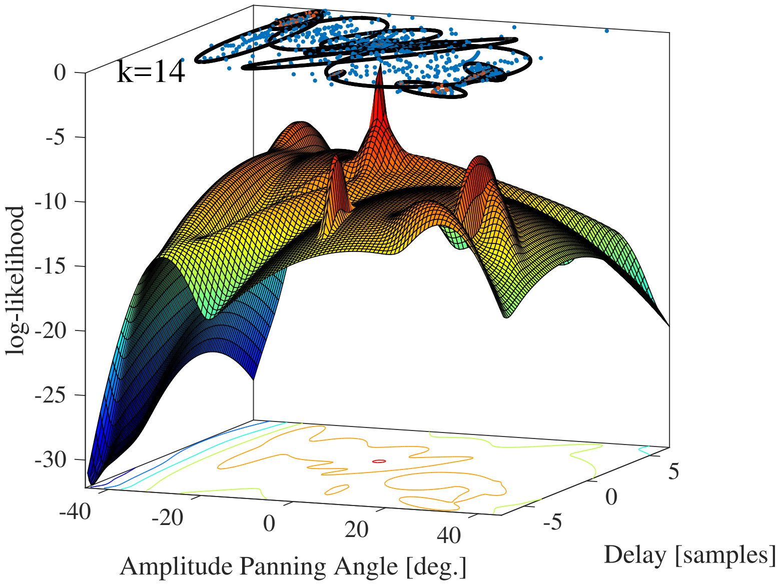
Estimation of Source Parameters and Segmentation of Stereophonic Music Mixtures.

May 2017

Jacob Møller Hjerrild

Master's Thesis



Aalborg University
Sound & Music Computing
Rendsburggade 14
DK-9000 Aalborg

AALBORG UNIVERSITY
STUDENT REPORT

Architecture Design & Media Technology
Rendsburggade 14
DK-9000 Aalborg
http://smc.aau.dk

Title:
Estimation of Source Parameters and Segmentation of Stereophonic Music Mixtures.

Theme:
Sound and Music Computing Master Thesis

Project Period:
Spring Semester 2017

Author:
Jacob Møller Hjerrild

Jacob Møller Hjerril

Supervisor:
Mads Græsbøll Christensen

Number of Prints:
none

Number of Pages:
93

Date for Submission:
May 22$^{\text{nd}}$, 2017

Abstract:

In this report, we propose a novel source parameter estimator for stereophonic mixtures, allowing for panning parameter estimation on multi-channel audio, even if the source pitches and harmonic amplitudes are unknown. The presented method does not require prior knowledge of the number of sources present in the mixture. The estimator is formulated using an unsupervised learning framework, using Bayesian statistics, allowing for optimal segmentation of the stereophonic signal, based on maximum a posteriori modelling of source parameters.

In the proposed method, we model the distribution of panning parameters with a Gaussian mixture model (GMM). Then we estimate the model parameters by using the maximum a posteriori (MAP) estimation based on the expectation-maximization (EM) algorithm. In order to avoid one cluster being modeled by two or more Gaussians, we utilize a sparse distribution modeled by the Dirichlet distributions as the prior of the GMM mixture probabilities, along with a model pruning algorithm. Moreover, to obtain a better time segmentation of the stereophonic mixtures, we propose to apply a segmentation scheme that guarantees the global optimality, based on the cost function of the maximum a posteriori model. The developed estimator is evaluated through simulations on synthetic signals as well as on real audio signals. These simulations show that the developed estimator performs good in terms of source parameter estimation and number or sources in the stereophonic mixture.

The content of this report is not public available, and publication may only be pursued due to agreement with the author.

Preface

This master's thesis is written by Jacob Møller Hjerrild, at the Department of Architecture, Design and Media Technology on Aalborg University during the 10th semester in the project period spanning from February 1, 2017 to May 22, 2017. During the project period, I was affiliated with the Audio Analysis Lab group at Aalborg University. The thesis is concerned with the estimation of source parameters in stereophonic mixtures. This is a new problem and it has only been an active part of research during the recent year in the Audio Analysis Lab, where it has been shown that more efficient and a more precise multi-pitch estimator can be achieved by knowing the stereophonic panning parameters when applied to stereophonic music. The few approaches so far for solving this estimation problem have been based on tools from parametric multi-pitch estimation, that requires a preceeding pitch estimate in order to estimate the panning parameters. In this thesis, however, a different approach is taken based on unsupervised learning, using Bayesian statistics. The unsupervised learning approach offers the advantage that no prior knowledge of pitch is needed for solving the estimation problem. By taking this approach, new knowledge from unsupervised learning of stereophonic mixtures has been added into the Audio Analysis Lab group and this may be important for future research work. Furthermore, the problem of estimating the stereophonic panning parameters, is in general a new research subject within the audio signal processing community and therefore the proposed solution brings novelty and has no explicit precursor.

The contents of the thesis is a description of the proposed solution. An introduction in Chapter 1, is concerned with the definition of the stereophonic parameters and the signal model parameters and assumptions. Chapter 1 also serves the purpose of defining the measurement space which is basis for the unsupervised learning by clustering. Chapter 2 describes the clustering as a maximum likelihood approach. Chapter 3 concludes Chapter 2 by applying Bayesian maximum a posterior model order selection that is basis for a proposed component annihilation. Chapter 4 is concerned with the proposal of a stereophonic segmentation scheme based on the parameters of the Gaussian mixture model. In chapter 5, the proposed solution is evaluated trough experimentation. The thesis is concluded in chapter 6. In the appendices, a few intermediate test results are shown.

The reader should be aware of the following typographical conventions of this thesis: All figures, tables and equations are referred to by the number of the chapter they are used in, followed by a number indicating the number of figure, table or equation in the specific chapter. Hence, each figure has a unique number, which is also printed at the bottom of the figure along with a caption. An example is Figure 2.1, which means the first figure in Chapter 2. The same applies to tables and equations, the latter of which have no captions. Appendices are referred to by capital letters instead of chapter numbers. At the very end of the report, a bibliography is listed which contains all sources of research used for reference.

I would like to thank my supervisor Prof. Mads Græsbøll Christensen who have guided and inspired me through this master project as well as the projects conducted on the 7th and 8th semester. They have been a significant part of what I have achieved during the learning process that I went through in the recent years. This guidance also played a great part in setting up a three month internship at the company AM3D A/S, which was a great learning experience and an eye opener for me, in terms of working in the industry of signal processing as an engineer. Lastly, I am grateful towards all staff members in the Media Technology of Aalborg University who have always been helpful and welcoming which has been of valuable support for me, to accomplish the realization of my masters degree.

Nomenclature

$(\cdot)^T$	matrix transpose
$\arg\max_{\mathbf{x}} f(\mathbf{x})$	the value of $\mathbf{x}$ which maximizes $f(\mathbf{x})$
$\det(\cdot)$	determinant
$(\hat{\cdot})$	estimator or an estimate
$\mathbb{E}$	expectation operator
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$	$\mathbf{x}$ has a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$
$\boldsymbol{X}$	matrix
$\boldsymbol{x}$	vector
$\mathbf{I}_N$	the $N \times N$ identity matrix
j	imaginary unit
$p(\cdot)$	probability distribution
$p(\cdot, \cdot)$	joint probability distribution
x	scalar
$x^{(i)}$	the value of x at the ith iteration
AIC	Akaike information criterion
BIC	Bayes information criterion
BSS	blind source separation
CH	Calinski-Harabasz
DAW	digital audio workstation
DFT	discrete Fourier transform

EM	expectation-maximization
GMM	Gaussian mixture model
MAP	maximum a posteriori
MDL	minimum description length
ML	maximum likelihood
MLE	maximum likelihood estimate
MMDL	mixture-MDL
PA	public adress
VRC	variance ratio criterion

Contents

Chapter 1

Introduction

The stereophonic source parameters of the mixture created in recording studios, have recently been shown to improve multi-pitch estimation [1]. Pitch estimation is important and has many applications such as separation [2], enhancement [3], transcription [4], classification [5], and source localization [6]. The latter of which is closely related to the virtual source positioning technique [7], which has a central role in this thesis.

1.1 Motivation

The problem of estimating source parameters in stereophonic mixtures, is a new problem and it has only been an active part of research during the recent years in the Audio Analysis Lab. The reason for this new interest is that it has been shown that improvement in precision can be achieved for the multi-pitch estimator, by knowing the stereophonic panning parameters [8] when applied to music mixtures. Music mixtures involves interdependent harmonic structures, between sources, and spectral overlap is a common problem within multi-pitch estimation of musical content [9, 10, 11]. Music mixtures are mainly available in stereo and Weiss et al. [12] proposed a novel stereophonic maximum likelihood multi-pitch estimator, that utilizes the panning parameters, by assuming that these are known. To the authors knowledge, only one approach exists for solving this problem of estimating panning parameters [13], which is based on convex optimization techniques, but it requires a preceeding pitch estimate in order to estimate the panning parameters. No solution exists for explicit estimation of stereophonic amplitude and delay panning parameters without pitch information. This thesis proposes such an algorithm.

