdot sin(\alpha) \tag{4.2.2}$$

This complex function has a periodicity of π and its magnitude peaks in correspondence with the angle $\sigma(f)$ with a value equal to $|U_1(f)|$. The behaviour of $\tilde{X}(f,\alpha)$ varying $\sigma(f)$ for a fixed f can be seen in Fig. 4.1 and 4.2.

If $U_1(f)$ is distributed in $\mathbf{X}(f)$ by some convolutive phenomena, difference in phase between $X_1(f)$ and $X_2(f)$ occurs, which in turn affects $\tilde{X}(f,\alpha)$ as depicted in Fig. 4.3.

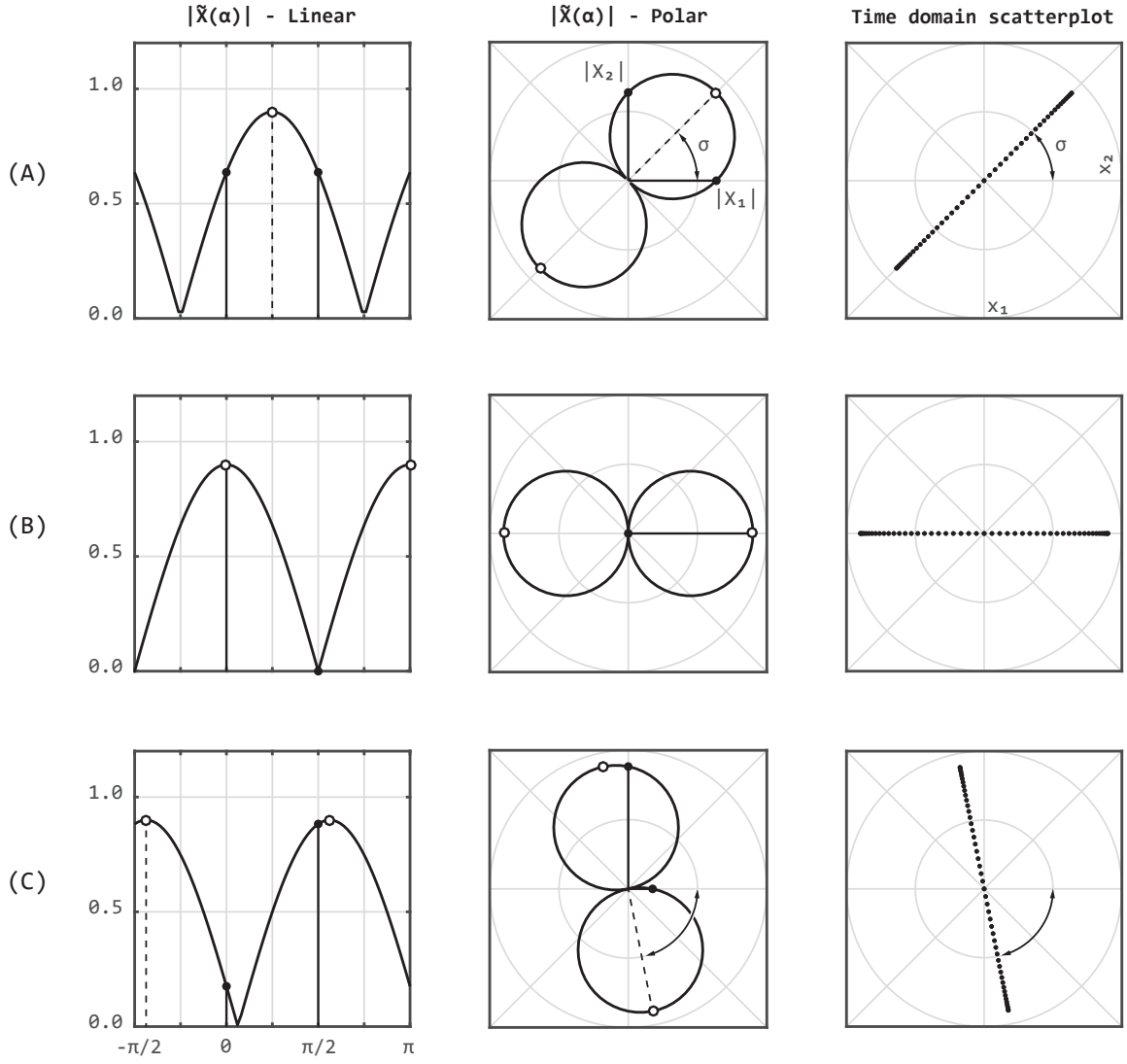


Figure 4.2: $\left|\tilde{X}(f,\alpha)\right|$ (first two columns) and scatterplot in the time domain (third column) for a fixed f. In the first two rows X_1 and X_2 share same phase but different magnitude, while in the third row they have different magnitude and opposite phase. In **A** $X_1 = X_2$, in **B** $X_2 = 0$ and in **C** $X_1 = -0.2X_2$. For a legend of symbols and axes see Fig. 4.1.

Despite those phase differences, $\left|\tilde{X}(f,\alpha)\right|$ will always peak in correspondence of the principal component angle (as long there is one).

4.2.2 Bivariate Spectrum

To better represent the information present in the two spectra contained in $\mathbf{X}(f)$, a more compact transformation can be used. First, the angle $\sigma(f)$ where energy is concentrated can be inferred by finding the peaks of $\left|\tilde{X}(f,\alpha)\right|$, that are the zeroes of its first derivative, occurring at ($\Im$ and $\Re$ denote Imaginary and Real parts):[1]

$$\sigma(f) = \frac{1}{2}\arctan\frac{2\,\Im X_1\,\Im X_2 + 2\,\Re X_1\,\Re X_2}{|X_1|^2 - |X_2|^2} \tag{4.2.3}$$

Then, note that the linear correlation $C(f)$ between bases in $X_1(f)$ and $X_2(f)$ can be computed as a function of their phase difference:

$$C(f) = \cos(\angle X_1(f) - \angle X_2(f)) \tag{4.2.4}$$

Finally, instead of computing the whole $\tilde{X}(f,\alpha)$ surface, just the principal components for each f can be saved as a *principal spectral content* (PSC) $\overline{X}(f)$ which somehow discards the relational information:

$$\overline{X}(f) = \tilde{X}(f,\sigma(f)) \tag{4.2.5}$$

then the relational information contained in $\sigma(f)$ and $C(f)$ can be stored as a new *Relational vector* $R(f)$:

$$R(f) = |C(f)| \cdot e^{i\sigma(f)} \tag{4.2.6}$$

finally all of this information can be packed into a vector $\vec{X}(f)$ called *bivariate spectrum* (BS)

$$\vec{X}(f) = \{\overline{X}(f),\, R(f)\} \tag{4.2.7}$$

which is more meaningful than $\mathbf{X}(f)$ since it discards no information, but organizes it in a more straightforward form separating spectral content from relational content:

- $\overline{X}(f)$ accounts for the overall magnitude and phase of the spectral content of the bivariate mixture (note that by simply summing $X_1(f)$ and $X_2(f)$ destructive phase interference may occur)
- $\sigma(f)$ accounts for the *balance* of the energy, i.e. it provides information about how each input mixture contributes to the PSC, and can be used to retrieve the principal components of the signal
- $C(f)$ provides information about correlation and phase differences of the input mixtures at a single base frequency level

[1] for the sake of readability, the f argument has been omitted for X_1 and X_2

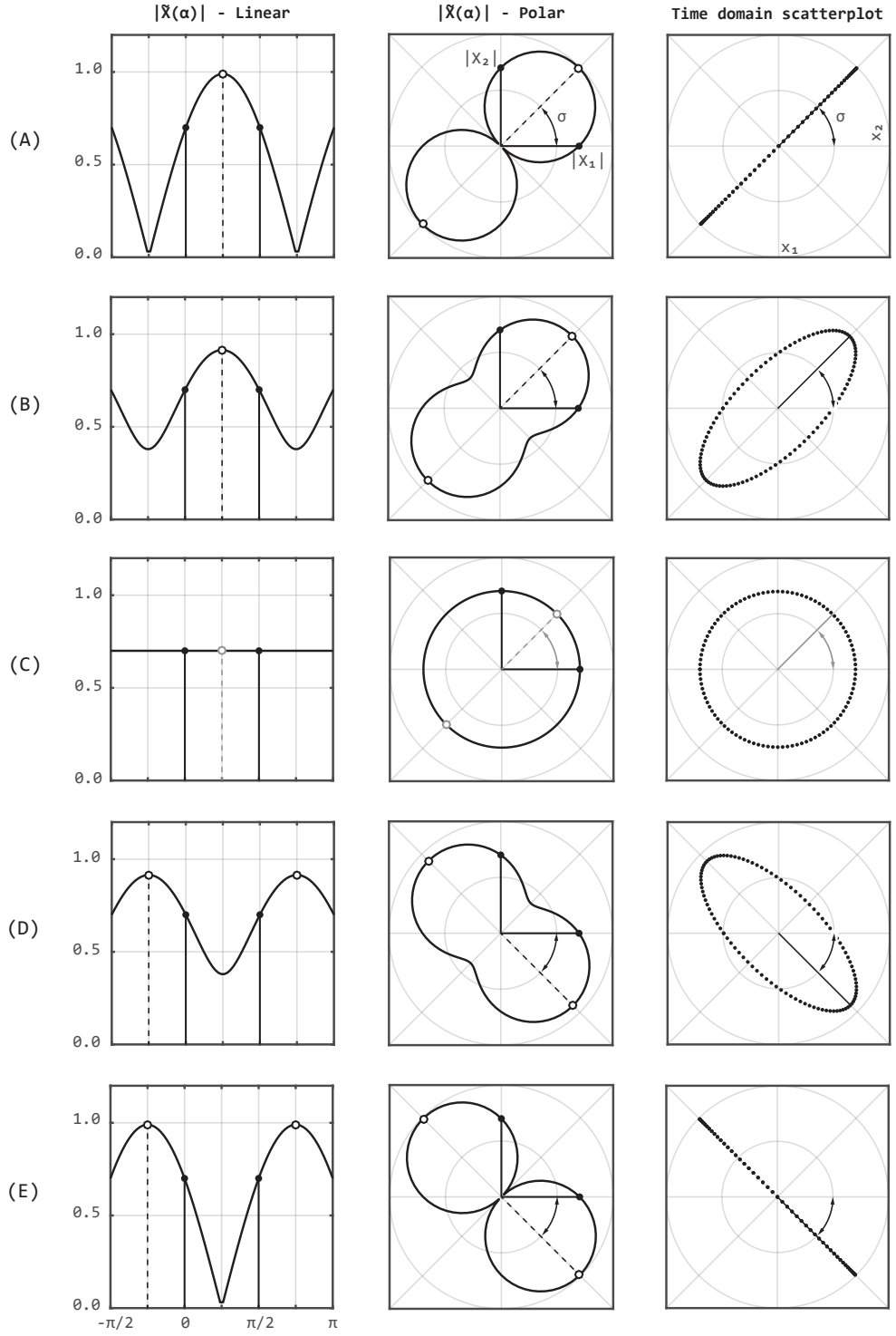


Figure 4.3: $\left|\tilde{X}(f,\alpha)\right|$ and scatterplot in the time domain for a fixed f with X_1 and X_2 with same magnitude but different phase. Phase difference in each row is **A**: 0; **B**: $\frac{1}{4}\pi$; **C**: $\frac{1}{2}\pi$; **D**: $\frac{3}{4}\pi$; **E**: π. Note the negative angle for anti-phase cases **D** and **E**, also characterized by a destructive interference in the space between X_1 and X_2. Also note the ambiguity in case **C** of uncorrelated signals. For a legend of symbols and axes see Fig. 4.1.

4.3 Properties and Applications

A way to visualize the BS together with properties and examples of use for BMS are provided in the next sections. What follows is a recap of the notation introduced, from now on the argument f will be omitted for improved readability

- U_i is a phenomenon observed by $\mathbf{X}$ (Eq. 4.2.1)
- $\mathbf{X} = \{X_1, X_2\}$ is the input bivariate signal, composed of two mixtures
- $\tilde{X}(\alpha)$ is the bivariate mixture space (Eq. 4.2.2)
- $\vec{X} = \{\overline{X}, R\}$ is the bivariate spectrum (Eq. 4.2.7)
- $\overline{X}$ is the principal spectral content (Eq. 4.2.5)
- R is the relational content vector (Eq. 4.2.6)
- σ is the value of α that maximizes $\tilde{X}(\alpha)$ (Eq. 4.2.3)
- C is the correlation between X_1 and X_2 (Eq. 4.2.4)

In this chapter also the following symbols will be introduced as manipulators of the input mixture

- γ is the power which $|C|$ is usually raised to weight correlation relevance
- $(\cdot)^{\{M\}}$ is a masking operation M, in particular:
  - $(\cdot)^{\{\theta_l^h\}}$ is a selection of components within $\theta - l < \sigma < \theta + h$ (Eq. 4.3.1)
  - $(\cdot)^{\{+\}}$ is a selection of in-phase components (Eq. 4.3.2b)
  - $(\cdot)^{\{-\}}$ is a selection of anti-phase components (Eq. 4.3.2c)
  - $(\cdot)^{\{1\}}$ is a selection of clear, correlated components (Eq. 4.3.3a)
  - $(\cdot)^{\{0\}}$ is a selection of cluttered, uncorrelated components (Eq. 4.3.3b)
- $X_\theta = \tilde{X}(\theta)$ is a new observation of $\mathbf{X}$ at angle θ (Eq. 4.3.4)
- $\mathbf{X}_\theta = \{X_\theta, X_{\theta+\frac{\pi}{2}}\}$ is a new bivariate signal after a rotation θ (Eq. 4.3.5)

Examples and applications of BMS manipulation and visualization can be found in the GitHub repository.

4.3.1 Signal Manipulation

$\tilde{X}(\alpha)$, σ and C can be exploited to perform two different kinds of data manipulation: *spectral masking* and *mixture resampling*.

Spectral masking is a simple way to control the magnitude of specific bases of the mixture by using σ as key to mask parts of the spectrum, thus isolating components at particular positions of the BMS. For example, given an angle θ and upper and lower bounds h and l, a simple notation of masking can be introduced as (in principle also non-binary masks are possible):

$$mask = \begin{cases} 1 & \text{if } \theta - l < \sigma < \theta + h \\ 0 & \text{otherwise} \end{cases} \tag{4.3.1a}$$

$$X^{\{\theta_l^h\}} = X \cdot mask \tag{4.3.1b}$$

Special masks can be realized by using C as key, for example to split any X in $X^{\{+\}} + X^{\{-\}}$ containing respectively positively and negatively correlated components. The first may also be referred as the *in-phase* signal, while the latter is the *anti-phase* signal

$$mask = \begin{cases} 1 & \text{if } C \geq 0 \\ 0 & \text{if } C < 0 \end{cases} \tag{4.3.2a}$$

$$X^{\{+\}} = X \cdot mask \tag{4.3.2b}$$

$$X^{\{-\}} = X \cdot (1 - mask) \tag{4.3.2c}$$

C can also split X in $X^{\{1\}} + X^{\{0\}}$ containing respectively correlated and uncorrelated components. The first may also be referred as the *clear* signal, while the latter is the *cluttered* signal (γ is a *clarity parameter* used to emphasize relevance of $|C|$)

$$X^{\{1\}} = X \cdot |C|^{\gamma} \tag{4.3.3a}$$

$$X^{\{0\}} = X \cdot (1 - |C|^{\gamma}) \tag{4.3.3b}$$

Finally, it may be useful to define some simple syntaxt for expressing basic boolean operations between two masks a and b such as

- $a \cdot b$ or logical *AND*: $X^{\{a,b\}}$
- $a + b$ or logical *OR*: $X^{\{a\},\{b\}}$
- $1 - a$ or logical negation: $X^{\{\bar{a}\}}$

Mixture resampling is a resynthesis process which relays on $\tilde{X}(\alpha)$ to generate new mixtures or to rotate the mixture space. A new mixture X_θ can be synthesized by choosing any θ as argument for $\tilde{X}(\theta)$.

$$X_\theta = \tilde{X}(\theta) \tag{4.3.4}$$

In principle, mixture resampling may provide new observations of $\mathbf{X}$. Those observations may seem to not overcome the constraint of ICA to provide as many mixtures as the latent sources, since new observations are basically a linear combination of the input ones. However, a recent work by Fitzgerald *et al.* [70] demonstrates how, under certain conditions, it is possible to get over these limitations by working on ensembles of resamplings in combination. This aspect will be properly addressed in the future.

Finally, rotation is realized simply by resampling the input at an angle θ and $\theta+\frac{\pi}{2}$, in order to create new orthogonal output mixtures.

$$\mathbf{X}_\theta = \{X_\theta,\, X_{\theta+\frac{\pi}{2}}\} \tag{4.3.5}$$

4.3.2 Distributions of Components in the BMS

Among the information that can be collected from the BS, it is worth citing the distribution of components in the BMS. However collecting the mere distribution of σ can produce misleading graphs, since bases with different magnitudes have the same influence on the distribution. So, to plot the components dispersion correctly, it may be useful to weight the distribution by $\left|\overline{X}\right|$ to account for the actual signal content. Nevertheless, in some cases, also the $\left|\overline{X}\right|$ weighting can be misleading, since bins with low $|C|$ are placed at an angle σ which is very likely incorrect (see Fig. 4.3, row *c*). To overcome this issue, it may be useful to consider the more sophisticated weighting $\left|\overline{X}^{\{1\}}\right|$ to see the distribution of correlated components, or $\left|\overline{X}^{\{0\}}\right|$ to see the contribution of uncorrelated components. Differences are shown in Fig. 4.4.