1.2 Source Positioning

The stereophonic virtual source positioning is based on human spatial perception and psycho physics of sound [14]. Virtual positioning of sound sources can be described with amplitude and time-delay ratios between the stereophonic channels. The amplitude and time-delay ratios are basic for the human auditory spatial perception and these features are also applied within array processing [15, 16] and blind source separation (BSS) [17] of speech signals. BSS methods are well suited to speech mixtures, due to the sparse structure of speech which leads to the assumption of W-disjoint orthogonality [18]. The BSS algorithm of [19] builds a weighted histogram in time-frequency domain, and requires manual inspection for the parameter estimates or prior knowledge of the number of speakers. Time-frequency amplitude ratios have also been applied for stereo upmixing techniques [20], without explicit estimation of source parameters. The BSS methods are generally operating in the time-frequency domain, which implicitly requires a uniform time segmentation of the signals. However, usually we can not know the number of sources in advance and the signal content is changing over time and therefore a varying segment length can be appropriate.

The following proposal is a method that does not require prior knowledge of the number of sources present in the mixture. The estimator is formulated using an unsupervised learning framework, using Bayesian statistics, allowing for optimal segmentation of the stereophonic signal, based on maximum a posteriori modelling of source parameters. In the proposed method, the distribution of panning parameters are modelled with a Gaussian mixture model (GMM). The model parameters are estimated by using the maximum a posteriori (MAP) estimation based on the expectation-maximization (EM) algorithm. With suitable priors on the parameters, the MAP estimator can be used for model selection [21] and [22] of the components in the Gaussian mixture. In order to avoid one cluster being modelled by two or more Gaussians, this approach is modeled by the Dirichlet distributions as the prior of the GMM mixture probabilities, along with a model pruning algorithm. To obtain a better time segmentation, we propose to apply a sterophonic segmentation scheme that guarantees the global optimality, based on the cost function of the maximum a posteriori model.

The rest of this report is organized as follows: The remaining part of Chapter 1, is concerned with the definition of the stereophonic parameters and the signal model parameters and assumptions. Chapter 1 also serves the purpose of defining the measurement space which is basis for the unsupervised learning by clustering. Chapter 2 describes the clustering as a maximum likelihood approach. Chapter 3 concludes Chapter 2 by applying Bayesian maximum a posterior model order selection that is basis for a proposed component annihilation. Chapter 4 is concerned with the proposal of a stereophonic segmentation scheme based on the parameters of the Gaussian mixture model. In chapter 5, the proposed solution is evaluated trough experimentation. The thesis is concluded in chapter 6. In the appendices, a few intermediate test results are shown.

1.3 The Panning Parameters

The panning parameters discussed in this report is a product of the mixing process applied in sound studios. To enhance the sound quality and to ease the virtual perceived separation of sound sources in a stereo mixture the sound engineer can apply various effects, such as amplitude and delay panning. Other effects such as reverb, equalization and dynamic effects are usually also applied, but are of no interest in the remaining report. Amplitude and delay panning is exactly the two parameters that we estimate in this report, since they carry spatial information that is equivalent to direction and positioning in a real geometric setup.

1.3.1 Delay Panning

The delay panning parameter is directly related to the time delay that humans experience when a sound signal is propagated through the air from some source and received at each ear at separate time instances. Such a delay changes the perceived direction of the sound source [14]. It has been shown that a constant time delay to one of the speakers is frequency dependent in terms of virtual source positioning [23]. Though amplitude panning is the traditional post processing “way to go” for sound engineers, delays are being added as part of post processing mixing procedure both to correlate phases of microphones and to change directivity of sources. Lastly, delays longer than 1 ms can be applied to achieve depth and dimension, by virtually placing the sound source mostly in the channel where the signal arrives first [24].

1.3.2 Amplitude Panning

Amplitude panning is the general method for altering the perceived direction of a sound source in a sound field between two or more loud speakers. Amplitude panning is an approximation of source localisation and its application ranges from stereophonic amplifiers to multichannel speaker setup and professional multi-channel mixing desks, DAW’s etc. Most often the user/engineer of a mixing desk can configure the the perceived direction of each individual sound source in the mix by turning one knob, attached to a trim pot that controls the signal voltage level to each speaker output. If the desk is digital, the user has a similar digital knob or slider interface.

Amplitude panning can be applied to multi-speaker setups, while the most common speaker configuration is a stereo setup, consisting of a left and a right speaker, with two audio channels being played back (one for each loudspeaker), whether it is a home audio hi-fi system, PA (Public Adress) system, headphone system etc. The stereophonic configuration is shown in Figure 1.1, where the listener is placed in orego,

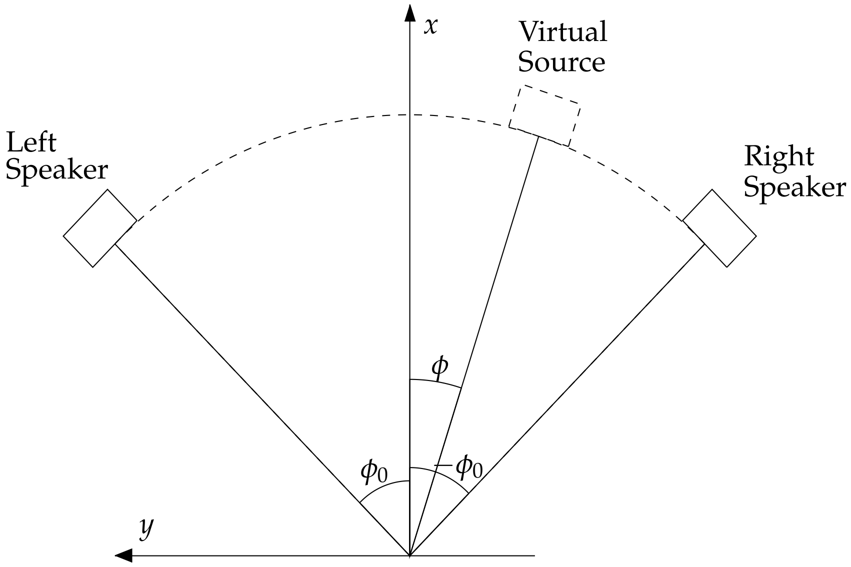


Figure 1.1: Stereophonic Configuration

fronting the x-axis in ($x = y = 0$). Amplitude panning in the sterephonic configuration is explained in the following sub section.

1.3.3 Stereophonic Amplitude Panning

Figure 1.1 shows the stereophonic sound configuration patented by Blumlein[25]. The listener is situated equidistant to each speaker in orego. The listener perceives an illusion of an auditory event, that is placed in a specific point on a two dimensional arc between the two speakers. The auditory event is moved by changing the signal amplitudes of the signal in the left and right channel. Amplitude panning is described by Ville Pulkki [7] in a vector based framework that allows two- and three dimensional speaker setups. Amplitude panning can be formulated at time t, by applying a signal $x(t)$ to both loudspeakers with different amplitudes, and gain factors for left and right channel respectively. In general the signal $x_i(t)$ is then

$$x_i(t) = g_i x(t), \qquad i = 1, 2, \cdots, N \tag{1.1}$$

where $x_i(t)$ is the signal applied to the i^{th} loudspeaker and g_i is the gain factor of the corresponding channel and $N = 2$ is the number of speakers in stereo configuration. While the virtual source is moving along the arc, the distance to the the listener should be constant. For the stereophonic configuration the vectorial distance of the gain factors g_1 and g_2 equals a constant C

$$g_1^2 + g_2^2 = C \tag{1.2}$$

The relation between the gain factors and the perceived virtual source direction has been derived for panning in the stereophonic configuration by Bauer [26] as the "stereophonic law of sines", where the acoustic shadow of the head is not taken in to account and the sine law is assumed valid at all frequencies. For the sine law, the listener is situated symmetrically between the speakers in orego, facing along the x-axis in Figure 1.1. The stereophonic sine law is described by the ratio of the difference and sum of the gain factors as,

$$\frac{\sin\phi}{\sin\phi_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{1.3}$$

where ϕ is the perceived angle and ϕ_0 is the speaker base angle. It is required that $0^\circ < \phi_0 < 90^\circ$, $-\phi_0 \leq \phi \leq \phi_0$ and $g_1, g_2 \in [0, 1]$. An extension of the sine law is the tangent law, originally proposed by Bernfeld [27] as

$$\frac{\tan\phi}{\tan\phi_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{1.4}$$

The tangent law behaves similar to the sine law with very small difference, taking some of the listeners head complexity into account. Ville Pulkki [7] formulates the vector based approach as a reformulation of the tangent law, called the vector based amplitude panning (VBAP). Figure 1.2 visualizes the vector based framework of the stereo configuration, that is used in the remaining of this report to describe the estimated amplitude panning angle, shown in results and in figures.