Note that in case of $STFT$ inputs, the same distributions may be computed also in a *per-frame* or *per-frequency* base, providing more punctual information about the mixture, as shown in Fig. 4.5.

4.3.3 BS-Enhanced Spectrogram

When analysing signals, visualizing data is part of the process. Usually spectrograms (i.e. plotting $STFT$ across time and frequency using colours for coding magnitude) are the first tool that comes in mind for inspecting signal frequency content over time. Unfortunately, the relationship between the components of bivariate signals are hard to see with this method. A naive approach is to display the sonogram of the sum of X_1 and X_2. In this way, those components which are not in phase are cancelled from

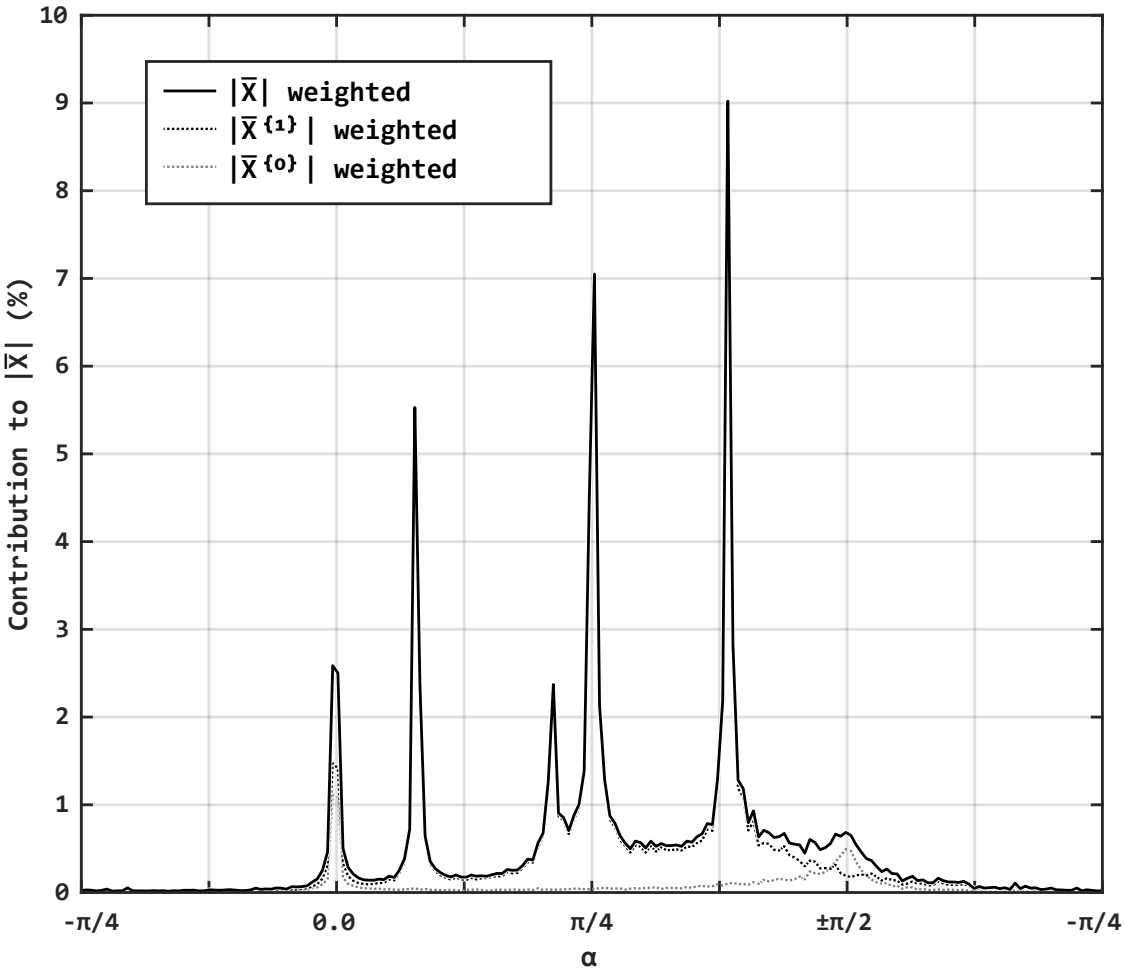
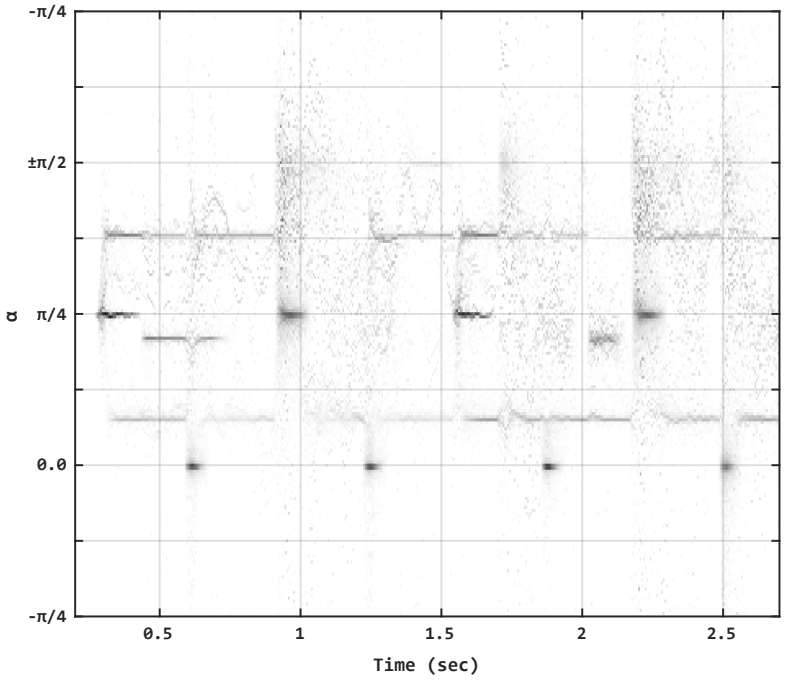
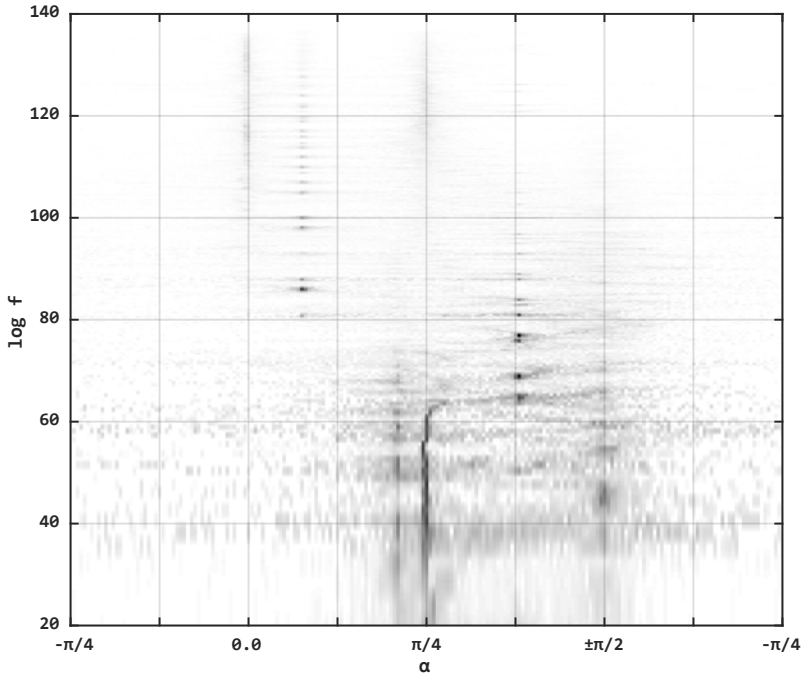


Figure 4.4: Distribution of total signal energy over the BMS of the same signal depicted in Fig. 4.6. Solid line displays overall distribution, while dotted lines shows energy distribution of correlated and uncorrelated components. Horizontal axis has been shifted to display the common components in the centre of the picture.



(a) Energy distribution over time and mixture space.



(b) Energy distribution over frequency and mixture space.

Figure 4.5: Total signal energy distribution for the same signal depicted in Fig. 4.6. α axis has been shifted to display the common components aligned with the centre of the picture.

the plot. So, a more refined way to mix the spectra is to sum only the magnitude. Anyhow, in the classic $|X_1|+|X_2|$ spectrogram, no relational information is preserved. At the same time, the display of two separate sonograms for the inspected bivariate mixture is not really easy to interpret and introduces a lot of redundancy.

A new kind of sonogram, visible in Fig. 4.6, is introduced as a way to represent BS, which encodes $\left|\overline{X}\right|$ with brightness, σ with hue and $|C|$ with saturation. Of course magnitude can be expressed in dB and the frequency axis can be in a logarithmic scale. Since variation in saturation may be hard to notice, correlation visualization may be tuned by choosing proper γ. Empirical tests show that values between $\gamma = 0$ (ignoring correlation and thus relative phase information) and $\gamma = 4$ (strongly emphasizing correlation) may fit most of the situations.

This method reduces the redundancy of having two separate sonograms and highlights mixtures differences, without discarding any information but absolute phase (relative phase is encoded with $|C|$ and σ sign). Moreover low-level visual cues such as brightness, hue and saturation are processed by our brain faster than the pattern recognition task needed to compare two separate sonograms [71].

For example, if a colourful image appears, it means that the observed signals are strongly correlated, while if a grey-scale image appears, it means that signals have a very low absolute correlation.

This kind of spectrogram helps interpreting the bivariate couple more as a continuum than a discrete set of mixtures, and packs a wide range of information in a compact area, letting the user recognize signal properties at a glance.

On the other hand, the subjective visual brightness of different colours could reduce the perception accuracy of the components magnitude, but this issue is not critical, since precise magnitude measurements are more affordable in a classical 2D magnitude profile plot.

4.3.4 Relation With Other Methods

Let's compare this work with some of the SMC techniques described in Section 3.2.

The azimuth space described in ADRess [57] is very similar to BMS, but with some key difference. As it can be seen comparing Fig. 3.3 with Fig. 4.7, the azimuth space is only defined between the centre position and a channel side, while BMS is defined in proper angles between $\pm\frac{1}{2}\pi$, spanning across all possible signal positions. Moreover, sources are found in ADRess by scanning the whole space for some minima, then source magnitude is properly reconstructed and phase is taken as that of the input. In the BMS sources are visible as peaks, whose position can be found with Eq. 4.2.3 without scanning the whole space, while exact source magnitude and phase at position θ can be properly computed as a mixture resampling operation $\tilde{X}(\theta)$.

Moreover, excluding or adding back cluttered components from the masked signal

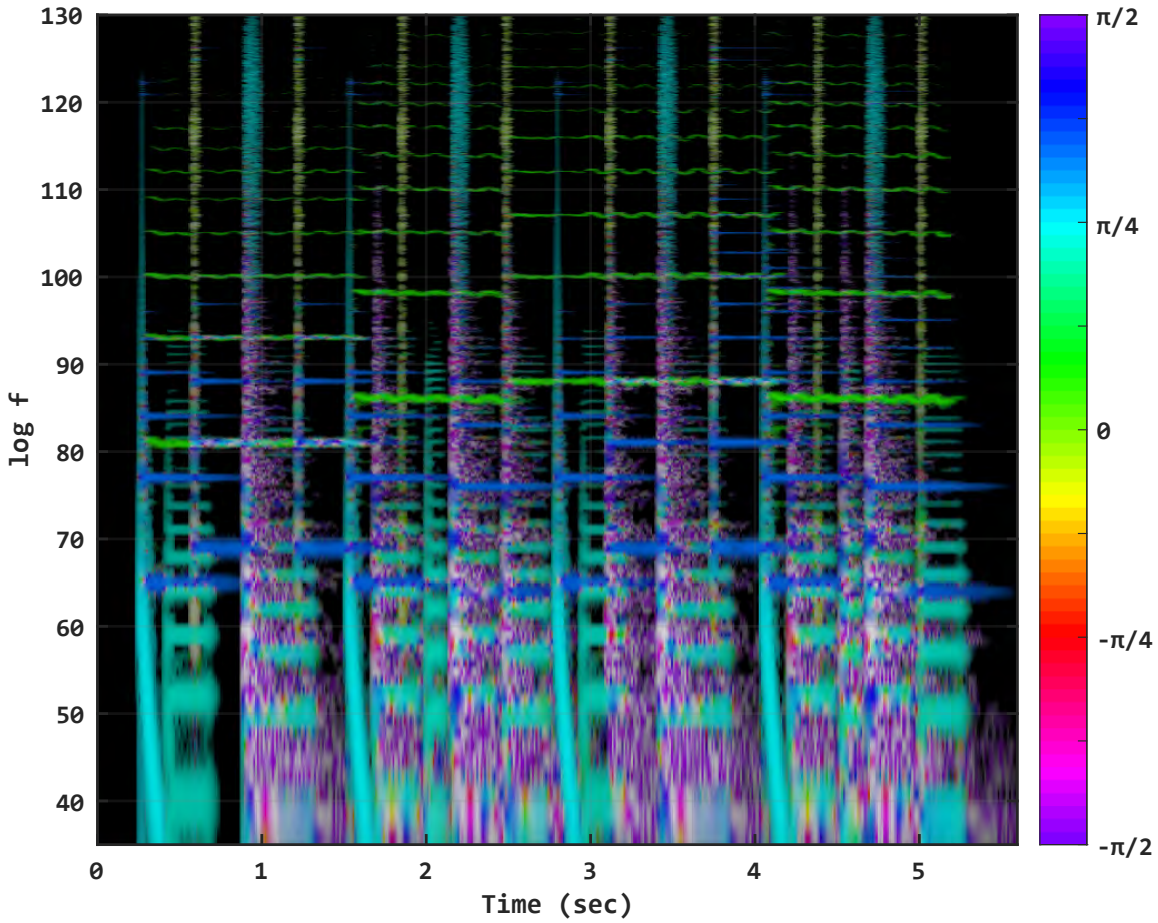


Figure 4.6: BS-enhanced spectrogram. Hue is linked to σ, saturation to $|C|^{\gamma}$ and brightness to $\log(|\overline{X}|)$. In this example an audio file containing different instruments panned in the stereo image is shown. Visible sources are: a violin in green; a kick drum and a snare in light blue; a bass guitar in teal; an electric piano in dark blue; a hi-hat in yellow, and a taiko drum in purple. $\sigma = 0$ is left, $\sigma = \pi/2$ is right and $\sigma = \pi/4$ is the middle position. $\sigma = -pi/4$ denotes anti phase signals.

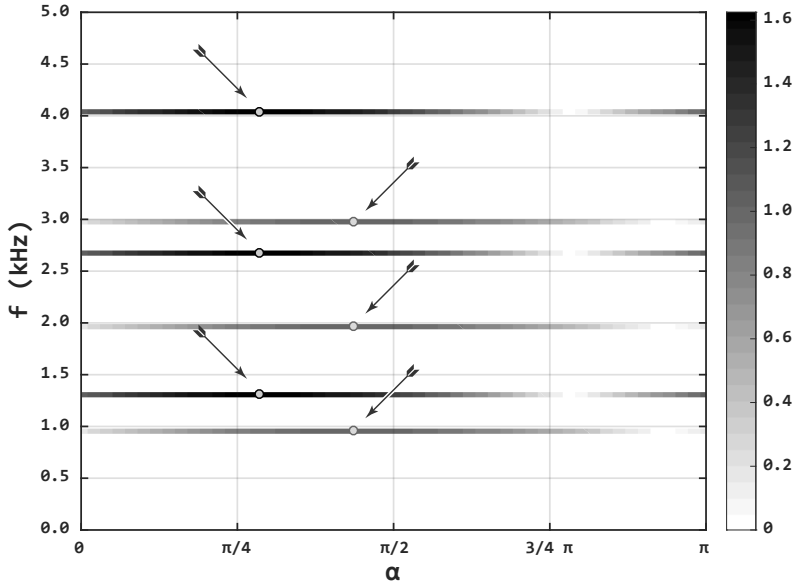


Figure 4.7: Sources as peaks of the BMS, compared with sources as minima of the azimuth space of Fig. 3.3.

may help in properly reconstructing the sources. Two new extraction methods of a source U_θ may then be defined as

$$U_\theta = X_\theta^{\{\theta_l^h\}} \tag{4.3.6}$$

as a simple masking and resampling operation (M&R), and

$$U_\theta = \begin{cases} X_\theta^{\{+,\theta_l^h\}} & \text{if excluding all cluttered components} \\ X_\theta^{\{+,\theta_l^h\},\{-\}} & \text{if including cluttered components outside } \theta_l^h \end{cases} \tag{4.3.7}$$

as masking and resampling with taking into account also cluttered and clear components (M&R+C).