Gain Vector Relation to Virtual Sound Source Positioning

To ease the understanding of the amplitude panning parameter it is convenient for the human reader to consider the parameter as a perceived angle in a carthesian coordinate system, since a music listener is normally placed in front of two speakers as mentioned in Section 1.3.3. To present the gain ratios as angles we apply the stereo vector base virtual sound source positioning [7]. A backwards amplitude panning algorithm serves the purpose of estimating the gain parameters. As visualized in Figure1.1, each loudspeaker has a base angle $\phi_0 = \pm 45^\circ$ to the x-axis direction that the listener is facing towards; the listener is situated equidistant to each speaker in $(x = y)$. The angle ϕ describes the virtual source position respective to the x-axis. The trigonometric functions are used for the panning gain since they fit the unit circle, thus they retain unity power along an arc as $1 = \cos^2 + \sin^2$. The gains are then

$$g_x = \cos\theta \tag{1.5}$$

$$g_y = \sin\theta \tag{1.6}$$

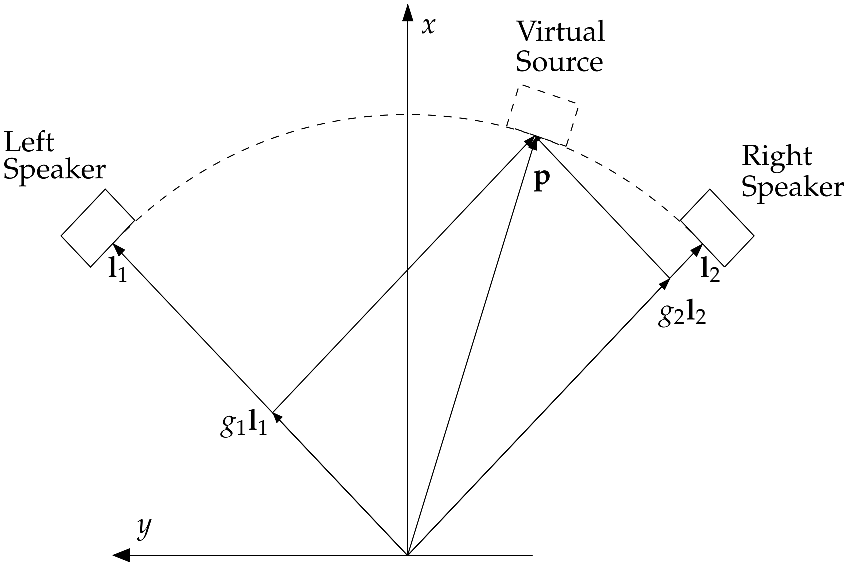


Figure 1.2: Stereophonic configuration with vector formulation

where $\theta = \phi + \phi_0$. If we define a loudspeaker base matrix **L**

$$\mathbf{L} = [\mathbf{l}_1\,\mathbf{l}_2]^T \tag{1.7}$$

consisting of two unit length loudspeaker vectors $\mathbf{l}_1 = [l_{11} l_{12}]^T$ and $\mathbf{l}_2 = [l_{21} l_{22}]^T$ pointing toward each speaker. In Figure 1.2, the unit vector **p** points towards the virtual source as a linear combination of the gained loudspeaker vectors

$$\mathbf{p}^T = \mathbf{gL} \tag{1.8}$$

This equation can be solved for the gain vector, by applying the inverse loudspeaker base matrix

$$\mathbf{g} = \mathbf{p}^T\mathbf{L}^{-1} \tag{1.9}$$

The loudspeaker base matrix **L** is unitary and $\mathbf{L}^{-1}$ exists under the conditions $0° < \phi_0 < 90°$, $-\phi_0 \leq \phi \leq \phi_0$ and $g_1, g_2 \in [\,0, 1]$. Finally, we can estimate the panning angle $\hat{\theta}$ as

$$\hat{\theta} = \arctan \frac{p(1)}{p(2)} \tag{1.10}$$

The amplitude panning angle applied to sources in a stereo mixture, can be estimated as shown from the obtained gain factors. The trigonometric functions used in this computation, estimates within the domain of the loudspeaker base matrix **L** with a span of 90°. In professional studios the aperture of loudspeakers is typically 60°. However, this relation between to the tangent law is linear and is simply solved by normalization to a wider domain by multiplication. This was a description of the panning parameters that is the subject of estimation in the following thesis. The panning parameters are used for source identification, which means that we estimate the number of sources in the stereophonic mixture, using these. In the following sections the signal model and the associated assumptions are defined.

1.4 Signal Model

In this section, the signal model and assumptions are introduced. Consider an M-channel music mixture consisting of K unknown sources corrupted by white Gaussian noise at time instant n. The data in the m^{th} channel is represented as $\mathbf{x}_m(n) \in \mathbb{R}^N$,

$$\mathbf{x}_m(n) = [x_m(n) \quad x_m(n+1) \quad \dots \quad x_m(n+N-1)]^T \tag{1.11}$$

for $m = 1, \dots, M$. The signals captured by channel m, relating to the k^{th} source are attenuated by gain coefficient $g_{m,k}$ and delayed by $\tau_{m,k}$ depending on their perceptional virtual positioning, given by the panning parameters. The signal mixture is modelled as a linear superposition of K attenuated and delayed sources embedded in noise $\mathbf{e}_{m,k}(n)$,

$$\mathbf{x}_m(n) = \sum_{k=1}^{K} g_{m,k}\mathbf{s}_k(n - f_s\tau_{m,k}) + \mathbf{e}_{m,k}(n) \tag{1.12}$$

where $g_{m,k}$ and $\tau_{m,k}$ are the attenuation and delay applied to the source $\mathbf{s}_k(n)$, respectively and f_s is the sampling frequency. Considering stereophonic mixtures with $M = 2$ for a stereo loudspeaker setup, amplitude panning is the traditional procedure [26, 7] for virtual source positioning. In the post-processing of a every music production, delays can be added to enhance the spatial perception [14]. The trigonometric functions are often used for the panning attenuation because they induce a constant perceived distance between listener and the virtual source, described by $1 = \cos^2 + \sin^2$. The gains for channel m are expressed as [28],

$$g_m = \begin{cases} \cos\theta_k, & \text{for } m = 1 \\ \sin\theta_k, & \text{for } m = 2 \end{cases} \tag{1.13}$$

where $\theta = \phi + \phi_0$ is a sum of the perceived angle ϕ and the speaker base angles $\pm\phi_0 = 45°$. Under the conditions $0° < \phi_0 < 90°$, $-\phi_0 \leq \phi \leq \phi_0$ and $g_1, g_2 \in [\,0, 1]$ the gains can be expressed as,

$$\mathbf{g}_k = \mathbf{p}_k\mathbf{L}^{-1} \tag{1.14}$$

where the unit-vector $\mathbf{p}$ points towards the virtual source with $\mathbf{L}$ as a unitary loudspeaker base matrix. For the stereophonic mixture ($M = 2$), we simplify notation by modelling attenuation and delay parameters as ratios between the frequency representations of active sources in the two channels.

1.4.1 Estimating the Panning Parameters

When only source k is active, the frequency representation in each stereophonic channel is,

$$S_{1,k}(\omega) = \sum_{n=1}^{N} s_k(n)e^{-j\omega n}, \tag{1.15}$$

$$S_{2,k}(\omega) = \sum_{n=1}^{N} \gamma_k s_k(n)e^{-j\omega n\delta_k}, \tag{1.16}$$

ω is the frequency grid, $\delta_k = f_s\tau_k$ is the relative delay of source k between the channels and γ_k is the relative attenuation factor corresponding to the ratio of attenuation of source k between the channels. The panning parameters γ_k and δ_k, that are associated with active sources in each frequency point can be computed as,

$$(\gamma_k, \delta_k) = \left(\left| \frac{S_{2,k}(\omega)}{S_{1,k}(\omega)} \right|, \frac{1}{\omega}\angle\frac{S_{1,k}(\omega)}{S_{2,k}(\omega)} \right) \tag{1.17}$$

where we must ensure that,

$$|\omega_{\max}\delta_{\max}| < \pi \tag{1.18}$$

to avoid phase ambiguity. Our aim is to estimate the panning parameters (γ, δ) for all K sources, along with an optimal segment length N, given only the stereophonic mixture in (1.4). The kth panning parameter is associated with only the kth source component, under the assumption that only one source is dominant at each frequency point. This is described by the approximate disjoint orthogonality expressed as [18],

$$S_{1,k}(\omega)S_{1,i}(\omega) \approx 0 \quad \forall\omega, k \neq i \tag{1.19}$$