A comparison of the three methods can be seen in Fig. 4.8.

The test has been run over the signal in Fig. 4.6, with the same masking windows for ADRess, M&R, and M&R+C. $STFT$ settings are the same described in [57], and the scores are those from PEASS software.

It turns out using the correct phase information in the resynthesis process slightly improves the separation quality, while considering clear and cluttered components may improve the separation of some sources, but in general it can be used to trade-off between a good score in interference IPS and artefacts APS.

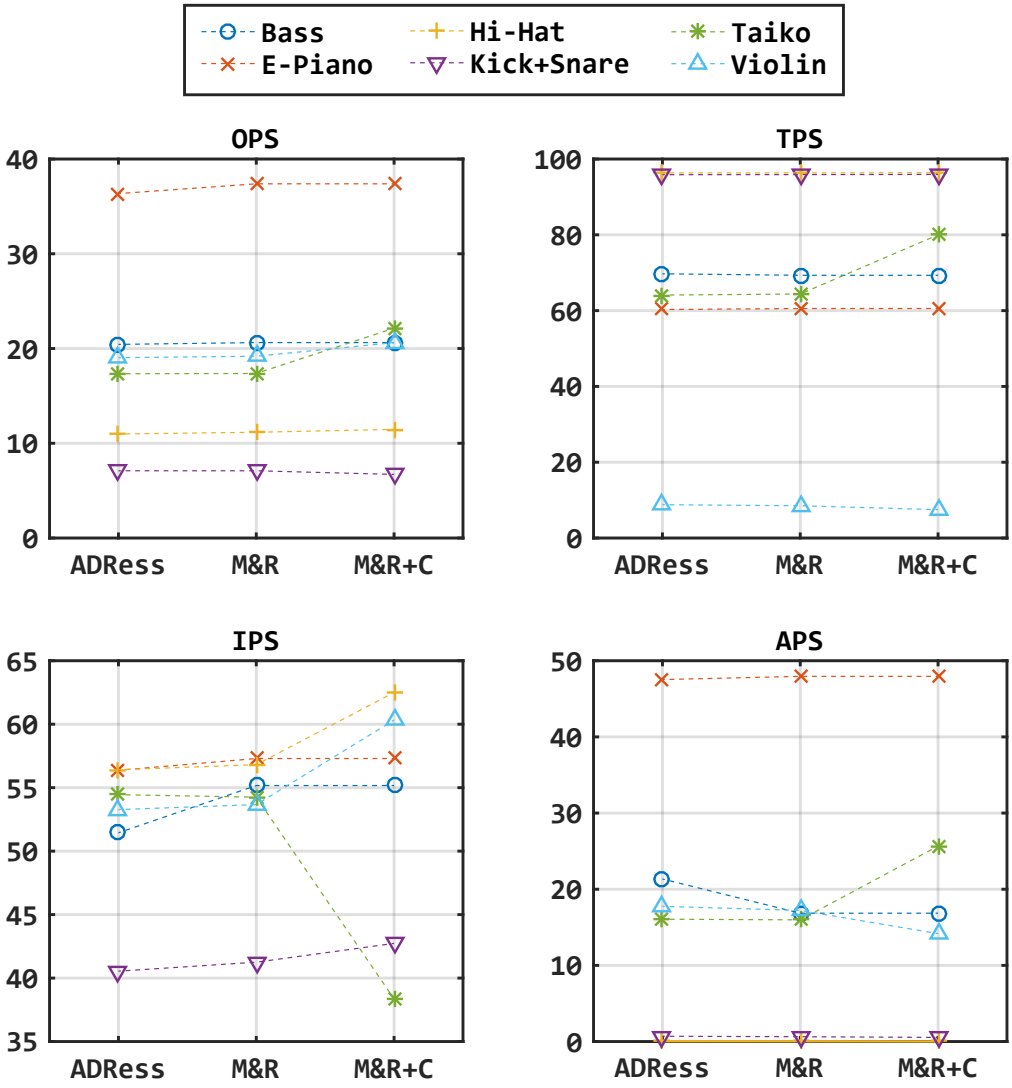


Figure 4.8: Comparison of PEASS results for three different source separation strategies based on stereo panning. Input is composed of a mixture of 6 sources.

Unfortunately, PEASS scores do not recognize the great improvement of transient quality that arises from using the correct phase instead of the original phase (more tests can be found in Section 6.1).

For what concerns execution time, ADRess took 0.624 seconds to perform the separation of a 6 seconds excerpt, while M&R and M&R+C took respectively 0.281 and 0.471 seconds. The test was ran in Matlab with a 64 bit operating system over a desktop computer equipped with an Intel $i7$ processor running at 3.1 Ghz and 16 Gb of ram memory. The azimuth space investigation of ADRess has been replaced with σ function, which was also used in M&R and M&R+C. A profiling of the execution time showed that the most time consuming task was the arctangent function needed to compute σ, C and the phase of input bins required by ADREss resynthesis technique.

Regarding the distinction between direct and diffuse signal realized by Kraft [61], note that the principal components (considered as direct signal) are all contained by definition in $\overline{X}$, while diffuse signal (defined as orthogonal to the direct one) can be retrieved by computing $\tilde{X}(\sigma + \frac{1}{4}\pi)$. Also in this case, phase must not be guessed, since all functions are defined natively in the complex domain. Endowing Kraft's assumptions, direct and wet signals U_d and U_w can be defined as the PSC and its orthogonal resampling

$$U_d = X_\sigma = \overline{X} \tag{4.3.8a}$$

$$U_w = X_{\sigma+\frac{\pi}{4}} \tag{4.3.8b}$$

Another method (referred as PSC+C) could consist in considering in part of the ambience signal also the uncorrelated components of the signal

$$U_d = X_\sigma^{\{1\}} = \overline{X}^{\{1\}} \tag{4.3.9a}$$

$$U_w = X_{\sigma+\frac{\pi}{4}} + X_\sigma^{\{0\}} \tag{4.3.9b}$$

Again, to test the improvements of correct phase estimation and C weighting, a PEASS test has been run over a sample signal, composed of a mixture of a dry voice with some reverberation (achieved by convolution with the impulse response of a large hall). Results are shown in Fig. 4.9.

It is clear how correct phase interpolation drastically improves the performance regarding the direct signal, while phase interpolation and C masking does not seem to help with the wet signal. Nevertheless the degradation in the wet signal seems to be outweighed by the improvements in the dry one.

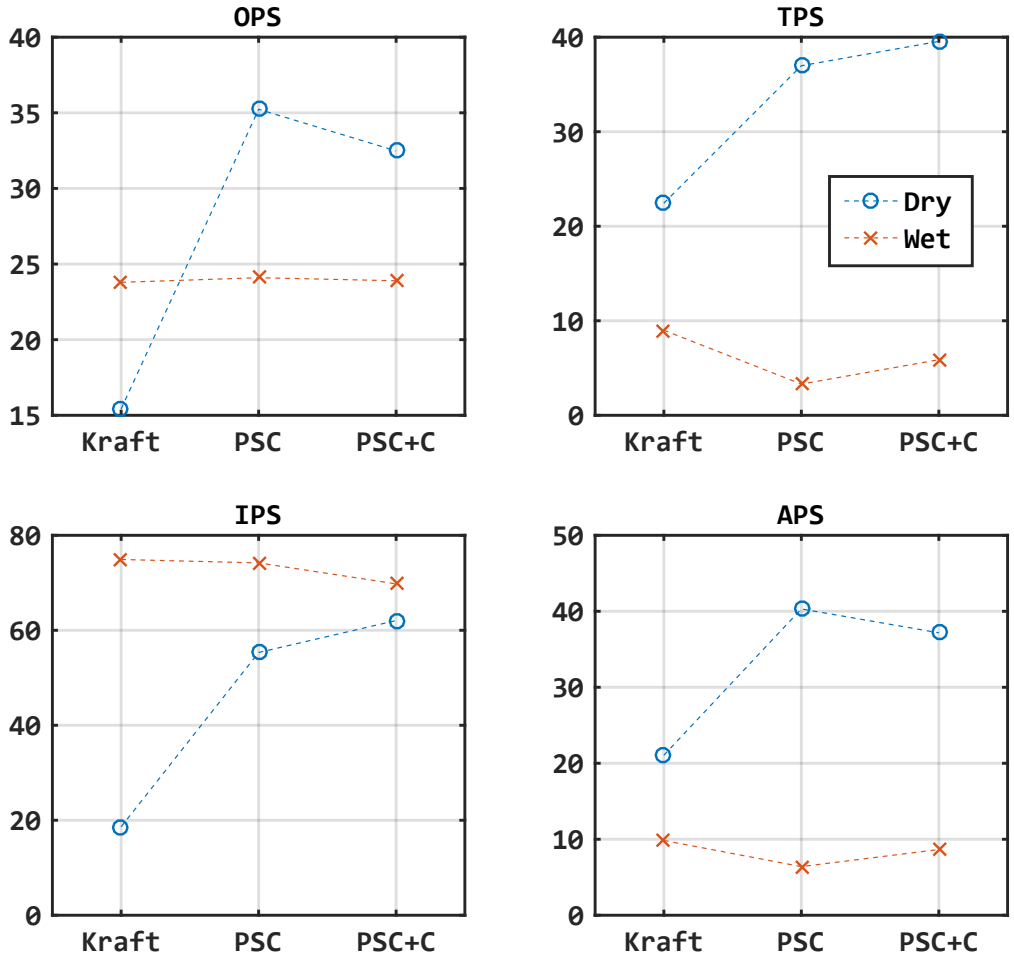


Figure 4.9: Comparison of PEASS results for three different dry/wet separation strategies based on frequency domain stereo decomposition. Input is composed of a mixture of voice and reverb.

With respect to execution time, Kraft took 0.212 seconds to separate 8 seconds of audio, while other methods took respectively 0.120 and 0.155 seconds. As usual as before, arctangent is the most time consuming operation.

Finally, note the similarity between the metrics of Eq. 3.2.2 and the BS defined in Eq. 4.2.7. They basically provide the same kind of information but, again, the main difference is that phase information is properly encoded as argument of $\overline{X}$, thus the original signal $\mathbf{X}$ can be reconstructed properly. Morover, as displayed in the third column of Fig. 4.2 and 4.3, BS information are suitable also for describing the ellipses parameters delineated in Section 3.1.

4.3.5 Other SMC Applications

With respect to music production techniques, decomposition of stereo audio mixtures through masking and resampling may be useful in upmixing, mastering, restoration, and noise suppression processes. For example, it is possible to generate multichannel audio $\mathbf{S}$ from a stereo track resampling $\mathbf{x}$ in more than two points. For example

$$\mathbf{S} = \{X_{\frac{\pi}{4}},\, X_{\frac{\pi}{8}},\, X_{\pi\frac{3}{8}},\, X_{-\frac{\pi}{8}},\, X_{-\pi\frac{3}{8}}\} \tag{4.3.10}$$

corresponding respectively to centre, left, right, surround-left, and surround-right channels.

Moreover, $X^{\{+\}}$ and $X^{\{+\}}$ can be exploited to perform selective phase correction along the spectrum, such as creating a mono compatible stereo signal $\mathbf{M}$ without interfering too much with the original stereo image

$$\mathbf{M} = \{X_0,\, X^{\{+\}}_{\frac{\pi}{2}} - X^{\{-\}}_{\frac{\pi}{2}}\} \tag{4.3.11}$$

that is inverting phase only for those components which are in anti phase.

Eventually, $\overline{X}$ can be used as a conservative mono version of the signal which minimizes phase interference usually introduced with channels summation.

Finally, from the study of $\vec{X}$ statistics it is possible to introduce new audio features based on stereophonic properties of the signal. An investigation that is left to future works on this topic.

Chapter 5

Spectro-Temporal Structure Field

5.1 Rationale

Sonograms ($STFT$s magnitude) are a useful tool for observing signal spectral behaviour in time, but it is hard to gather objective information regarding the features caught by the eye. One way to do this can be to apply the Radon transform to the $STFT$ matrix of a signal with the purpose of findings vertical, horizontal and oblique features of the spectrum (transients, periodicities and frequency modulations). Unfortunately the Radon transformation discards information regarding time and frequency. Starting from the study of a back-projection strategy that brings feature information back to the time frequency domain, a new analytical transformation has been found, which avoids completely the computation of the Radon transform and can add a layer of information upon the $STFT$ regarding the distribution of energy in linear patterns. [1]

This approach for spectral description is based on a more general idea regarding any $\mathbb{R}^2$ function, realized in two steps: first a *Signal Energy Angular Distribution* (SEAD) is computed for each point of the input function, that is a measure of how much energy is surrounding the point. Then the SEAD of each point is condensed into a single vector aligned with the direction where most of the function energy lies, thus the input function is paired with a *Linear Structure Field* (LSF) embedding into each point information about straight lines drawn by the function. If the input for these processes is the magnitude of an $STFT$ representation of an audio file, this technique provides information about the spectro-temporal organization of sound such as described in Sections 2.3.2, 3.3, 3.4, and 3.5, and can be referred as the *Spectro-Temporal Structure Field* (STSF).

[1] At a first sight, the same information contained in a STSF may seem similar to the detection of collinearity of $|STFT|$ peaks, which can be probably computed more efficiently. Nevertheless STSF is thought to produce a non-sparse (and theoretically continuous) output.

5.2 Definition

5.2.1 Signal Energy Angular Distribution

First a definition of some useful functions is called for. Let $P = (x_P, y_P)$ be a point inside the domain of an $\mathbb{R}^2$ function $f(x, y)$, and let $L(P, \alpha)$ be a family of straight lines, all intersecting in P with angle α:

$$L(P, \alpha) : (x - x_P) \sin \alpha = (y - y_P) \cos \alpha \tag{5.2.1}$$

then let $W_P(x, y)$ be an isotropic weighting function that masks $f(x, y)$ in some neighbourhood of P, but assigning a null weight to the point itself (the Gaussian kernel of this example is just one of the many weighting functions that may be used):

$$W_P(x, y) = \begin{cases} 0 & \text{if } (x, y) = P \\ \dfrac{1}{2\pi\sigma^2} e^{-\frac{(x-x_P)^2+(y-y_P)^2}{2\sigma^2}} & \text{otherwise} \end{cases} \tag{5.2.2}$$

The SEAD $\mathcal{D}\{f, P, W_P\}$ of the point P of a given function $f(x, y)$ (such as a picture) is a new function $\mathring{P}(\alpha)$, that is a line integral that scans all the straight lines $L(P, \alpha)$ and for each α outputs a vector whose magnitude is the sum of all the points of $f(x, y) \cdot W_P(x, y)$ along $L(P, \alpha)$ and whose angle is simply α:

$$\mathring{P}(\alpha) = \mathcal{D}\{f, P, W_P\} = \int_{L(P,\alpha)} f(x, y) \cdot W_P(x, y) \cdot e^{i\alpha} \, ds \tag{5.2.3}$$

since $L(P, \alpha) \equiv L(P, \alpha + k\pi)$, then $\mathring{P}(\alpha)$ can be restricted to the domain:

$$-\frac{\pi}{2} \leq \alpha < +\frac{\pi}{2} \tag{5.2.4}$$

An illustration of SEAD can be found in Fig. 5.1 and 5.2. The pictures show the accumulation around P for each angle α. In a sense, SEAD can be considered as a *local* or *introspective* Radon transform, since it endows all projections of $f(x, y)$ on a single point of itself.

5.2.2 Linear Structure Field

Having a complex function for each point of $f(x, y)$ may be redundant, so a strategy to reduce the information of SEAD into a field (one vector for each point of $f(x, y)$) is called for.

The main desideratum is to obtain a *Linear Structure (vector) Field* (LSF) such that if P is part of a linear pattern it should be paired with a vector $\hat{f}(P)$ that reflects

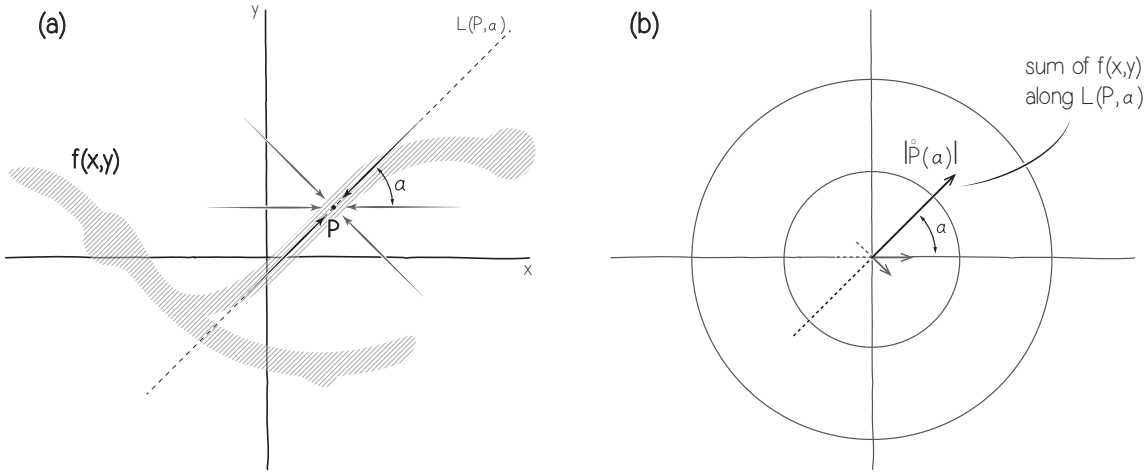


Figure 5.1: Input function $f(x, y)$ on the left, and SEAD $\mathring{P}(\alpha)$ of one point $P \in f(x, y)$ on the right (only 3 angles have been calculated in this example). The value of SEAD is equal to the integration of the input function over the line that passes through P with angle α.

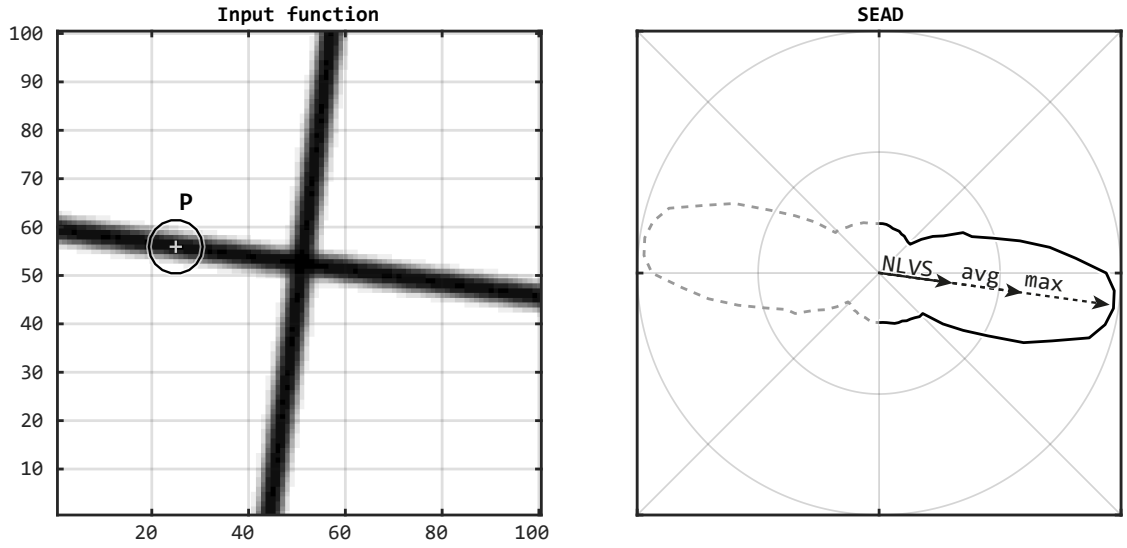
the angle of that pattern. Moreover $\left|\hat{f}(P)\right|$ should be proportional to the certainty of being over a single pattern. Finally, values of the vector where the function is zero are not of interest, while points of the function that form curved patterns should be paired with vectors at a tangent to that pattern.

In principle, by taking the α that maximizes $\mathring{P}(\alpha)$, the leading direction should be selected, and this may work for P in Fig. 5.2a. Nevertheless, the situation depicted in Fig. 5.2b shows how this approach may fail in recognizing ambiguous conditions, the same could be said for the average $\mathring{P}(\alpha)$, which returns a strong vector pointing where there is no energy accumulation of $f(x, y)$.

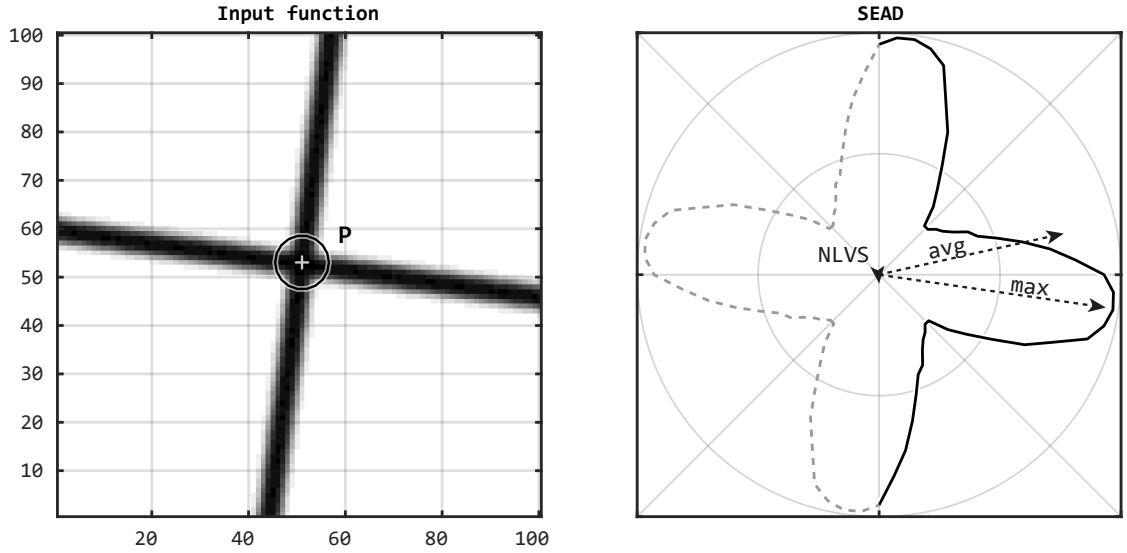
The main problem that arises when summing SEAD vectors, is that orthogonal components should cancel out each other instead of deviating the result. Similarly vectors with opposite direction should sum up instead of cancelling out, since their direction does not matter as long as they lay on the same straight line.

Therefore, to turn $\mathring{P}(\alpha)$ into a proper LSF, a method called *non-linear vector summation* (NLVS) is proposed.

NLVS consists in doubling the angle of all vectors before summing them up, then the angle of the result is halved to bring it back in the original domain (for better



(a) P is lying on a linear pattern.



(b) P is lying on the intersection of two linear patterns.

Figure 5.2: Input function $f(x, y)$ on the left, and SEAD $\mathring{P}(\alpha)$ on the right. All points of SEAD are the tip of vectors with angle α and magnitude $\mathring{P}(\alpha)$. A proper way to sum them up is called for.

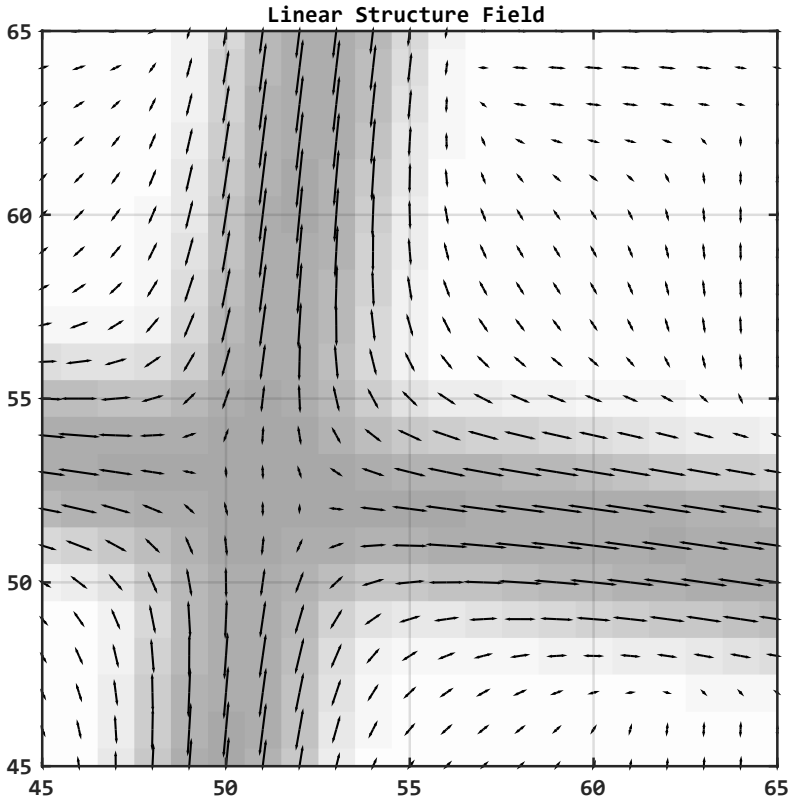


Figure 5.3: A detail of the LSF plotted on-top of the original function, vectors point toward the direction of the linear pattern where they lay.

readability an auxiliary variable ς is used):

$$\varsigma(P) = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} \left| \mathring{P}(\alpha) \right| \cdot e^{i2\alpha} \, d\alpha \tag{5.2.5a}$$

$$\hat{f}(P) = |\varsigma(P)| \cdot e^{i\frac{1}{2}\angle\varsigma(P)} \tag{5.2.5b}$$

The LSF of a detail of the function shown in Fig. 5.2 can be seen in Fig. 5.3

The main difference with the function gradient or Sobel operators is that the LSF does not work only on the *edges* of the function, but instead it can work out linear patterns also when the gradient is zero, i.e. in those points where the function is constant. The coverage area (that is the locally flat area of a function where LSF is still able to recognize the presence of a pattern) is proportional to the size of the W_P function, while the precision of the answer (i.e. how reliable are the pointed patterns) has an inverse proportion to the same parameter, since increasing the window size can make SEAD consider also patterns which are actually far away from P. In a

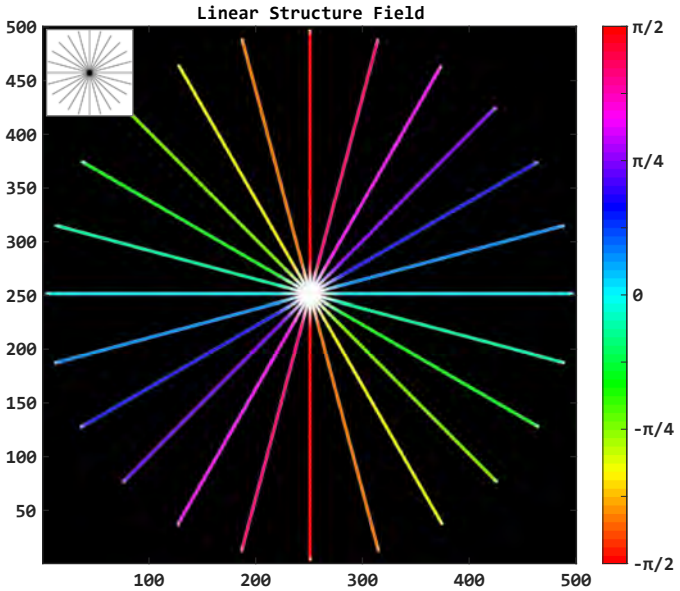


Figure 5.4: Thanks to LSF, lines with different angles are shown in different colours, except for the intersection area, where vectors have minimum magnitude (input image is shown on top left corner). This image can also be used as colour legend.

discrete context, LSF quality also depends on the sampling resolution of $\mathring{P}(\alpha)$ on the α axis. these parameters will be explored in Section 5.4.

Finally, with respect to the BS, a way to visualize LSF may be to encode $\angle\hat{f}(P)$ with hue, $\left|\hat{f}(P)\right|$ with saturation and $f(P)$ with brightness, as shown in Fig. 5.4 and 5.5.

5.2.3 Spectro Temporal Structure Field

If $f(x, y)$ is the $|STFT|$ of a signal $x(t)$, then $\hat{f}(x, y)$ could be called *Spectro-Temporal structure field* (STSF), and can be written as $\hat{X}(t, f)$, where t is time and f is frequency. An example is shown in 5.6: amplitude is represented in decibels and then rescaled to $0 \ldots 1$ values before taking the $\hat{X}$ (f and t arguments will be omitted for brevity). The STSF is computed on a linear-frequency representation of the signal, and then displayed in log-scale. Note that in STFT, linear patterns correspond to a frequency gliding in time, thus $\angle\hat{X}$ can be converted to modulation speed $v(P)$, that is the Hz per second of continuous frequency changes:

$$v_{hz/s}(P) = \frac{\Delta f}{\Delta t} \tan(\angle\hat{X}(t, f)) = \frac{\Delta f}{\Delta t} \cdot \frac{\Im\hat{X}}{\Re\hat{X}} \tag{5.2.6}$$

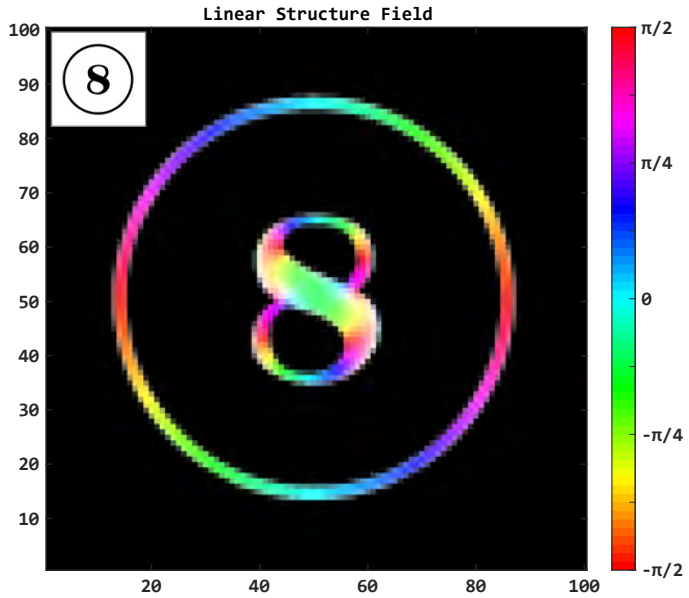


Figure 5.5: An 8 surrounded by a circle. Colours depict the tangents of the curves (as in Fig. 5.4, input image is shown on top left corner).

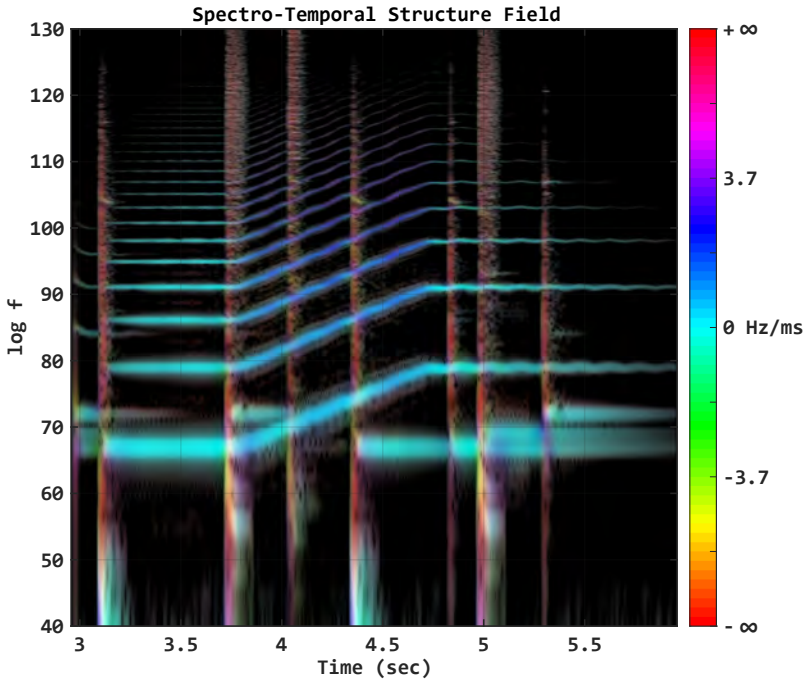


Figure 5.6: STSF of a segment of the same signal of Fig. 3.6a. Colour encodes the angle of lines in a time-frequency space, thus the speed of changes in frequency

where Δt is the time resolution of the STFT (i.e. the hop size), while Δf is the frequency resolution. Unfortunately, as it can be seen from Fig. 5.6, speed expressed in Hz/sec colours different harmonics of the same sound differently, for this reason a logarithmic measure may be preferable.

Once Δn is defined as the frequency resolution expressed in semitones:

$$\Delta n(f) = 12 \log_2 \frac{f + \Delta f}{f} \tag{5.2.7}$$

Speed information can be represented in log scale as semitones per second, thus resolving the issue that linear speed retain for harmonic signals:

$$v_{n/s}(P) = \frac{\Delta n(f)}{\Delta t} \cdot \frac{\Im \hat{X}}{\Re \hat{X}} \tag{5.2.8}$$

This operation avoids the computation of $\hat{X}(t, f)$ in the log frequency domain, a conversion which from a computational perspective calls a trade off between memory usage and proper high frequencies resolution.

As shown in Fig. 5.7, it is possible to see the distribution of log-speed to check for accumulation around certain points, corresponding to the concept of *common fate* used in CASA. Note that, as well as the one realized for the BMS, the distribution is weighted by $\left|\hat{X}\right| \cdot |X|$, to account only for relevant STFT bins.

5.3 Properties and Applications

Now some properties and applications of the discussed techniques will be investigated. Examples and applications of LSF and STSF can be found in the GitHub repository.

5.3.1 Radon as a Distribution

By definition $\hat{f}$ is a set of straight lines passing through P with angle $\angle \hat{f}$. Being able to convert those lines in to the same form used by the Radon transform (i.e. coefficients p and ϕ of Eq. 2.1.16) enables the ability to approximate the Radon transform by plotting the joint distribution of these parameters, as shown in Fig. 5.8.