Subject to this assumption, we apply a segmentation of the signal $\mathbf{x}_m(n)$ into segments of size N, that provides an improved separation of the clusters in (1.4.1). The optimal segmentation scheme is described in Section 5.1.1. The estimated amplitude and delay ratios, which often is refered to as the measurement vectors, are described from the spectral content of each channel in the stereophonic mixture as,

$$(\hat{\gamma}, \hat{\delta}) = \left(\left| \frac{X_2(\omega)}{X_1(\omega)} \right|, \frac{1}{\omega}\angle\frac{X_1(\omega)}{X_2(\omega)} \right) \tag{1.20}$$

where $X_m(\omega)$ is the discrete Fourier transform of $\mathbf{x}_m(n)$. In this domain K is unknown and we can expect the parameters to cluster in some form. Music mixtures often have a long duration of several minutes and we assume that such mixtures have stationary panning parameters throughout the full mixture i.e. a 3 minute song. We will collect measurement vectors and perform segmentation to select parts of the signal that carries relevant information of the measurement vectors. A great part of the noisefloor in the spectrum is removed which also lowers computional complexity. We define an indicator function $b(\omega)$ as,

$$b(\omega) \begin{cases} 1, & |X_1(\omega)||X_2(\omega)| > |\mathbf{X}_1|^T|\mathbf{X}_2|/N) \\ 0, & \text{otherwise} \end{cases} \tag{1.21}$$

where $X_m(\omega)$ is the pre-whitened DFT of $\mathbf{x}_m(n)$. It is possible to pick a specific number of measurement vectors by increasing the threshhold on the indicator function and improve on computational complexity. This was a description of the signal model and the main assumptions in this contex. We will end this chapter by introducing a time-panning domain visualization, which we call the panogram. The next chapter will explain the clustering of measurement vectors.

1.4.2 Visualizing the Amplitude Angle as a Panogram

A visual output of the panning angle can be used to identify the various sources in a stereo mix based on their panning coefficient. This can be acomplished via the amplitude panning ratio in (1.4.1). The computation is very fast and the output is shown in Figure 1.3 for a multi-pitch mixture of two instruments, trumpet and horn, playing the notes C4 (262 Hz) and F#4 (370 Hz), respectively. This specific mixture is also used in an experiment in [12]. The algorithm for the panogram in Figure 1.3 is based on searching through the power ratio of the absolute discrete Fourier transform of the two stereo channels. Consider a stereo input signal $x(n)$ consisting of both $x_1(n)$ and

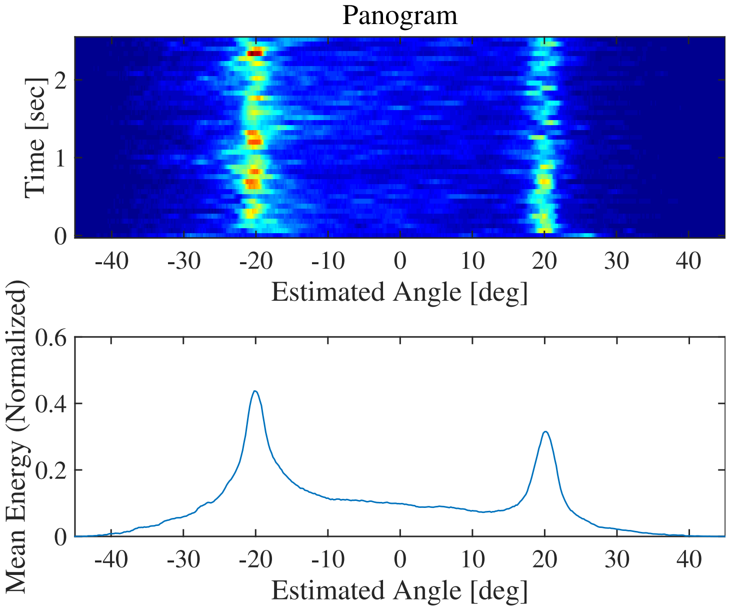


Figure 1.3: Panogram of a multi-pitch mixture of two instruments, trumpet and horn.

$x_2(n))$ at time instant n. At each time instant we compute the amplitude ratio $\gamma_n(\omega)$ using 1.4.1 and lastly marginalize by summing over all frequencies and at each time instant n, the panogram $p(\theta)$ is a vector function of θ and can be expressed the power at each panning angle. More Panograms can be found in Appendix B. The visual inspection and manual peak finding in an objective function or histogram created from time-frequency domain ratios, is used within the research area of blind source separation [18, 17, 19] and also used for channel upmix texhniques [20], however the aim in this thesis is to automatically estimate the panning parameters. Hence, we do not consider the approach of BSS and resynthesis of source signals.

In this chapter the signal model and assumptions has been defined. The measurement space has been described as consisting of stereophonic panning parameters that we could expect to cluster in some form. Therefore, we continue in the next chapter with the definition of the approach that is utilized for clustering algorithm, which is based on the Bayesian posterior modelling. Therefore, we start by defining the clustering problem as a likelihood description.

Chapter 2

Clustering

Once the measurement space containing the distribution of estimated panning parameters is well defined, it is the aim to estimate the number of sources and the source parameters as an unsupervised learning task, with no prior information given of the source parameters. The problem of finding clusters in a set of measurement vectors can be approached by using probabilistic techniques or non-probabilistic techniques. An example of a non-probabilistic clustering technique is the k-means algorithm [29]. The immediate K-means algorithm requires an input that specifies the number of clusters to estimate. We have modelled the source parameter distribution using the probabilistic clustering technique, as a mixture of Gaussians.

2.1 Estimation of Source Parameters

The source parameters are estimated by maximizing the likelihood. The maximum likelihood estimates are the parameters of the model that describe the observed measurement vectors the best, i.e. the parameters that maximizes the probability of the observed data, $\mathbf{x}$, given the parameters,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \underset{\boldsymbol{\theta}}{\arg\max}\ p(\mathbf{x}|\boldsymbol{\theta}) \tag{2.1}$$

where $\hat{\boldsymbol{\theta}}$ is a vector containing the model parameters. In the following the probabilistic model is described along with the K-means clustering algorithm that is used for initialization. We describe the maximum likelihood estimator, using latent variables and finally we consider the model order selection as both a probabilistic and non-probabilistic method.

2.2 Finite Mixtures

The following section is a brief description of the general model of finite mixtures, which the Gaussian mixture model belongs to. We have found the Gaussian mixture to be well suited for modelling the source parameter distribution. The research issue of order selection is relevant, when aiming to jointly estimate source parameters and number of sources in the stereophonic mixture. We can describe the sterophonic mixture as a finite mixture of K random sources described as probability density functions,

$$p_k(\mathbf{x}), \quad k = 1, \ldots, K \tag{2.2}$$

We observe a set of random independent distributed samples, coming from these probability density functions. We define the prior probability of observing data from source s_k as $p(s_k) = \alpha_k$, and the conditional probability of the data given source s_k is $p(\mathbf{x}|s_k) = p_k(\mathbf{x})$, thus the joint probability $p(\mathbf{x}, s_k)$ is expressed as $\alpha_k p_k(\mathbf{x})$. Finally, the unconditional probability density is,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}) \tag{2.3}$$

Which means we that the unconditional density is a finite mixture of component densities $p_k(\mathbf{x})$ weighted by their prior, referred to as the mixing probabilities which we denote α_k for the kth source. The mixing probabilities has the general constraint of summing to one.

2.2.1 Parameterization of the finite mixtures of Gaussians

The unknown parameter vector is denoted by $\boldsymbol{\theta}$. In general for a finite mixture model it will be consisting of the mixing probability and the unknown parameters,

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k, \alpha_1, \alpha_2, \ldots, \alpha_k\}$$

The conditional densities related to the source components are then given by,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k p(\mathbf{x}|\boldsymbol{\theta}_k)$$

By assuming that sources are Gaussian distributions with arbitrary covariance the conditional density is modelled as,

$$p(\mathbf{x}|\boldsymbol{\theta}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k)$$

the parameter vector contains the mean $\boldsymbol{\mu}_k$ and covariance $\mathbf{C}_k$ for $i = k, \ldots, K$,

$$\boldsymbol{\Theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k, \mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k, , \alpha_1, \alpha_2, \ldots, \alpha_k\}$$

The aim is now to estimate the parameter set $\boldsymbol{\theta}$ from the given observations. An order selection procedure will estimate the given number of sources, while panning parameters are given by the mean of the mixture components. However, the task of assigning points to mixture components is not trivial to do automatically, since the observed data with unknown classes, can be clustered in to an arbitrary number of classes, dependent on the choice of model and how the model is being fitted to the observations. The aim in such an unsupervised learning task, by model based clustering is that each component models one cluster.

$$\sum_{k=1}^{K} \alpha_k = 1 \quad \text{and} \quad 0 \leq \alpha_k \leq 1 \tag{2.4}$$

as any probability the mixing probability is required to take a value between 0 and 1. The finite mixture in this general form is possible to parameterize with the unknown parameters and by applying some model to the distribution, we can build a convenient estimator, as we will do in the following section.