ϕ is actually $\angle \hat{f}$, while distance from the centre p can be calculated using ϕ and P coordinates:

$$p = \sqrt{x_P^2 + y_P^2} \cdot \sin\left(\alpha - \arctan \frac{y_P}{x_P}\right) \tag{5.3.1}$$

Again, the distribution is weighted by $\left|\hat{f}\right| \cdot |f|$, to account only for relevant vectors.

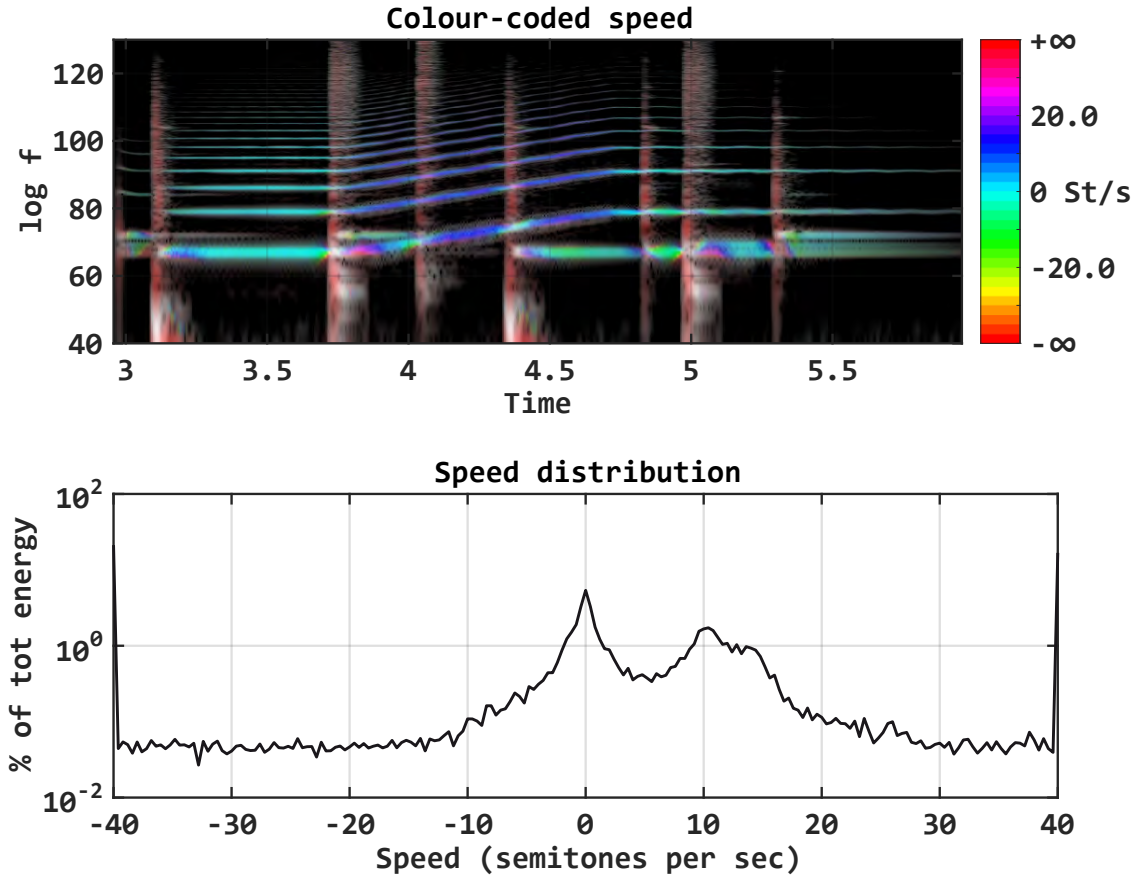


Figure 5.7: Same signal of Fig. 5.6, but coloured by log-speed (semitones per second). The distribution shows peaks for $\pm\infty$ (percussive sounds), 0 (pitched sounds), and 10 (the gliding sound).

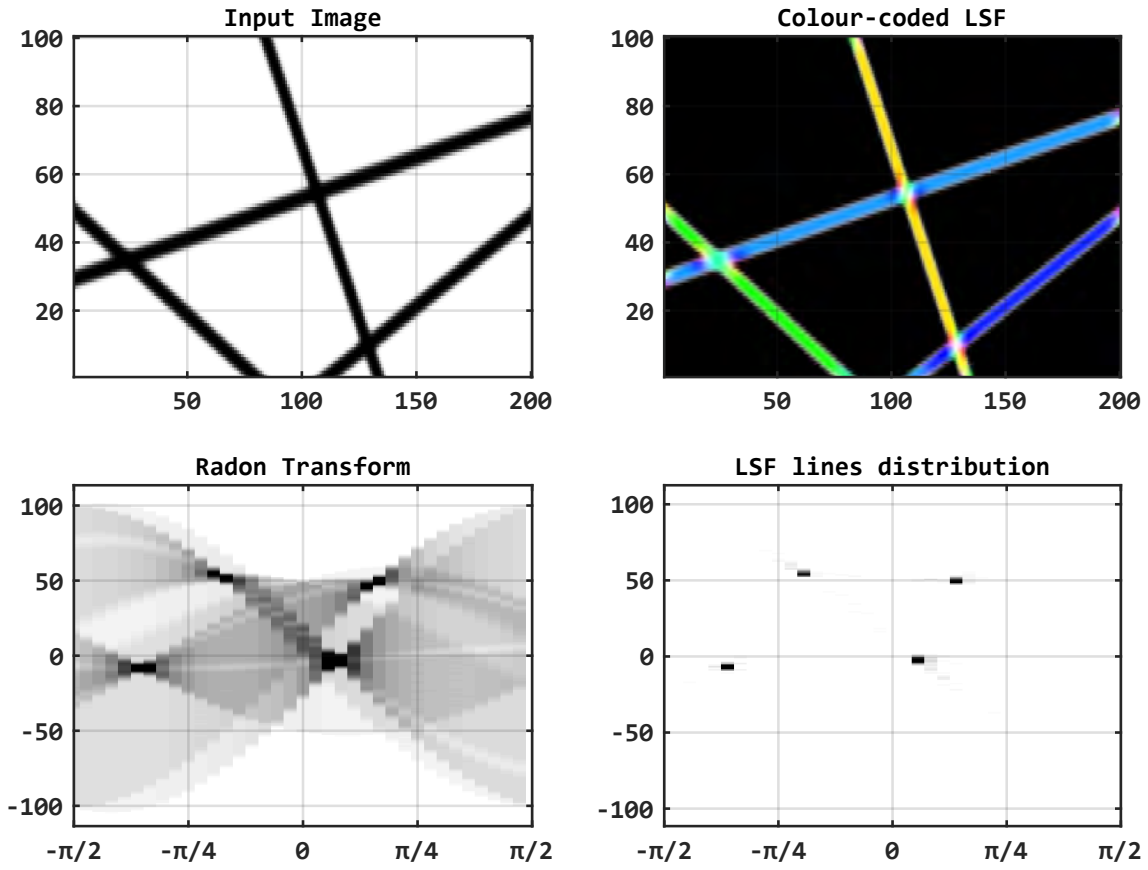


Figure 5.8: Top: an input function and its LSF. Bottom: the Radon transform of the function and the joint distribution of a transformation of LSF vectors

Grey areas visible in the Radon transform shown in Fig. 5.8 corresponds to the information packed in $\mathring{P}$, which is discarded in $\hat{f}$. In principle from $\mathring{P}$ it should be possible to reconstruct a complete Radon transform, but this possibility will not be explored since the mere approximation of its peaks is usually sufficient.

This concept, together with the speed distribution, demonstrates how useful STSF representation can be in context such as those described in Section 3.5, where the $STFT$ of a signal is Radon-transformed to gather some features about the underlying audio signal.

5.3.2 Anisotropic Masking

In the audio processing context, segregation between percussive and pitched material may be useful in many situations as described in Chapter 3. Thanks to the STSF it

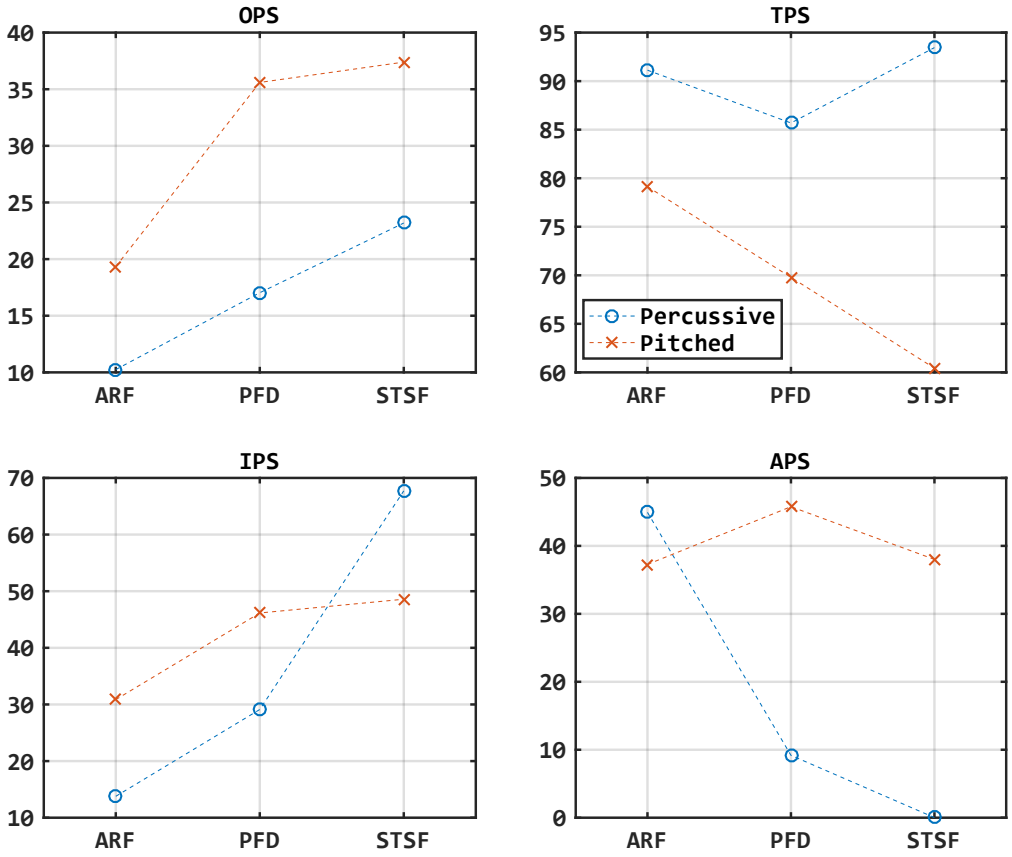


Figure 5.9: PEASS results for 3 algorithms of pitched v.s. percussive audio separation. Input sources are the same of Fig. 3.6a, grouped by type (i.e. percussive or pitched).

is possible to create masks based on spectral features which not only helps in separating transients from periodic sounds, but can also enable the possibility of selecting portions of sound where frequency changes at a specific St/sec rate. Moreover masks parameter can vary in time, which is not possible when using methods such as those described in Section 3.4.

To compare the separation capabilities of the STSF, a PEASS test has been run, comparing the results provided by 3 algorithms: The Auditory Receptive Field based separation (ARF) [51, 52], the Percussive Feature based (PFD) [63], and the STSF based. Results are shown in Fig. 5.9. It is hard to properly compare the results of the 3 algorithms, since they do slightly different things, but in general it seems that STSF may provide similar outcomes with respect to the other strategies. One thing that the PEASS framework fails to highlight is that STFT is the only one that completely rejects the glissando signal from the extraction of the percussive signal (more tests

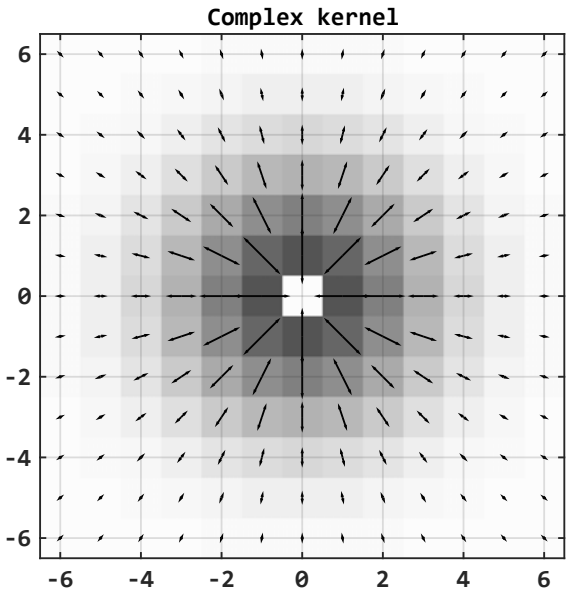


Figure 5.10: Complex kernel for the representation of the LSF as a *quasi-convolutive* operation

can be found in Section 6.1). Execution time are 1.57 seconds for ARF, 0.16 for PFD and 0.15 for STSF, but STSF time only accounts for the masking time. It does not take into account the computation of the STSF, since once it is computed it can be used for any masking instance. Unfortunately it took about 77 seconds to compute the STSF, an issue that will be discussed in the next section.

5.4 Optimization

The computation of $\hat{f}$ (and especially $\mathring{P}$) may be very slow due to the great number of integrals calculated for each point of the function, especially when high α resolution is required. Moreover, in a discrete context, pixels close to the point may be overweighted due to the aliasing of the sampling line $L(P, \alpha)$.

Looking at Eq. 5.2.3, Eq. 5.2.5b, and Fig. 5.1, it may be noticed that, except for the non-linear summation, those operations are similar to a complex convolution of $\hat{f} \approx f * h$, with h being a complex function depending on W_P, and which sums are calculated by scanning the function *radially*. In particular $h(x, y)$ can be written as:

$$h(x, y) = W_0(x, y) \cdot e^{i \tan \frac{y}{x}} \tag{5.4.1}$$

where W_0 is W_P for $P = (0, 0)$. This kernel h is depicted in Fig. 5.10.

Now, in a discrete domain, a *convolution-like* form could be very useful since it avoids aliasing issues, it computes $\hat{f}$ without explicitly finding each $\mathring{P}$ (thus avoiding the sampling operation over α).

For these reasons, to express the calculation of $\hat{f}$ in a simpler way, it is useful to introduce a custom operator called *NLV-Convolution* $f \circledast h$, which replaces the implicit summations of the convolution with the NLVS used in Eq. 5.2.5, that is defined as the sum of two complex numbers $a \oplus b$:

$$\varsigma = |a|\, e^{i2\angle a} + |b|\, e^{i2\angle b} \tag{5.4.2a}$$

$$a \oplus b := |\varsigma| \cdot e^{i\frac{1}{2}\angle\varsigma} \tag{5.4.2b}$$

such that $\hat{f}$ can be calculated as:

$$\hat{f} = f \circledast h \tag{5.4.3}$$

In this form, STSF can be seen as an approximated pre-computation of many Auditory Receptive Fields, which response can be retrieved by masking the STSF, based on the desired ARF angle (consequently, LSF can be seen as a pre-computation of many Gabor filters).

Fig. 5.11 compares average execution time and overall accuracy for the two implementations (with different resolutions of α for the canonical implementation). The test was run upon a random set of synthetic pictures of 100×100 pixels generated automatically together with theoretical ground truth gr. Error was measured as absolute difference between result pixels and gr pixels.

The test shows how, for W_P size set to 12, the computational cost is comparable for a $\mathring{P}$ computed on 8 equally spaced α angles, while accuracy becomes comparable from a resolution of 9 equally spaced α angles. In conclusion, it seems that the canonical method is preferable only when expected α angles are known a priori or when low α resolution is required, so that the function can be sampled only on a small number of angles, otherwise the NLV-Convolution method is more reliable.

Finally,a profiling of the test executed on a 100×100 image with 8 angles resolution, underlines how the highest computational cost comes from: the line sampling for the canonical implementation; the NLVS for the NLV-Convolution implementation; and from the arctangent function, which is used in both implementations. In particular the canonical approach calls $atan2(\cdot)$ 10000 times, and the line sampling function is called 44264 times, while in the NLV-Convolution method no line sampling function is called, but $atan2(\cdot)$ is called 30000 times: 3 times for each of the 10000 calls to the NLVS function. This means that a fast $atan2(\cdot)$ approximation should drastically improve performance, especially for the NLV-Convolution method, while a fast way to sample a function over a straight line should also improve the canonical implementation.