2.2.2 The Gaussian mixture model as a likelihood

As discussed in Section 2.4, the K-means assigns every measurement vector uniquely to one cluster as a hard assignment. However, it is not clear that a measurement vector which is placed midway between two cluster centers is assigned appropriately, relative to the cluster center which can affect the precision of the parameter estimates. By using probabilistic models such as the Gaussian mixture model (GMM), the assignments can reflect this level of uncertainty as a soft assignment of measurement vectors to clusters. Furthermore, the mixture model is good at representing class conditional densities in supervised learning, because mixtures can approximate arbitrary densities, i.e. two strongly non-Gaussian classes, can be modelled by mixtures of each class conditional density [30]. On the contrary, in the unsupervised learning task it is a matter of fitting the model sparsely to the data without overfitting to parameter space. Therefore, the Gaussian mixture model will firsly be described as a likelihood, followed by an interpretation as an a posteriori distribution, penalizing higher model orders.

Using the GMM framework, the full parameter space is modelled as a Gaussian mixture distribution i.e. a linear superposition of Gaussians,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C}) \tag{2.5}$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \frac{(2\pi)^{-\frac{d}{2}})}{|\mathbf{C}_k|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \tag{2.6}$$

where $\boldsymbol{\mu}_k$ is the mean and $\mathbf{C}_k$ is the covariance of the kth Gaussian. The mixing probabilities $\{\alpha_1, \ldots, \alpha_K\}$ are constrained to

$$\sum_{k=1}^{K} \alpha_k = 1, \qquad 0 \leq \alpha_k \leq 1 \tag{2.7}$$

and can be interpretted as the prior probabilities of having the class k,

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \alpha_k \frac{(2\pi)^{-\frac{d}{2}})}{|\mathbf{C}_k|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{C}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\} \tag{2.8}$$

where each θ_k is the parameter specifying the kth component. The parameter vector is defined as,

$$\boldsymbol{\Theta} \equiv \{\alpha_1, \ldots, \alpha_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \, \mathbf{C}_1, \ldots, \mathbf{C}_K\} \tag{2.9}$$

The parameter vector specifies the full mixture as the complete set of parameters. Observing a set of N independent distributed samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the log-likelihood function corresponding to a K-source mixture is,

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{X}) = \ln p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathbf{C}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{C}_k)\right\} \tag{2.10}$$

Maximizing the log-likelihood of (2.10), turns out to be a complex problem mainly due to the summation inside the logarithm. The logarithm function of (2.10) does not act directly on the Gaussian, but also on the summation over k. If we differentiate the log-likehood and set it to zero it will not have a closed form solution. However, we can maximize the likelihood function with the expectation-maximization (EM) algorithm. In the following section we proceed with a general description of the EM in the context of fitting a mixture of Gaussians to the measurent vectors.

2.2.3 Fitting the Gaussian mixture as a maximum likelihood solution

A powerful method for finding the maximum likelihood solutions to models with latent variables is the EM algorithm [31, 32]. Due to the inner sum of (2.10), it is necessary to view the problem by defining a K-dimensional binary latent variable $\boldsymbol{z}$ that for a given n has k latent variables where only one of these is equal to 1, while the rest are equal to 0. This means that the vector $\boldsymbol{z}$ has K possible states and $z_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$. We can then view α_k as the prior probability $p(z_k = 1) = \alpha_k$, i.e. the probability of z_k equals 1. In these terms the marginal distribution over $\boldsymbol{z}$ can be written in the form,

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \alpha_k^{z_k} \tag{2.11}$$

The conditional distribution of $\mathbf{x}$ given $\boldsymbol{z}$ is also a Gaussian and can be described by,

$$p(\mathbf{x}|\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k)^{z_k} \tag{2.12}$$

We are now able to work with the joint distribution $p(\mathbf{x}, z) = p(z)p(\mathbf{x}|z)$. By summing the joint distribution over all possible states of z, we can obtain the marginal distribution of $\mathbf{x}$ as,

$$p(\mathbf{x}) = \sum_{z} p(z)p(\mathbf{x}|z) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

Which is equivalent to the form of the Gaussian mixture expressed as a linear superposition of Gaussian distributions as given by (2.2.2), only now there is a corresponding latent variable for each measurement vector $\mathbf{x}_n$. Observing a set of N independent distributed samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the log-likelihood function corresponding to a K-source mixture can now be expressed for the complete measurent vectors $\{\mathbf{X}, \boldsymbol{Z}\}$ containg both the observed data $\mathbf{X}$ and the latent variable $\boldsymbol{Z}$ [33]. The log-likelihood is then expressed as,

$$\ln p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathbf{C}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \{\ln\{\alpha_k\} + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{C}_k)\} \tag{2.13}$$

Since the logarithm now acts directly on the Gaussian distribution, it leads to much simpler solution for the maximum likelihood. In practice, the values of the latent variables are unknown, thus we consider the expectation with respect to the posterior distribution of the latent variables, which takes the form,

$$p(\boldsymbol{Z}|\mathbf{X}, \alpha_k, \boldsymbol{\mu}_k, \mathbf{C}_k) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} (\alpha_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{C}_k)^{z_{n,k}} \tag{2.14}$$

Where $\alpha_k = \frac{1}{N} \sum_{n=1}^{N} z_{n,k}$. The expected value of the complete data log-likelihood function is now,

$$\mathbb{E}_{\boldsymbol{Z}}[\ln p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathbf{C})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \beta(z_k) \{\ln\{\alpha_k\} + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{C}_k)\} \tag{2.15}$$

Where $\beta(z_k)$ is a quantity that plays an important role as the conditional probability of z given $\mathbf{x}$. By viewing α_k as the prior probability of $z_k = 1$ and $\beta(z_k)$ as the corresponding posterior once we have observed $\mathbf{x}$. The quantity $\beta(z_k)$ is also referred to as the responibility htat component k takes for explaining the observation of $\mathbf{x}$.

$$\beta(z_k) \equiv p(z_k = 1|\mathbf{x}) = \mathbb{E}[z_k] = \frac{\alpha_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k)} \tag{2.16}$$

The responsibility $\beta(z_k)$ applies different weight for each parameter estimate, which turns out to be crucial for the model selection procedure for the mixture model, as described in Section 3.0.2.

2.3 Maximizing the likelihood

Now that we have defined the log-likelihood by using latent variables to describe the complete data, we are ready to apply the EM-algorithm for the Gaussian mixture models. The condition that must be satisfied at the maximum of a likelihood function is found by setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathbf{C})$ in (2.10) to zero. First the mean parameter:

$$\frac{d}{d\boldsymbol{\mu}_k} \ln p(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathbf{C}) = 0 \tag{2.17}$$

$$\sum_{n=1}^{N} \frac{\alpha_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_j \alpha_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \mathbf{C}_j)} \mathbf{C}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{2.18}$$

Where it is interesting that the responsibility of (2.2.3) appears naturally, and the expression is equivalent to,

$$\sum_{n=1}^{N} \beta(z_{n,k}) \mathbf{C}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{2.19}$$

When we multiply by $\mathbf{C}_k$ we can rearrange the expression to,

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \beta(z_{n,k}) \mathbf{x}_n \tag{2.20}$$

where $N_k = \sum_{n=1}^{N} \beta(z_{n,k})$, can be interpretted as the effective number of points assigned to cluster k. Therefore, $\boldsymbol{\mu}_k$ for the kth Gaussian component is obtained by taking a weighted mean of all the points in the measurent vectors. The weight is given by the posterior probability $\beta(z_{n,k})$ that component k was for generating $\mathbf{x}_n$.
The maximum likelihood solution for the covariance $\mathbf{C}_k$ is found by,

$$\frac{d}{d\mathbf{C}_k} \ln p(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathbf{C}) = 0 \tag{2.21}$$

$$\mathbf{C}_k = \frac{1}{N_k} \sum_{n=1}^{N} \beta(z_{n,k})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{2.22}$$

where each measurement vector also is weighted by the responsibility $\beta(z_{n,k})$. The maximum likelihoog solution for the mixing probability is derived in [33] and it is,

$$\alpha_k = \frac{N_k}{N} \tag{2.23}$$

where $N_k = \sum_{n=1}^{N} \beta(z_{n,k})$. This means that the mixing coefficient for the kth component is given by the average responsibility which the component takes for explaining the measurement vectors. It is now possible to proceed with the EM-algorithm to obtain the maximum likelihood estimate for the particular case of the Gaussian mixture model.

EM-algorithm for the complete measurent vectors

1. Choose an initial value for the parameter vector $\boldsymbol{\theta}^{\text{old}}$.

2. (**E-step**). Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$, by evaluation of the responsibilties of the current parameter values.
$$\beta(z_{n,k}) = \frac{\alpha_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_j \alpha_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \mathbf{C}_j)} \tag{2.24}$$

3. (**M-step**). Evaluate $\boldsymbol{\theta}^{\text{new}}$, re-estimating the parameters using the current probabilities,
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \beta(z_{n,k}) \mathbf{x}_n$$
$$\mathbf{C}_k = \frac{1}{N_k} \sum_{n=1}^{N} \beta(z_{n,k})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$
$$\alpha_k = \frac{N_k}{N}$$
where $N_k = \sum_{n=1}^{N} \beta(z_{n,k})$.

4. Evaluate the log-likelihood
$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \ln p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathbf{C}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{C}_k) \right\}$$
Check for convergence. If no convergence, then update, $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and go to step 2.

We have considered how to use the EM-algorithm to maximize the likelihood, when there are discrete latent variables. This has been derived for the Gaussian mixture model. In the following we will also use the EM-algorithm for finding the the maximum a posterior solutions (MAP). Since our aim is to estimate a given number of clusters and their respective source parameters from the measurent vectors alone, we will use the MAP model. The MAP model adds a prior $p(\boldsymbol{\theta})$ to the log-likelihood expression in (2.10). The prior is defined over the parameters and a suitable choice of the prior will improve the model selection.

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\arg\max} \left\{ \ln p(\mathbf{X}|\boldsymbol{\theta}) \; + \ln p(\boldsymbol{\theta}) \right\} \tag{2.25}$$

Model selection is explained in Section 3. In the following section we will describe the non-probabilistic clustering method of K-means. We use this algorithm for initialization and furthermore, the model selection criteria of Calinski-Harabasz fits well to the K-means algorithm and we will explain this connection also.

2.4 Initialization using *K*-Means Clustering

Given the unlabelled measurent vectors, the aim is to estimate the corresponding unknown parameter vector $\boldsymbol{\theta}$, which can be done using the non-probabilistic method of K-means. Once we have estimated which points go to which cluster, we can estimate a Gaussian mean and covariance for that cluster. It is unlikely that the guess is right the first time, but based on the initial estimates of parameters, it is possible to make a better guess at pairing points with components, in an iterative procedure using the EM-algorithm. We consider the problem of identifying clusters of measurement vectors in a multidimensional space. We observe N observations of the measurent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$. In general the variable $\mathbf{x}$ is D-dimensional. However, we have defined two parameters in this study ($D = 2$). Each cluster center is represented by $\boldsymbol{\mu}_k$ after we have assigned each point in the measurent vectors to a given cluster. The assignment of a measurement vector $\mathbf{x}_n$ to cluster k is described by the binary indicator variable $b_{n,k} \in \{0, 1\}$. The aim is to minimize the sum of squares distance from each measurement vector to its closest center vector $\boldsymbol{\mu}_k$. We can now describe a cost function J as,

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} b_{n,k} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 \tag{2.26}$$

Finding the values of $b_{n,k}$ and $\boldsymbol{\mu}_k$ that will minimize J is done by using an iterative optimization procedure, involving two steps for each iteration. To begin the iterations, some initial values are assigned to $\boldsymbol{\mu}_k$. The two iterative steps are,

- Minimize J with respect to $b_{n,k}$, with $\boldsymbol{\mu}_k$ fixed.
- Minimize J with respect to $\boldsymbol{\mu}_k$, with $b_{n,k}$ fixed. .

The assignment of the nth measurement vector to the closest cluster center can be expressed as,

$$b_{n,k} = \begin{cases} 1, & \text{if } k = \arg\min_j ||\mathbf{x}_n - \boldsymbol{\mu}_j||^2 \\ 0, & \text{otherwise} \end{cases} \tag{2.27}$$

Since the cost function J is a quadratic function of $\boldsymbol{\mu}_k$ we differentiate with respect to $\boldsymbol{\mu}_k$ and set it to zero,

$$\frac{d}{d\boldsymbol{\mu}_k} J = 2 \sum_{n=1}^{N} b_{n,k}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{2.28}$$

and solve for $\boldsymbol{\mu}_k$,

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} b_{n,k}\mathbf{x}_n}{\sum_{n=1}^{N} b_{n,k}} \tag{2.29}$$

which expresses that $\boldsymbol{\mu}_k$ is the mean of all measurement vectors $\mathbf{x}_n$ assigned to cluster k. The iteration over these two steps are guaranteed to reach convergence. The

K-means assigns every measurement vector uniquely to one cluster, and it is not clear that a measurement vector which is placed midway between two cluster centers is assigned appropriately, but by using probabilistic models such as the Gaussian mixture model (GMM), the assignments can reflect this level of uncertainty. For initialization of the EM-algorithm by deliberately overfitting, i.e. choosing a K much larger than the expected value, the K-means algorithm assures that the true parameter are among the estimates, making it convenient to use it for initialization of the GMM-EM algorithm before applying the MAP model selection to the GMM-model.

Model selection using K-means

It is possible to evaluate the k-means for different number of clusters and then choose the optimal number of clusters based on the variance ratio criterion [34]. The variance ratio criterion (VRC) is based on the ratio between the overall between-cluster variance and the overall within-cluster variance. We run a short experiment with 7 sources. From the scatter plot in Figure 2.1, we can see that the correct number of clusters have been found in this specific case. However, the K-means clustering algorithm can be stuck in a local minimum rather than the global and it is therefore dependent on the initialization to be well considered. An initialization of the K-means have been proposed as the K-means++ algorithm by [35], a variant that chooses centers at random from the measurement vectors, but weighs the measurement vectors according to their squared distance, squared from the closest center that has already been chosen. This gives a faster convergence and overcomes some of the local minimum problems. Although the K-means clustering algorithm offers no accuracy guarantee, its simplicity

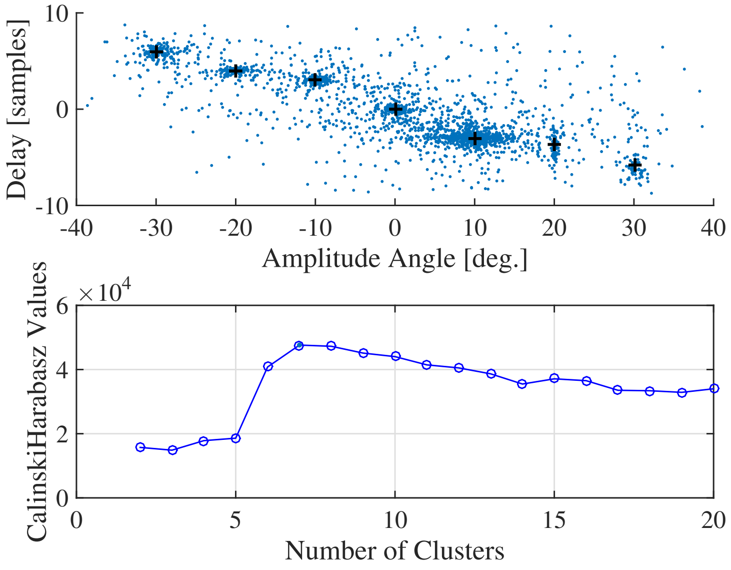


Figure 2.1: Model selection using the Calinski Harabasz criterion on mixture of seven sources.

is very appealing in practice, thus it is widely used for clustering.

Calinski Harabasz Evaluation

The selected clusters in the measurent vectors shown with black plus signs in Figure 2.1, were subjected to a cluster validation algorithm called the Calinski-Harabasz [34] or the Variance Ratio Criterion (VRC), which is similar to the Inter-Intra class distance [36]. The validation algorithm selects the subset of clusters that maximizes the cluster separability. It is based on the Euclidean distance measure between measurement vectors in the measurent vectors. The assumption of mutually exclusive clusters leads to the assumption that the expectation vectors of the different cluster centroids are discriminating [36]. The optimal measure is a monotonically increasing function of the distance between expectation vectors and an increasing function of the scattering around the expectatations. The conditional expectation of the measurement vectors given the cluster is the sample mean $\hat{\boldsymbol{\mu}}_k$:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{x}_{k,n} \tag{2.30}$$

where $\mathbf{x}_{k,n}$ are measurement vectors from cluster $\mathbb{C}_k$. The unconditional expectation of the measurement vector $\mathbf{x}$ is the sample mean of the full measurent vectors $\hat{\boldsymbol{\mu}}$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N_s} \sum_{n=1}^{N_s} \mathbf{x}_n \tag{2.31}$$

where $N_s = \sum N_k$ is all samples in the set. The scattering of vectors from a given class $\mathbb{C}_k$ is:

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} (\mathbf{x}_{k,n} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_{k,n} - \hat{\boldsymbol{\mu}}_k)^T \tag{2.32}$$

It is analogous to a covariance matrix. The scatter matrix describing the noise is called the within-scatter matrix $\mathbf{S}_w$. Averaged over all classes it describes the average scatter within classes.

$$\mathbf{S}_w = \frac{1}{N_s} \sum_{k=1}^{K} N_k \mathbf{S}_k \tag{2.33}$$

Complementary to the within-scatter $\mathbf{S}_w$ is the between scatter matrix $\mathbf{S}_b$ that describes the scattering of class dependent sample means around the overall average:

$$\mathbf{S}_b = \frac{1}{N_s} \sum_{k=1}^{K} N_k (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^T \tag{2.34}$$