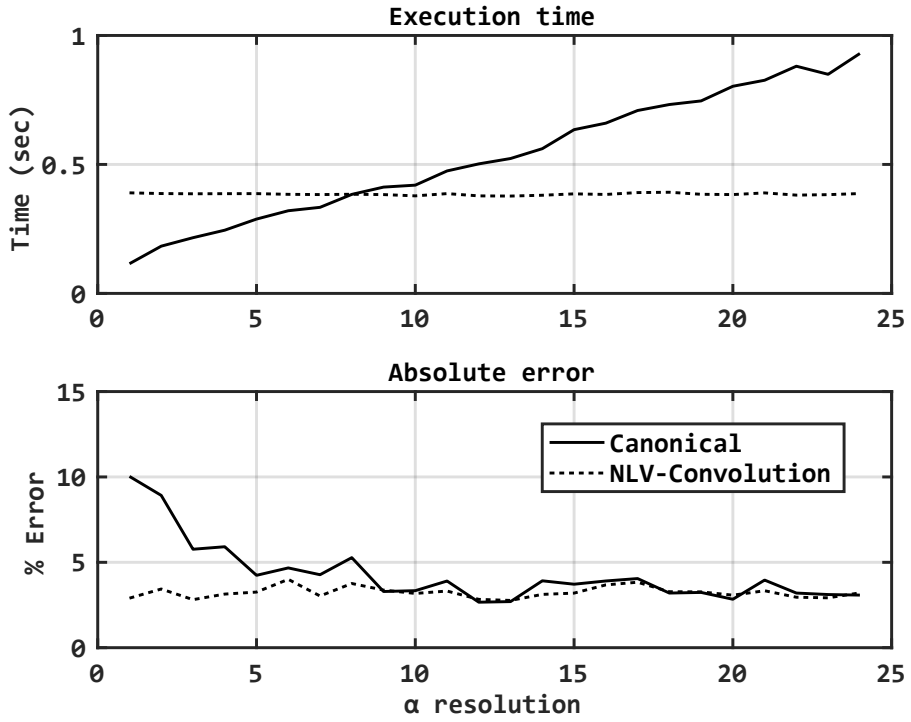


Figure 5.11: Test results for different computing methods. Finer approximations can be achieved by choosing the right parameters for W_P function

Chapter 6

Combined Methods Use Cases

6.1 Source Separation

6.1.1 Experimental Setup

To evaluate the usability of proposed techniques in the real world, a typical SMC task will be explored, that is the separation of vocal signal within a complex mixture. In particular the MASS database [44] will be used as dataset, since it contains real world samples, thus providing a realistic scenario (details about the dataset are provided in Section 2.3.4). In case of instrumental songs, an alternative pitched instrument has been separated, such as bass or guitar.

The proposed architecture is composed of a first *masking and resampling* operation in the BMS, followed by a rejection of percussive sounds, realized by attenuating components with almost-vertical STSF (percussive signals, such as drums, are frequently found in the same azimuth position of vocals [57, 63]).

As a baseline, ADRess algorithm will be used to isolate the azimuth portion containing lead vocals, followed by PFD for rejecting percussive signals. Moreover, also ARF-based separation is performed over ADRess output as an alternative to PFD as second stage baseline.

Both baseline and proposed techniques are enhanced by a final hi-pass filter that should remove any component under 100 Hz. In the case where a non-vocal signal is targeted, the cut-off frequency is modified according to the lowest note playable by the instrument.

All algorithms have been implemented from scratch in the Matlab environment, and are available for download in the repository hosted on GitHub. The *azimuth discrimination* part of ADRess has been realized by computing BS angles, in such a way that component position in the azimuth plane are expressed in the same measurement units of the proposed technique. PFD has been implemented by meticulously

following the original definition [63]. ARF has been implemented as a set of convolutions with different kernels in the log-frequency domain, using bin numbers (instead of Semitones and Seconds) as measurement units. Finally, BMS and STSF separation algorithms have been implemented with no exploitation of speed conversion or other metrics of the proposed spaces, since a fine source separation is not the aim of this work. Indeed, a basic implementation is tested as a *working space* alternative to the ones of the baseline.

The *STFT* is performed on L1-normalized input, divided into frames of 2048 samples, a hop size of 512 samples, and with input and output Hann windowing. ADRess and BMS mask parameters are set to equivalent values, manually selected for each song, while PFD, ARF, and STSF mask parameters are set to fixed values, that aim at isolating pitched signals.

6.1.2 Results and Discussion

The results of each algorithm have been analysed with PEASS software, whose output is shown in Fig. 6.1, and whose raw results are reported in Appendix B. In particular, 5 techniques are visible: ADRess and BMS labels are relative to the output of the first stage, while +PFD, +ARF, and +STSF are relative to the three final outputs. Finally, since every song behaved very differently, PEASS results have been normalized by song.

ADRess and BMS performs very similar, with no significant differences in any of the scores. On the second stage of the processes, PFD, ARF and STSF performs quite differently: PFD seems to sacrifice part of the target sound and the presence of artefacts in order to achieve a better separation from other sources, STFT discards significantly less target signal and it is milder in terms of separation, but is able to improve the signal to artefact ratio, resulting in a more natural sound even in presence of artefacts coming from the previous stage. Finally, ARF seems to mediate between PFD and STFT results, significantly improving the overall perceptual score. One thing that does not emerges from the PEASS scores is that STSF is more linear in the frequency response, in opposition to ARF and PFD which tend to *darken* the sound, rejecting a consistent amount of high frequencies.

The fact that the results are dependent on the analysed song might indicate that the proposed methods works better for some classes of signals. For what concerns PFD, ARF and STFT, analysing algorithms scores in detail reveals how ARF consistently gets the best OPS, PFD gets the best IPS and STFT gets the best TPS and APS, thus confirming the previous observations.

Dependency on the song kicks in when talking about ADRess and BMS. In musical genres such as pop, rock (and some hip-hop songs) chorus effects and unison instruments are frequently used. These effects can spread a source all-over the stereo

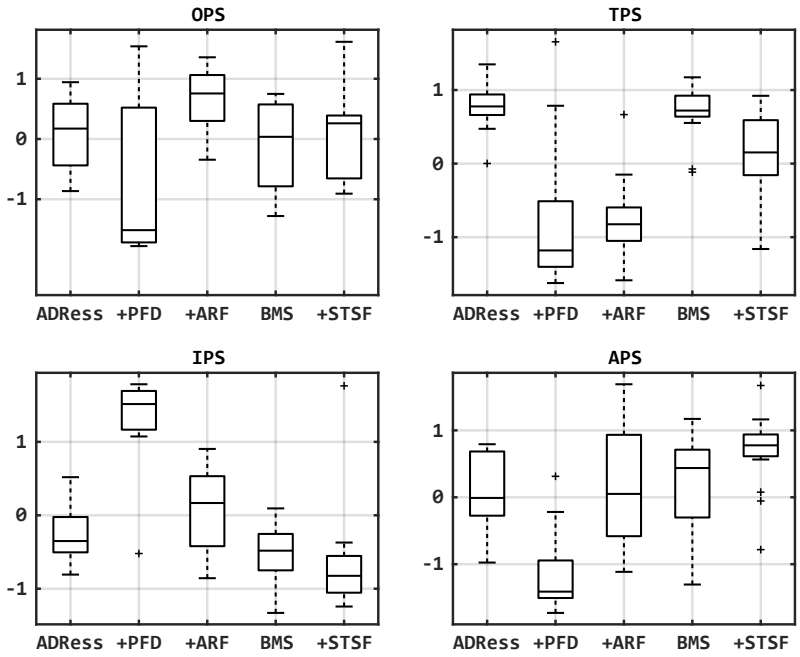


Figure 6.1: Normalized PEASS results for a voice extraction task performed on the MASS dataset. Raw PEASS scores can be found in Appendix B

image and can decorrelate stereo channels. When this happens, the differences between ADRess and BMS get stressed. ADRess ignores everything is outside the target mask, and this may include uncorrelated components of the target source, while BMS can also consider uncorrelated elements as *ubiquitous* in the stereo image, including those which do not belong to the target source. It is easy to see how these characteristics may be either positive or negative depending on the circumstances, but in general test data support the idea that ADRess is slightly biased towards the minimization of interferences, while BMS minimizes artefacts and target loss. Data also show that, in case sources are not spread by the aforesaid effects, ADRess and BMS behaves very similarly, with almost identical scores.

In general, data show that the OPS of any algorithm drastically drops when the target is not a vocal signal, or when a metal song is analysed: in both situations the scores that drag OPS down are TPS and APS. On the other hand, as expected, best case scenarios are those where a minimalistic set of instruments is played (such as in bossanova or reggae music), or where loud leading vocals are mixed in the centre of the stereo image, such as in some hip-hop songs.

Given the aforesaid results, and considering that the proposed techniques can be parametrized far more than the baseline approaches, the proposed spaces can be seen as promising helpers in the field of source separation and other SMC tasks.

6.2 Information Visualization

BS-enhanced spectrograms and colour-coded STSF can be effective ways to visually represent a variety of spectro-temporal information. But in case they are used simultaneously to seek for patterns inside the $|STFT|$, the task of looking at two pictures at the same time may be hard to accomplish. In such a situation, what is relevant is not the actual value of the BS or STSF, but some kind of *distance* measure among bins properties, ideally packed into a single image. These premises seem to point towards Multidimensional Scaling, but at the cost of a high computational expense. What follows is instead an information processing and visualization technique based on PCA, where data dimensionality is reduced preserving Euclidean distance among bins properties, and then represented into a colour space able to express most of this distance information.

Relational BS information R can be seen in polar coordinates as described in eq. 4.2.6, but bases at $\sigma = 0$ are actually close to those at $\sigma = \pi$ due to the periodicity of eq. 4.2.2. Moreover, bases with low C values should be considered very close to each other since, for those bases, angle σ loses its meaning. These proximity features can be rendered in Cartesian coordinates by:

$$\begin{aligned} x_a &= |C| \cos(2\sigma) \\ y_a &= |C| \sin(2\sigma) \end{aligned} \tag{6.2.1}$$

The same can be said for the bases in STSF, where bases at $\angle\hat{f} = \frac{\pi}{2}$ are close to those at $\angle\hat{f} = -\frac{\pi}{2}$, since they are both vertical. Furthermore, also bases with low $\left|\hat{f}\right|$ should be considered similar, since they share the property of being upon no linear patterns. Again, in Cartesian coordinates this translates to:

$$\begin{aligned} x_b &= \left|\hat{f}\right| \cos(2\angle\hat{f}) \\ y_b &= \left|\hat{f}\right| \sin(2\angle\hat{f}) \end{aligned} \tag{6.2.2}$$

Now suppose to take the STSF of the PSC $\left|\overline{X}\right|$, which is notated as $\hat{\boxtimes}$, and to augment the aforesaid Cartesian representations by also considering bases level as a proximity feature. The signal is now projected on a four-dimensional space O:

$$\begin{aligned} x &= |C| \cos(2\sigma) \\ y &= |C| \sin(2\sigma) \\ u &= \left|\hat{\boxtimes}\right| \cos(2\angle\hat{\boxtimes}) \\ v &= \left|\hat{\boxtimes}\right| \sin(2\angle\hat{\boxtimes}) \end{aligned} \tag{6.2.3a}$$

$$O = \{x,\, y,\, u,\, v\} \cdot \left|\overline{X}\right| \tag{6.2.3b}$$

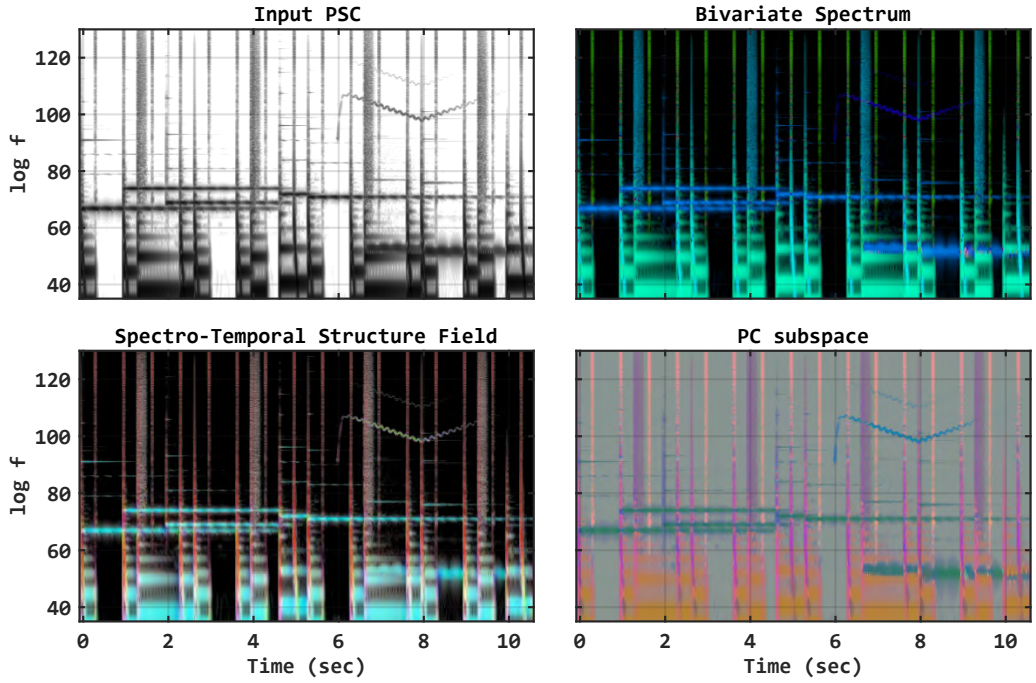


Figure 6.2: In the PC subspace only differences in colour are relevant, rather then the absolute colour value.

If two bases are close to each other in this space, it means they share both mixing properties (stereo position in an SMC scenario) and spectro-temporal properties (percussive or pitched sounds in an SMC scenario). On the other hand, in this space, bases with similar mixing properties but different spectro-temporal features are distant from each other.

Unfortunately, this space is not useful to approximate source separation, since the proper way to consider STSF in a bivariate context is to take it on a precise sampling of $\tilde{X}(\alpha)$ rather than the messy $\overline{X}$. Nevertheless, this space can be used to visually differentiate mixture information.

Let PC_{1-4} be the 4 principal components of O; a simplification of O called *Principal Component subspace* Ω can be defined as:

$$\Omega = \{PC_1,\ PC_2,\ PC_3\} \tag{6.2.4}$$

This three-dimensional space can be mapped onto RGB colours simply by scaling its values in the range [0..1] (that is a saturation scaling between [0..1]). An example is shown in Fig. 6.2.

	Most comfortable	Least comfortable
PCss	5	1
STSF	2	1
BS	2	3
PSC	1	5

Table 6.1: Users visualization preferences for the source-counting task

The proposed design is based on the idea that similar input properties will produce similar output colours, which the brain will cluster into consistent regions only in case of moderate differences in the original space. Moreover, cluster colours will reflect cluster distances in terms of BMS and STSF similarity.

To validate this approach, 10 music production experts were asked to count how many instruments they could spot by looking at different graphical representations of 4 music excerpts. The same 4 representations of Fig. 6.2 were rendered for each sample, so that, in total, each expert had to count sources in 16 different pictures, presented in random order. Experts were also asked if they suffer from some colour deficiency, and which representation they found to be the most and the least fitting for the proposed task.

Pictures were shown on a 21′ calibrated desktop-pc monitor. A *Google Form* version of the test can be found at the URL: `https://goo.gl/XZx3DS`. Raw data results can be found in Appendix B.

The music excerpts were chosen so that the number of sources was clearly distinguishable by listening to them[1]. In order of increasing mix complexity, the excerpts were taken from:

- The Beatles – *When I'm Sixty-Four* (4 sources);
- John Coltrane – *In a Sentimental Mood* (5 sources)
- A synthesized example with no stereo effects (8 sources);
- Amy Winehouse – *Back to Black* (8 sources).

Unfortunately the standard deviation of the error data shown in Fig. 6.3 is too large to draw definitive conclusions, suggesting that more investigation is needed. Nevertheless, some preliminary observations can be done.

First, by analysing the same picture, it seems that counting sources only by looking at the sonograms always leads to an underestimation of the total number of instruments. Nevertheless, on average Principal Component subspace (PCss) performed

[1]A source is defined as a percussive or pitched instrument in a particular stereo image position

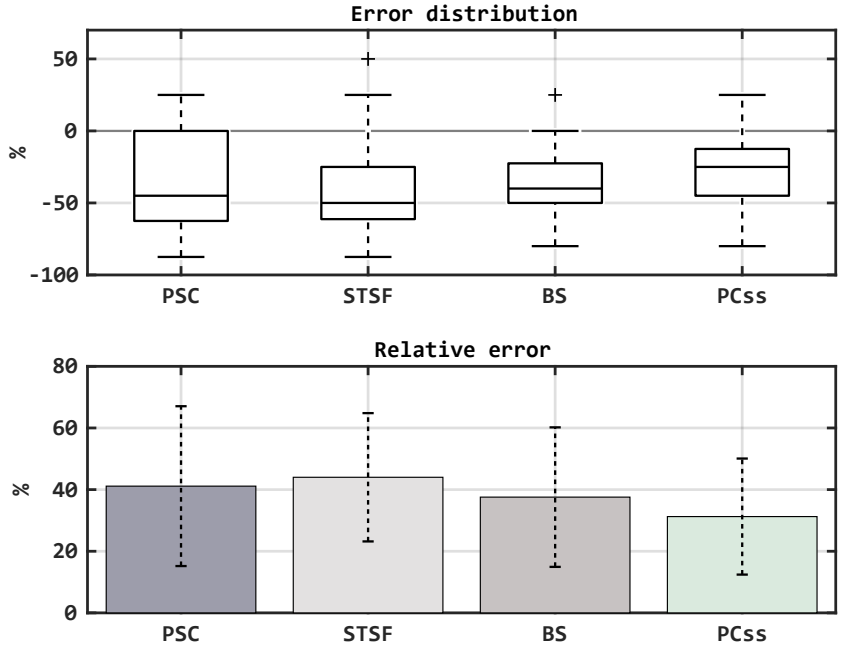


Figure 6.3: Error distribution and mean of relative errors for a source-counting task, grouped by visualization type. Raw data can be found in Appendix B

better than the other visualization strategies, presenting the lowest mean error together with the smallest standard deviation.