With these definitions we can express the Calinski-Harabasz criterion as,

$$\mathrm{CH}_k = \frac{\mathbf{S}_b}{\mathbf{S}_w} \frac{N_s - k}{k - 1} \tag{2.35}$$

To determine the optimal number of clusters, we maximize CH_k with respect to k. The optimal number of clusters is the solution with the highest Calinski-Harabasz index value. The rightmost fraction of (2.35) is different from the inter-intra class distance ratio of [36]. Basically, this fraction expresses that we maximize the criterion by explaining large amount of observations by few clusters. The scatter within and scatter between ratio can be regarded as a signal to noise ratio, but it does not alter the underlying tendency of the *K*-means clustering algorithm, to overfit the measurement space and get stuck in a local miximum. Therefore, it seems reasonable to use the probabilistic method for cluster evaluation and only the *K*-means for initialization. Figure 2.2 shows the correct estimated model order, however the *K*-Means clustering algorithm is stuck in a local minimum and has therfore missed one of the true clusters. Since it is a strong criterion for non-probabilistic model order selection which fits well to the *K*-means clustering, which is a minimizer of the squared error, we will test its ability to be used for segmentation in Section 5.1.2, only then we normalize the Calinski-Harabsz criterion to the measurement space as,

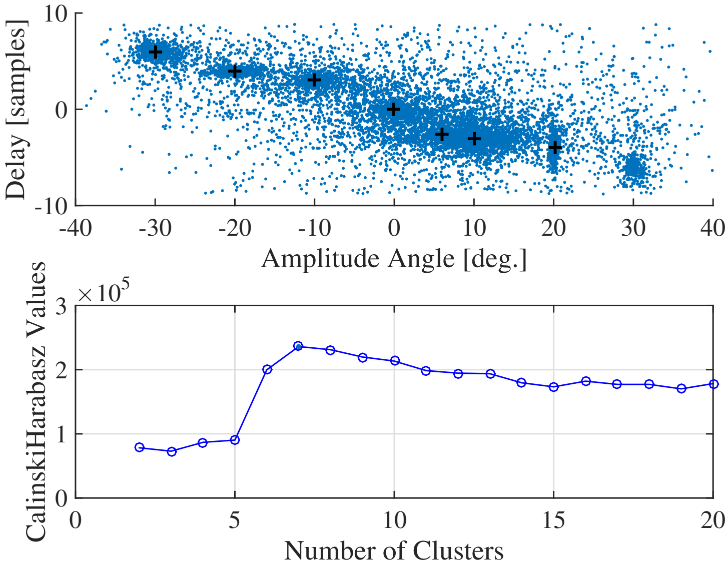


Figure 2.2: Model selection using the Calinski Harabasz criterion on a mixture of seven sources. In this specific evaluation the *K*-means is overfitting to the measurements compared to Figure 2.1.

$$\mathrm{CH}_k = \left(\frac{\mathbf{S}_b}{\mathbf{S}_w} \frac{N_s - k}{k - 1} \right) \frac{1}{N_s} \tag{2.36}$$

The normalized objective functions of the two given examples of Figure 2.1 and 2.2 are shown in Figure 2.3.

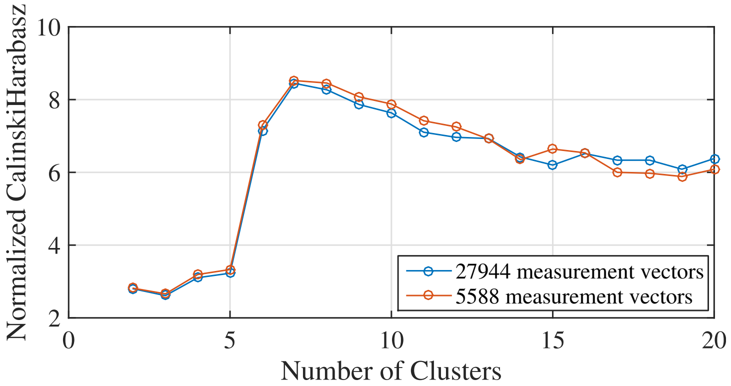


Figure 2.3: The normalized Calinski Harabasz objective functions of Figure 2.1 and 2.2.

Chapter 3

Model Order Selection

One advantage of the mixture model approach to clustering is that it allows the use of approximate Bayes factors to compare models. A thorough comparison of Bayes factors can be read in [37]. The model order selection and the segmentation can be done with a *maximum* a posteriori (MAP) criterion. The MAP estimator is,

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \{\ln p(\mathbf{x}|\boldsymbol{\theta}) \ + \ln p(\boldsymbol{\theta})\} \tag{3.1}$$

where $p(\theta)$ is the prior on the parameters and $\mathbf{x}$ is the observed data. We will introduce the MAP criterion in the following.
There exists several approaches for finding a solution to the model order estimate. Two of these are very often used [38, 39], namely the AIC and the MDL, the latter of which formally coincides with the BIC [40]. In the following we will describe the MDL as a special case MAP criterion in the following section. The AIC is given as [39],

$$\mathcal{M}_s = \arg\min_{\mathcal{M}_k} \{-\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M}_k) + N_p\} \tag{3.2}$$

The MDL is,

$$\mathcal{M}_s = \arg\min_{\mathcal{M}_k} \left\{ -\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M}_k) + \frac{N_p}{2} \ln N \right\} \tag{3.3}$$

where $\mathcal{M}_s$ is the selected model, $\mathbf{x}$ is the observed measurement vector, $p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M}_k)$ is the probability density function of the data given the model parameters and the model, $\boldsymbol{\theta}$ is the parameter vector and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood of $\boldsymbol{\theta}$ and N_p is the dimension of $\boldsymbol{\theta}$.

3.0.1 The Asymptotic MAP criterion

The principle of the MAP is choosing the model $\mathcal{M}$ that maximizes the posterior probability given the observed data $\mathbf{x}$,

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathcal{M}|\mathbf{x}) \tag{3.4}$$

expressed by using Bayes method,

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} \frac{p(\mathbf{x}|\mathcal{M})p(\mathcal{M})}{p(\mathbf{x})} \tag{3.5}$$

Choosing a uniform prior $p(\mathcal{M})$ to not favour any model beforeh and noting that once $\mathbf{x}$ is observed $p(\mathbf{x})$ is constant and the MAP model reduces to the likelihood of the observed data given the model,

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} p(\mathbf{x}|\mathcal{M}) \tag{3.6}$$

where the likelihood is dependent on the parameters, $\boldsymbol{\theta}$. In the Bayesian framework we obtain the marginal density of the measurents given the model, by integrating the parameters out [38],

$$p(\mathbf{x}|\mathcal{M}) = \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta} \tag{3.7}$$

The asymptotic approximation to this integral is found by assuming high amounts of data, when the most significant peaks occur in the likelihood function around the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$. (3.7) becomes equal to [39],

$$p(\mathbf{x}|\mathcal{M}) = (2\pi)^{N_p/2} \det\left(\widehat{H}\right)^{-1/2} p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M})p(\hat{\boldsymbol{\theta}}|\mathcal{M}) \tag{3.8}$$

where $\widehat{H}$ is the Hessian of the log-likelihood function when evaluated at the $\hat{\boldsymbol{\theta}}$,

$$\widehat{\mathcal{H}} = -\left.\frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{3.9}$$

By neglecting terms of order $\mathcal{O}(1)$, the assymptotic MAP expression is found by taking the negative logarithm of (3.8), where the term $2\pi^{N_p/2}$ can be assumed constant for asymptotic signal length N, while a weak prior on $p(\boldsymbol{\theta}|\mathcal{M})$ has been used [39] to obtain the MAP expression [38],

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} \left\{ -\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M}) + \frac{1}{2}\ln \det(\widehat{\mathcal{H}})\right\} \tag{3.10}$$

where the first term is the log-likelihood and the last term is the penalty added. The first term of the criterion decreases when the complexity of the model increases, and by contrast, the second term increases and acts as a penalty for using additional parameters to model the data. The penalty term is found by noting that the Hessian in (3.9) can be replaced by the Fisher information matrix since the error it introduces is smaller than the neglected terms of order $\mathcal{O}(1)$ [38, 39]. The Hessian is then,

$$\widehat{\mathcal{H}} \approx -\mathrm{E}\left.\left\{\frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right\}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{3.11}$$

Under the assumptions of the observed measurement being real, independent and identically distributed, we can write,

$$\det \widehat{\mathcal{H}} = \mathcal{O}(N^{\frac{N_p}{2}}) \tag{3.12}$$

The interested reader can find specific details on this assumption in [39]. The expression in (3.10) then reduces to,

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} \; -\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}, \mathcal{M}) + \frac{N_p}{2} \ln N \tag{3.13}$$

which is the MDL that formally coincides with the BIC. For the case of the multivariate Gaussian distribution with arbitrary covariance $N_p = d + d(d+1)/2$, where d is the dimensionality of the feature space. The expression in (3.13) is not valid for all signal processing families of models. In fact, for the Gaussian mixture model this rule will not be directly appropriate for model order selection without applying priors to the parameters as decribed in Section 3.0.2. Figure 3.1 and 3.2 shows the AIC and the asymptotic MAP, refered to as the BIC under the described assumptions. Both figures show the criterion applied to a mixture of seven sources from the SQAM database, using the MAP criterion implemented by the EM-algorithm. From both of the figures it is clear that the criterion results in a monotonically decreasing function of the model order. This tendency to overfitting is due to the fact that the measurement vectors does not have equal weight in each parameter estimate in the Gaussian mixture model.