In detail, Fig. 6.4 shows how PCss significantly improves results only for those excerpts made of 5 sources or more, while there is no significant improvement in case of the simpler excerpt, where colours seem to be misleading, and a classic sonogram may be enough to accomplish the requested task.

Finally, as shown in Table 6.1, users reported a moderate preference for the PCss when asked to tell which visualization strategy they felt more comfortable for the proposed task, while it was among the least mentioned when asked to tell which was the least adequate visualization strategy. Interestingly, even if BS was among the most disliked, it ranked second in terms of error.

These results suggests that PCss may be an interesting design, but more tests are needed to properly validate the proposed approach.

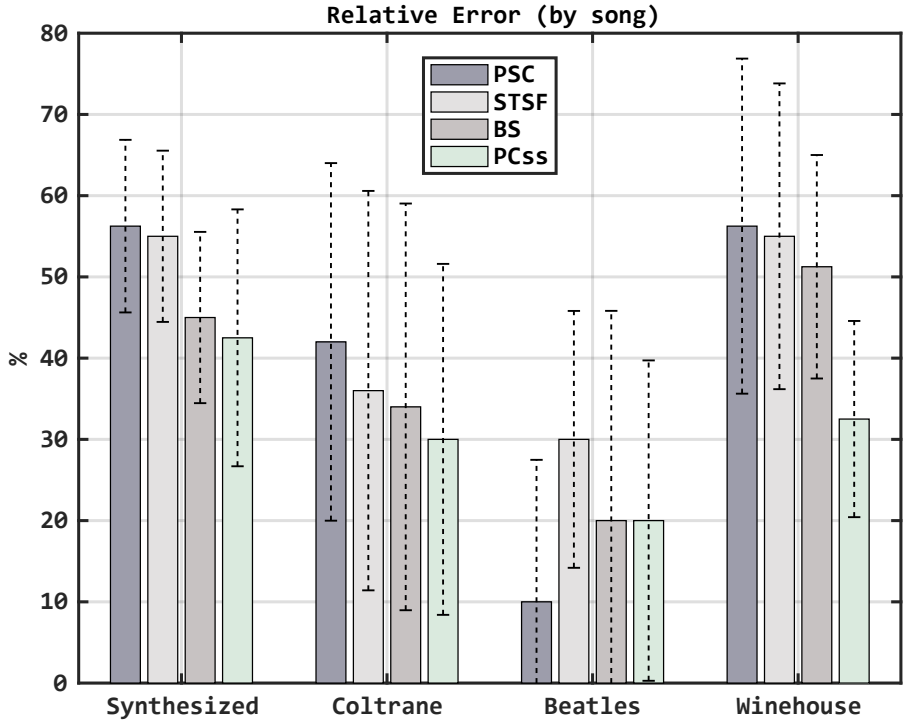


Figure 6.4: Mean of relative errors, grouped by song and visualization type. Raw data can be found in Appendix B

Chapter 7

Final Remarks

7.1 Conclusions

Two new techniques have been presented, one aimed at better representing bivariate spectral information in an effective way (as an answer to $\boldsymbol{Q1}$), and the other aimed at folding spectro-temporal information onto a single point of a STFT representation of a signal (as an answer to $\boldsymbol{Q2}$).

The first technique is based on the idea of Bivariate Mixture Space, that is an interpolation of two mixtures in the frequency domain. From this auxiliary space different representations of the signal can be rendered, such as the decomposition into Principal Spectral Content and Relational Content. Eventually, some of the techniques present in literature can be generalized by manipulations of the BMS, like masking and resampling, or by reading bivariate spectrum statistics.

The second technique is based on the idea of creating a Structure Field onto a function, that is a vector field where each vector is aligned with linear patterns created by the function itself. This technique is useful in image processing contexts whenever the Radon transform is called for, but it can also be applied to $|STFT|$ to obtain a Spectro-Temporal Structure Field, useful to track frequency modulations or transient presence. Again, this space can be considered a generalization of other techniques, such as an approximated pre-computation of any possible auditory receptive field output.

Both techniques can be combined in the context of SMC to perform source separation tasks, or more in general, they can be used in signal processing context as valid signal representation spaces, which in turn can be visualized very efficiently.

A Matlab implementation of these techniques is provided as a GitHub repository at `https://github.com/Kuig/LIM-Toolbox`, together with instructions to easily produce custom audio examples. The very same implementation has been used to

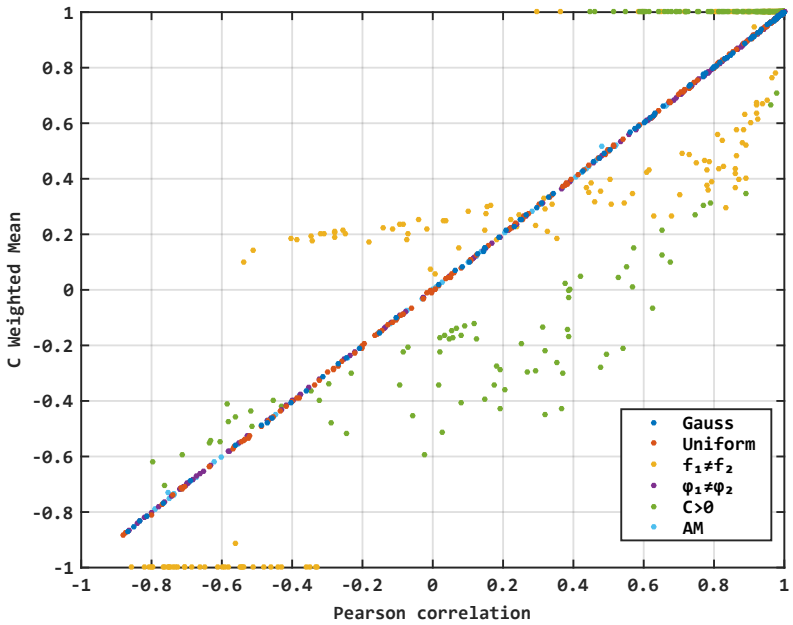


Figure 7.1: $\overline{C}$ v.s. Pearson Correlation for different classes of sounds. Zero-frequency and silent components have been excluded from the computation of $\overline{C}$, since correlation should not be defined for those values.

provide an experimental demonstration of how this proposal can compete with state-of-the-art signal processing algorithms such as Azimuth Discrimination and Resynthesis, Percussive Feature Detection, and Auditory Receptive Fields. The test-bed of the validation process consisted in a PEASS comparison of a vocal separation task performed on the MASS dataset, which demonstrated the validity of the proposed approach.

7.2 Future Works

Many aspects of the BMS should be explored in the future. For example, among the many $\vec{X}$ statistics that can be used as new features, it has been noticed that the weighted mean of C values $\overline{C}$ is a good approximation of linear correlation, as can be seen from fig.7.1, providing results within $\pm 0.016\%$ for most of the signals (Gaussian and uniform distributed noises, periodic signals with different random phase offsets and ring-modulated signals), except for 2 classes of signals, where error is within $\pm 5.69\%$ (mixtures of different frequencies and mixtures with only positively correlated components). In other words, $\overline{C}$ can be considered as a new measure of correlation based on spectral properties which deserves further attentions.

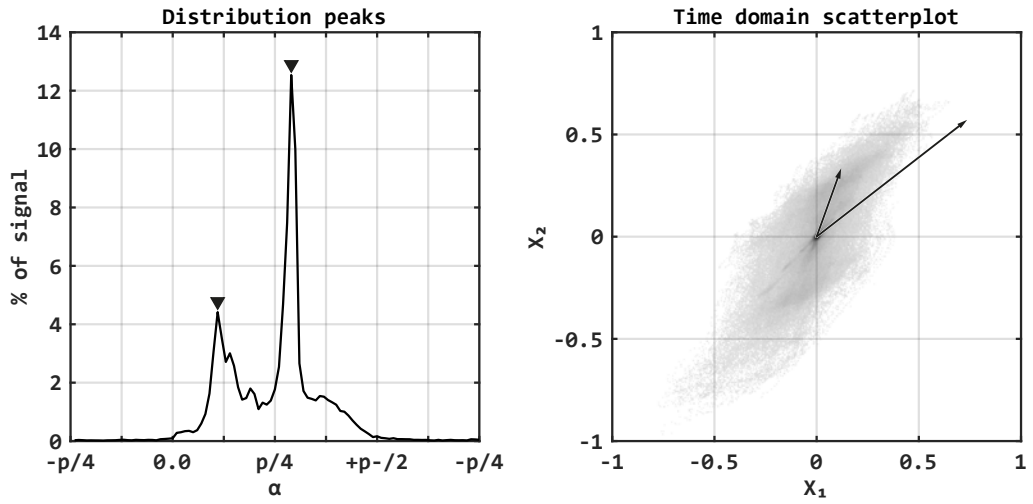


Figure 7.2: Distribution peaks are hints to independent components of the signal

Another interesting observation arises when plotting σ distribution peaks over a time domain scatterplot of the input mixtures. Those plots can give hints about the presence of independent components in the mixtures, which in turn can be approximated as the BMS signal present at σ distribution peaks, as shown in Fig. 7.2. Thus, a possible separation technique could exploit ICA for *cleaning up* a masked portion of the BMS.

In case of joint distributions such as those in Fig. 4.5a and 4.5b, it should be noticed how those matrices are actually an approximated non-negative decomposition of $\left|\overline{X}\right|$, as shown in Fig. 7.3. This is another aspect that deserves some attention in the future.

Finally, since the original form of BMS is blind to components that are distributed in the mixture space with delay operators, the computation of σ distributions can be extended to the lag-domain, by repeating the transformations for different offset of the input observations, preferring those offsets where a peak in the cross-correlation of the observations occurs.

Regarding SEAD and LSF, a detailed study regarding the relationship between the canonical form and the Radon transform is called for, as well as relationship of the NLV-convolutive method with the Gabor filtering techniques (or ARF if STSF is considered).

Some of the implications cited for the BMS also hold in the context of STSF, for example the same phenomena depicted in Fig. 7.3 can be observed in the joint distribution of angles $\angle\hat{f}$ in time and frequency.

LSF can be exploited to improve many image processing algorithms, such as feature detection, rescaling, vectorization and so on.

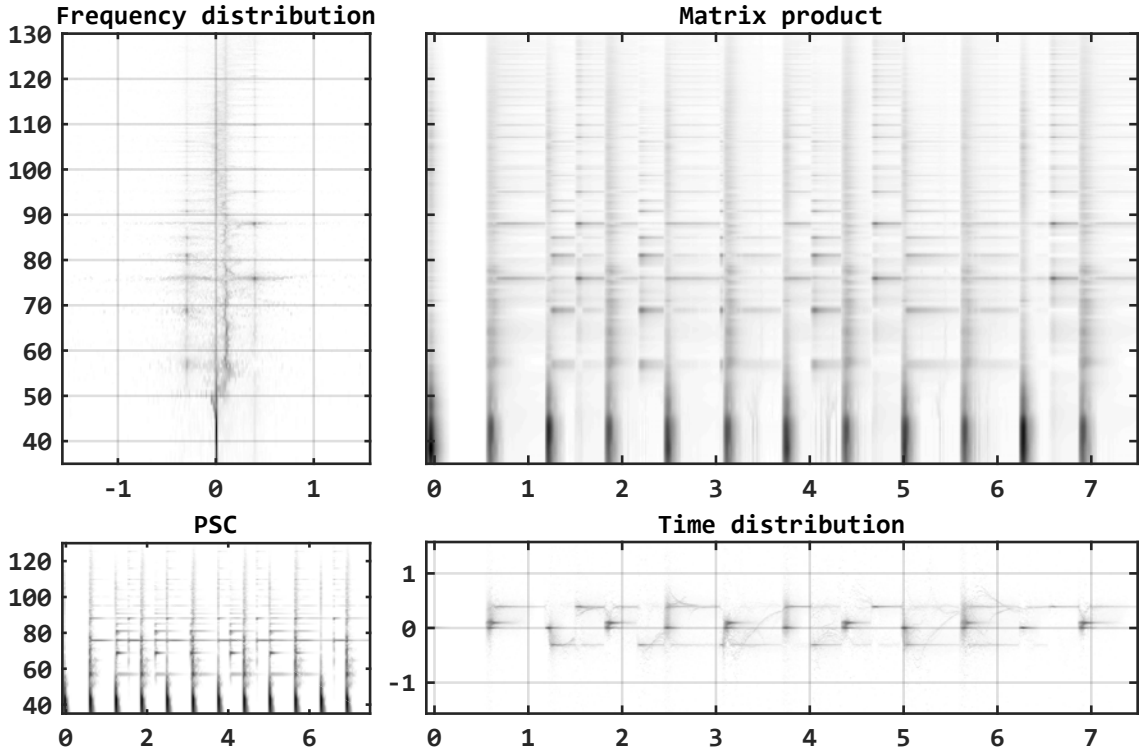


Figure 7.3: Product of the two weighted joint distributions of σ approximates $\left|\overline{X}\right|$

Bibliography

[1] Stumpf, Carl. "Tonpsychologie (Vol. 1-2)." Leipzig: Hirzel (1883-1890).

[2] Ehrenfels, Christian von. "Über gestaltqualitäten." Vierteljahresschrift für wissenschaftliche Philosophie 14.3 (1890): 249-292.

[3] Wertheimer, Max. "Musik der Wedda." Sammelbände der internationalen Musikgesellschaft 11.H. 2 (1910): 300-309.

[4] Sloboda, John A. "The musical mind: The cognitive psychology of music". Oxford University Press, 1985.

[5] Moura, Jose. "What is signal processing?[President's Message]." IEEE Signal Processing Magazine 26.6 (2009): 6-6.

[6] Baron Fourier, Jean Baptiste Joseph. "The analytical theory of heat." The University Press, 1878.

[7] Radon, J. "On determination of functions by their integral values along certain multiplicities." Ber. der Sachische Akademie der Wissenschaften Leipzig,(Germany) 69 (1917): 262-277.

[8] Lilly, Jonathan M., and Sofia C. Olhede. "Bivariate instantaneous frequency and bandwidth." IEEE Transactions on Signal Processing 58.2 (2010): 591-603.

[9] Deans, Stanley R. "The Radon transform and some of its applications". Courier Corporation, 2007.

[10] Ludwig, Donald. "The Radon transform on Euclidean space." Communications on Pure and Applied Mathematics 19.1 (1966): 49-81.

[11] Gel'Fand, I. M., M. I. Graev, and N. Ya Vilenkin. "Generalized Functions. Integral Geometry and Representation Theory, vol. 5." (1966).

[12] Duda, Richard O., and Peter E. Hart. "Use of the Hough transformation to detect lines and curves in pictures." Communications of the ACM 15.1 (1972): 11-15.

[13] van Ginkel, Michael, CL Luengo Hendriks, and Lucas J. van Vliet. "A short introduction to the Radon and Hough transforms and how they relate to each other." Delft University of Technology (2004).

[14] Helgason, Sigurdur. "The Radon transform on Euclidean spaces, compact two-point homogeneous spaces and Grassmann manifolds." Acta Mathematica 113.1 (1965): 153-180.

[15] Rowland, S. "Computer implementation of image reconstruction formulas." Image reconstruction from projections (1979): 9-79.

[16] Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2.11 (1901): 559-572.

[17] Hotelling, Harold. "Analysis of a complex of statistical variables into principal components." Journal of educational psychology 24.6 (1933): 417.

[18] Herault, Jeanny, and Christian Jutten. "Space or time adaptive signal processing by neural network models." Neural networks for computing. Vol. 151. No. 1. AIP Publishing, 1986.

[19] Comon, Pierre. "Independent component analysis, a new concept?." Signal processing 36.3 (1994): 287-314.

[20] Lawton, William H., and Edward A. Sylvestre. "Self modeling curve resolution." Technometrics 13.3 (1971): 617-633.

[21] Paatero, Pentti, and Unto Tapper. "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values." Environmetrics 5.2 (1994): 111-126.

[22] Anttila, Pia, et al. "Source identification of bulk wet deposition in Finland by positive matrix factorization." Atmospheric Environment 29.14 (1995): 1705-1718.

[23] Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788-791.

[24] Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." Advances in neural information processing systems. 2001.

[25] Alías, Francesc, Joan Claudi Socoró, and Xavier Sevillano. "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds." Applied Sciences 6.5 (2016): 143.

[26] Alan, Dower Blumlein. "Sound-transmission, sound-recording, and sound-reproducing system." U.S. Patent No. 2,093,540. 21 Sep. 1937.

[27] Eargle, John M. "Stereo/Mono Disc Compatibility: A Survey of the Problems." Journal of the Audio Engineering Society 17.3 (1969): 276-281.

[28] Bauer, Benjamin B. "Phasor analysis of some stereophonic phenomena." The Journal of the Acoustical Society of America 33.11 (1961): 1536-1539.

[29] Klapuri, Anssi, and Manuel Davy, eds."Signal processing methods for music transcription." Springer Science & Business Media, 2007.

[30] Plack, Christopher J., Andrew J. Oxenham, and Richard R. Fay, eds. "Pitch: neural coding and perception. Vol. 24." Springer Science & Business Media, 2006.

[31] Harold S. Powers, Randel, Don Michael. "Melody - The Harvard dictionary of music. Vol. 16." Harvard University Press, 2003.

[32] Brown, Guy J., and Martin Cooke. "Computational auditory scene analysis." Computer Speech & Language 8.4 (1994): 297-336.

[33] Wang, DeLiang, and Guy J. Brown. "Computational auditory scene analysis: Principles, algorithms, and applications." Wiley-IEEE Press, 2006.

[34] Bregman, Albert S., Christine Liao, and Robert Levitan. "Auditory grouping based on fundamental frequency and formant peak frequency." Canadian Journal of Psychology/Revue canadienne de psychologie 44.3 (1990): 400.

[35] Bregman, Albert S. "Auditory scene analysis: Hearing in complex environments." Thinking in sound: The cognitive psychology of human audition (1993): 10-36.

[36] Patterson, Roy D., and John Holdsworth. "A functional model of neural activity patterns and auditory images." Advances in speech, hearing and language processing 3.Part B (1996): 547-563.

[37] Slaney, Malcolm. "An efficient implementation of the Patterson-Holdsworth auditory filter bank." Apple Computer, Perception Group, Tech. Rep 35 (1993): 8.

[38] Darwin, C. J. "Auditory grouping in speech perception." International Journal of Psychology. Vol. 31. No. 3-4. 27 Church RD, Hove, East Sussex, England BN3 2FA: Psychology Press, 1996.

[39] Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte. "Performance measurement in blind audio source separation." IEEE transactions on audio, speech, and language processing 14.4 (2006): 1462-1469.

[40] Févotte, Cédric, Rémi Gribonval, and Emmanuel Vincent. "BSS_EVAL toolbox user guide–Revision 2.0." Technical Report (2005): 19.

[41] Emiya, Valentin, et al. "Subjective and objective quality assessment of audio source separation." IEEE Transactions on Audio, Speech, and Language Processing 19.7 (2011): 2046-2057.

[42] Cano, Estefanía, Derry FitzGerald, and Karlheinz Brandenburg. "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics." Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016.

[43] Ono, Nobutaka, et al. "The 2015 signal separation evaluation campaign." International Conference on Latent Variable Analysis and Signal Separation. Springer International Publishing, 2015.

[44] Vinyes, M. "MTG Mass database." `http://www.mtg.upf.edu/static/mass/resources` (2008).

[45] Fay, Richard R., ed. "The mammalian auditory pathway: Neurophysiology. Vol. 2." Springer Science & Business Media, 2013.

[46] Oliver, Douglas L. "Ascending efferent projections of the superior olivary complex." Microscopy research and technique 51.4 (2000): 355-363.

[47] Miller, Lee M., et al. "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex." Journal of neurophysiology 87.1 (2002): 516-527.

[48] Qiu, Anqi, Christoph E. Schreiner, and Monty A. Escabí. "Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition." Journal of Neurophysiology 90.1 (2003): 456-476.

[49] Elhilali, Mounya, et al. "Auditory cortical receptive fields: stable entities with plastic abilities." The Journal of Neuroscience 27.39 (2007): 10372-10382.

[50] Atencio, Craig A., and Christoph E. Schreiner. "Spectrotemporal processing in spectral tuning modules of cat primary auditory cortex." PloS one 7.2 (2012): e31537.

[51] Lindeberg, Tony, and Anders Friberg. "Scale-space theory for auditory signals." International Conference on Scale Space and Variational Methods in Computer Vision. Springer International Publishing, (2015).

[52] Lindeberg, Tony, and Anders Friberg. "Idealized computational models for auditory receptive fields." PloS one 10.3 (2015): e0119032.

[53] Hubel, David H., and Torsten N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." The Journal of physiology 160.1 (1962): 106-154.

[54] Hubel, David H., and Torsten N. Wiesel. "Receptive fields and functional architecture of monkey striate cortex." The Journal of physiology 195.1 (1968): 215-243.

[55] Wolfe, Patrick J., Simon J. Godsill, and Monika Dorfler. "Multi-Gabor dictionaries for audio time-frequency analysis." Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the. IEEE, 2001.

[56] Daugman, John G. "Two-dimensional spectral analysis of cortical receptive field profiles." Vision research 20.10 (1980): 847-856.

[57] Barry, Dan, Bob Lawlor, and Eugene Coyle. "Sound source separation: Azimuth discriminiation and resynthesis." 7th. International Conference on Digital Audio Effects (DAFx). 2004.

[58] Briand, Manuel, Nadine Martin, and David Virette. "Parametric representation of multichannel audio based on principal component analysis." Audio Engineering Society Convention 120. Audio Engineering Society, 2006.

[59] Goodwin, Michael M., and Jean-Marc Jot. "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement." 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. Vol. 1. IEEE, 2007.

[60] Vickers, Earl. "Frequency-domain two-to three-channel upmix for center channel derivation and speech enhancement." Audio Engineering Society Convention 127. Audio Engineering Society, 2009.

[61] Kraft, Sebastian, and Udo Zölzer. "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain." 18th International Conference on Digital Audio Effects (DAFx). 2015.

[62] Presti, Giorgio, Goffredo Haus, and Davide A. Mauro. "Visualization and manipulation of stereophonic audio signals by means of IID and IPD". Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2014.

[63] Barry, Dan, et al. "Drum source separation using percussive feature detection and spectral modulation." IEE Irish Signals and Systems Conference (ISSC). (2005): 13–17

[64] Salamon, Justin, and Emilia Gómez. "Melody extraction from polyphonic music signals using pitch contour characteristics." IEEE Transactions on Audio, Speech, and Language Processing 20.6 (2012): 1759-1770.

[65] Gómez, Emilia, et al. "Predominant Fundamental Frequency Estimation vs Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing." ISMIR. 2012.

[66] Bosch, Juan J., Ricard Marxer, and Emilia Gómez. "Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music." Journal of New Music Research (2016): 1-17.

[67] Ozer, Hamza, et al. "Steganalysis of audio based on audio quality metrics." Electronic Imaging 2003. International Society for Optics and Photonics, 2003.

[68] Ajmera, Pawan K., Dattatray V. Jadhav, and Raghunath S. Holambe. "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram." Pattern Recognition 44.10 (2011): 2749-2759.

[69] Kim, Myung Jong, and Hoirin Kim. "Automatic extraction of pornographic contents using radon transform based audio features." Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on. IEEE, 2011.

[70] Fitzgerald, Derry, Antoine Liutkus, and Roland Badeau. "PROJET-Spatial Audio Separation using Projections." 41st International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[71] Ware, Colin. "Information visualization: perception for design." Elsevier, 2012.

Appendices

Appendix A

Acronyms

What follows is a list of all acronyms used in this thesis, listed in alphabetical order. Important acronyms used very frequently (or used far away from their definition) ar written with a bold font. Also note that the beginning of Section 4.3 consists in a list that resumes the mathematical symbols about the BMS introduced by this thesis.

A1: Primary Auditory Cortex, defined in Section 2.4.1

ADRess: Azimuth Discrimination and Resynthesis, defined in Section 3.2

APS: Artefacts Perceptual Score, defined in Section 2.3.4

ARF: Auditory Receptive Fields, defined in Section 3.4

BMS: Bivariate Mixture Space, defined in Section 4.2.1

BS: Bivariate Spectrum, defined in Section 4.2.2

CASA: Computational Auditory Scene Analysis, defined in Section 2.3.3

DCT: Discrete Cosine Transform, defined in Section 3.5

FS: Fourier Series, defined in Section 2.1.1

ICA: Independent Component Analysis, defined in Section 2.2.2

ICC: Inferior Colliculus, defined in Section 2.4.1

ICLD: Inter Channel Level Difference, defined in Section 3.2

ICPD: Inter Channel Phase Difference, defined in Section 3.2

ILD: Interaural Level Difference, defined in Section 2.4.1

ITD: Interaural Time Difference, defined in Section 2.4.1

IPD: Interaural Phase Difference, defined in Section 2.4.1

IPS: Interference Perceptual Score, defined in Section 2.3.4

LR: Left-Right encoding, defined in Section 2.3.1

LSF: Linear Structure Field, defined in Section 5.1

MS: Mid-Side encoding, defined in Section 2.3.1

M&R: Mask and Rotate, defined in Section 4.3.4

M&R+C: Mask and Rotate, accounting for Correlation, defined in Section 4.3.4

NMF: Non-negative Matrix Factorization, defined in Section 2.2.3

NLVS: Non-Linear Vector Summation, defined in Section 5.2.2

OPS: Overall Perceptual Score, defined in Section 2.3.4

PC: Principal Component, defined in Section 2.2.1

PCA: Principal Component Analysis, defined in Section 2.2.1

PCss: Principal Component subspace, defined in Section 6.2

PEASS: Perceptual Evaluation of Audio Source Separation, defined in Section 2.3.4

PFD: Percussive Feature Detection, defined in Section 3.3

PSC: Principal Spectral Content, defined in Section 4.2.2

***Q*1**: How can the Fourier Transform be improved for considering the relationship that underlies bivariate signals? (defined in Section 1.3)

***Q*2**: How can the Short-Term Fourier Transform be improved for considering the relationship between neighbour frames? (defined in Section 1.3)

SEAD: Signal Energy Angular Distribution, defined in Section 5.1

SiSEC: Signal Separation Evaluation Campaign, defined in Section 2.3.4

SMC: Sound and Music Computing, defined in Section 2.3

SOC: Superior Olivary Complex, defined in Section 2.4.1

STFT: Short-Term Fourier Transform, defined in Section 2.1.1

STSF: Spectre-Temporal Structure Field, defined in Section 5.1

TPS: Target Perceptual Score, defined in Section 2.3.4

V1: Primary Visual Cortex, defined in Section 2.4.1

Appendix B

Experimental data

Input		OPS					TPS					IPS					APS				
Song	Taget	ADRess	+PFD	+ARF	BSM	+STSF	ADRess	+PFD	+ARF	BSM	+STSF	ADRess	+PFD	+ARF	BSM	+STSF	ADRess	+PFD	+ARF	BSM	+STSF
01_Pop_Rock	Brushes	8,32	9,04	8,65	8,44	9,89	14,78	15,64	1,82	14,28	8,78	51,88	50,99	51,99	50,36	60,00	2,15	1,19	5,13	2,06	2,53
02_Pop_Rock_nofx	Bass	7,62	11,51	11,72	7,65	7,25	10,13	6,67	3,22	10,61	8,50	48,30	64,46	41,34	48,33	43,26	7,20	5,21	6,07	7,20	11,31
03_Pop_Rock	Vocals	33,34	23,78	31,92	33,35	30,81	58,19	9,91	26,77	57,94	52,66	51,49	58,58	57,43	50,79	47,09	47,92	18,67	31,78	48,32	48,54
04_Pop_Rock_nofx	Vocals	26,26	22,69	27,73	26,22	26,70	51,72	11,03	20,78	52,13	58,56	48,03	57,22	52,44	47,85	45,72	41,21	20,68	29,08	41,66	40,99
05_Hip_Hop	Vocals	45,26	32,85	43,62	45,39	43,77	67,76	36,55	49,31	67,85	66,87	62,73	67,21	65,33	63,06	61,67	52,06	35,71	45,65	51,78	51,23
06_Hip_Hop	Vocals	18,98	32,84	18,15	8,33	8,23	75,99	46,93	39,78	66,62	49,41	18,89	42,34	18,55	12,18	3,91	75,23	48,06	60,31	86,74	82,61
07_Indie_Rock	Guitar	20,17	18,13	21,64	22,36	23,14	55,99	13,01	25,52	73,14	54,35	56,53	65,42	57,13	38,19	41,47	31,04	20,46	31,73	46,37	50,06
08_Indie_Rock	Vocals	27,63	23,81	27,75	27,55	26,53	44,19	15,51	19,15	44,04	45,36	45,96	54,46	49,46	46,25	42,85	40,38	24,53	29,70	40,16	41,08
09_Metal	Bass	5,51	6,62	6,92	5,17	5,09	34,19	18,34	5,34	35,05	25,63	45,99	58,10	43,56	45,45	37,85	3,33	7,75	16,30	3,91	13,87
10_Metal_nofx	Guitars	44,70	34,05	44,37	31,06	40,38	68,24	34,98	57,60	44,61	27,28	67,35	72,69	67,08	49,55	55,09	46,23	37,34	49,70	7,36	17,03
11_Metal	Vocals	30,81	23,40	33,82	15,78	18,23	76,50	13,36	16,44	77,50	51,12	34,20	48,86	42,78	12,32	12,39	43,85	25,33	17,05	77,87	69,42
12_Metal_nofx	Vocals	18,07	21,50	22,74	18,27	18,78	43,71	4,50	6,87	43,46	46,90	54,73	68,18	57,24	54,55	54,46	16,07	5,79	14,26	15,11	13,20
13_Bossanova	Vocals	43,04	32,11	43,60	45,35	42,88	75,20	38,89	57,93	69,91	61,83	57,92	70,64	58,48	61,84	57,75	45,42	37,49	51,84	50,95	56,81
14_Bossanova	Vocals	42,98	28,54	42,42	42,05	41,17	73,60	29,46	56,05	73,21	61,42	64,77	76,92	62,64	66,24	65,03	39,73	28,80	49,69	37,68	47,99
15_Reggae	Vocals	37,85	29,75	41,43	37,83	38,74	81,96	47,84	59,22	82,04	69,69	46,41	57,96	53,41	46,35	48,37	45,87	36,58	54,65	45,89	55,72
16_Reggae	Vocals	1,88	1,72	2,00	1,89	2,01	93,34	95,33	92,37	93,25	92,40	16,06	46,16	14,47	15,79	13,78	19,77	1,77	26,47	20,49	27,03

Input		OPS					TPS					IPS					APS				
Song	Target	ADRess	+PFD	+ARF	BSM	+STSF	ADRess	+PFD	+ARF	BSM	+STSF	ADRess	+PFD	+ARF	BSM	+STSF	ADRess	+PFD	+ARF	BSM	+STSF
01_Pop_Rock	Brushes				●	○	●	○				●				○	●		○		
02_Pop_Rock_nofx	Bass			○	●					●	○		○		●		●			●	○
03_Pop_Rock	Vocals			○	●		●				○	●	○							●	○
04_Pop_Rock_nofx	Vocals	●		○						●	○	●	○							●	○
05_Hip_Hop	Vocals				●	○				●	○		○		●		●				○
06_Hip_Hop	Vocals	●	○				●				○	●	○							●	○
07_Indie_Rock	Guitar				●	○				●	○	●	○							●	○
08_Indie_Rock	Vocals	●		○			●				○		○		●		●				○
09_Metal	Bass	●		○						●	○	●	○						○	●	
10_Metal_nofx	Guitars	●		○			●		○			●	○				●		○		
11_Metal	Vocals	●		○						●	○	●	○							●	○
12_Metal_nofx	Vocals			○	●		●				○	●	○				●		○		
13_Bossanova	Vocals			○	●		●				○		○		●					●	○
14_Bossanova	Vocals	●		○			●				○		○		●		●		○		
15_Reggae	Vocals	●		○						●	○	●	○							●	○
16_Reggae	Vocals				●	○	●	○				●	○							●	○

Table B.1: Table of raw PEASS results of experimental data described in Chapter 6. The bottom part marks with a black dot the maximum between BMS and ADRess, while black circles marks the maximum between ARF, PFD and STSF.

Gender	Age	Color blind	Use sonograms	Synthesized				Beatles				Coltrane				Winehouse				Like	Dislike
				PSC	STSF	BS	PCss	PSC	STSF	BS	PCss	PSC	STSF	BS	PCss	PSC	STSF	BS	PCss		
M	18-25	No	Frequently	2	3	4	4	2	2	3	3	4	2	3	4	5	5	5	6	PCss	PSC
M	26-35	No	Sometimes	4	4	5	5	5	5	**4**	**4**	**5**	**5**	4	**5**	3	5	4	5	PCss	PSC
F	26-35	No	Sometimes	4	4	5	7	3	**4**	3	5	4	4	4	4	3	3	4	7	PCss	PSC
M	26-35	No	Sometimes	4	4	4	4	3	3	2	**4**	4	3	4	3	3	3	3	4	PCss	PSC
M	26-35	No	Sometimes	4	4	4	3	1	1	1	1	4	6	2	2	10	10	3	10	PSC	BS
F	26-35	No	Never	3	3	4	4	3	2	5	3	4	2	3	4	1	1	3	4	BS	BS
M	36-55	No	Never	3	3	4	5	3	**4**	3	3	4	3	4	3	10	2	10	10	STSF	PSC
M	26-35	No	Sometimes	3	2	3	3	2	**4**	3	**4**	2	3	1	2	2	4	3	10	STSF	PCss
M	>55	No	Frequently	5	5	6	5	**4**	**4**	5	**4**	4	3	3	3	3	4	5	5	BS	BS
F	36-55	No	Sometimes	3	4	5	6	3	3	**4**	**4**	3	3	4	4	3	3	3	5	PCss	STSF

Table B.2: Table of raw results from the visualization validation described in Chapter 6.