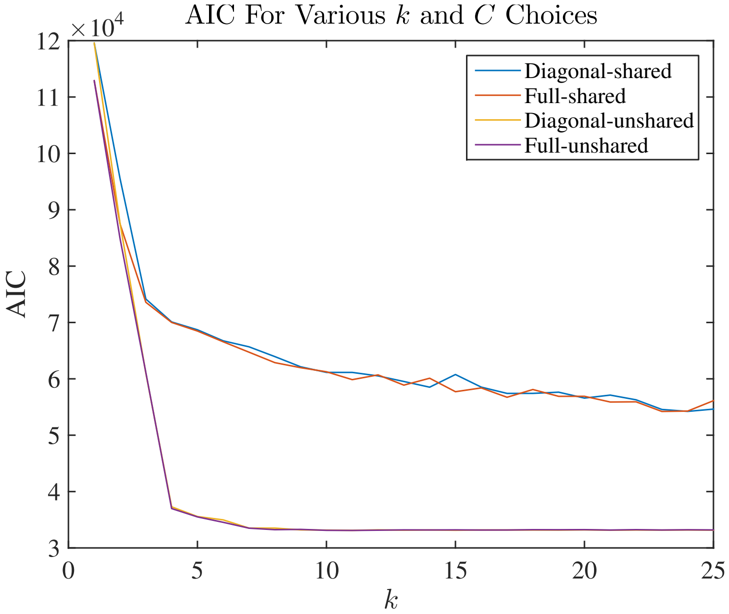


Figure 3.1: AIC, as a function of model order for fitting a mixture of seven sources with the Gaussian mixture model.

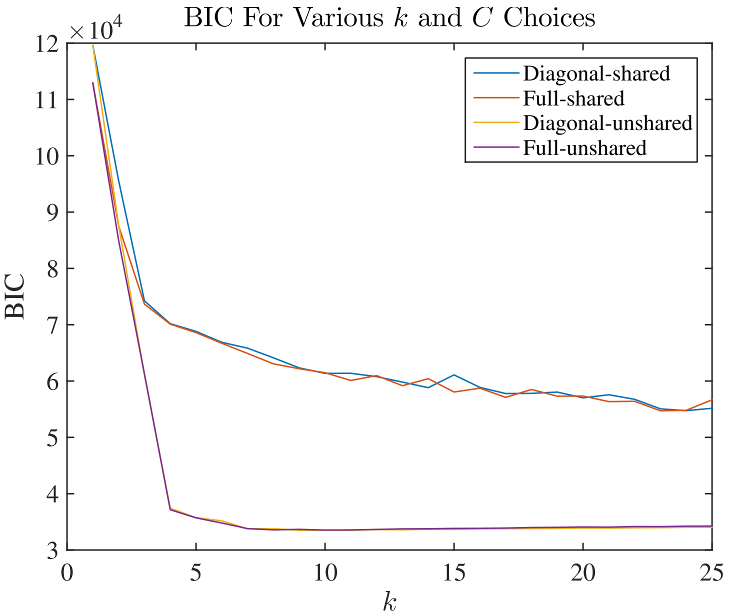


Figure 3.2: BIC, as a function of model order, for fitting a mixture of seven sources with the Gaussian mixture model.

The penalty term that is dependent on N can be altered to become appropriate for a Gaussian mixture model, which will be described in the following section.

3.0.2 Suitable Prior on the Mixing Probabilities

With suitable priors on the parameters, the MAP estimator can be used for model selection. In particular, [21] and [22] put the Dirichlet prior on the mixing probabilities, of the components in the Gaussian mixture model, and [41] applied the "entropic prior" on the same parameters to favor models with small entropy. All of these have in common that they used the MAP estimator to drive the mixing probabilities associated with unnecessary components toward extinction. Based on an improper Dirichlet prior, [30] suggested to use minimum message length criterion to determine the number of the components, and further proposed an efficient algorithm for learning a finite mixture from multivariate data which we have adopted for source estimation based on panning parameters. It is the model called mixture-MDL (MMDL), which is described in the following section.

3.1 Mixture MDL (MMDL)

If we recall the parameter vector of the Gaussian mixture model as,

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k, \alpha_1, \alpha_2, \ldots, \alpha_k\}$$

Once we have estimated one source parameter i.e. $\hat{\boldsymbol{\theta}}_k$ the sample size "seen" by this parameter is $N\alpha_k$ due to the mixing probability weighting [30]. The penalty term becomes dependent on not only the number of measurement vectors N, but also on the mixing probabilities α. The full derivation of the MMDL criteria is derived in [30], where it is described how the Fisher information for $\boldsymbol{\theta}_k$ for one observation from component k becomes $N\alpha_k\mathcal{I}(\boldsymbol{\theta}_k)$. The prior on the parameters of the asymptotic MAP expression for mixtures is,

$$p(\boldsymbol{\theta}_k) = \frac{k(N_p+1)}{2}\ln N + \frac{N_p}{2}\sum_k \ln(n\alpha_k) \tag{3.14}$$

We will adopt this criterion from [30]. The mixture-MDL is,

$$\hat{\boldsymbol{\theta}}_{k\text{MMDL}} = \underset{\boldsymbol{\theta}_k}{\arg\min}\ \{-\ln p(\mathbf{x}|\boldsymbol{\theta}_k)\ + p(\boldsymbol{\theta}_k)\} \tag{3.15}$$

The key observation of the MMDL is that the prior $p(\boldsymbol{\theta}_k)$ is not only a function of k and for a fixed k it is not a ML estimate. For fixed k, MMDL has a simple Bayesian interpretation [30]:

$$p(\{\alpha_1,\ldots,\alpha_k\}) \propto \exp\left\{-\frac{N_p}{2}\sum_{k=1}^{K}(\alpha_k)^{\frac{N_p}{2}}\right\} \tag{3.16}$$

Which is a Dirichlet-type improper prior, which can be used on the mixing probability in the maximum a posteriori (MAP) estimator for model selection.

The procedure is then as follows: We start with a large number of randomly initialized components and search for the MAP solution using the iterative procedure of the EM algorithm. The prior drives the irrelevant components to extinction. In this way, while searching for the MAP solution, the number of components is reduced until convergence is achieved. See [42] for details on Dirichlet type prior relation to the standard MDL. The MMDL minimization criteria and the complete algorithm that is implemented is described in [30],

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min}\ \mathcal{L}(\boldsymbol{\theta}|\mathcal{X}) \tag{3.17}$$

with

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{X}) = \underset{\boldsymbol{\theta}}{\arg\min}\left\{-\ln p(\mathbf{X}|\boldsymbol{\theta}) + \frac{N_p}{2}\sum_{k=1}^{K}\ln\frac{N\alpha_k}{12} + \frac{k}{2} + \ln\frac{N}{12} + \frac{k(N_p+1)}{2}\right\} \tag{3.18}$$

where $\alpha_k > 0$ and N_p is the number of parameters specifying each component. The MMDL mixture model is including the component-wise EM algorithm CEM^2 [43].

The expected number of measurement vectors generated by the kth component of the mixture is $N\alpha_k$, which is the sample size seen by the $\boldsymbol{\theta}_k$, thus the optimal (in the MDL sense) for each $\boldsymbol{\theta}_k$ is $N_p/2\log(N\alpha_k)$ [30]. The MMDL promotes sparseness in the sense that it is intialized with much higher k than expected, and the EM-MMDL will then set some $a_k = 0$ by killing the weakest component and then restart the CEM^2 algorithm [43].

3.2 Model Pruning by Component Annihilation

The following section is the proposed method for component annihilation for stereo-panning estimation. In the following description, this method is described as a post-processing procedure. However, it is desirable to implement the functionality of this method as part of the likelihood-model in the segmentation algorithm which still remains unsolved. In the end ef this section the model pruning will be described as a Bayesian interpretation.

3.2.1 Overfitting a Gaussian Mixture Model to Panning Parameter Space

The challenge of overfitting, a Gaussian mixture to the distribution space is ambiguous. It is the case that the MMDL will estimate a model order k that is equal to or larger than the true order. However, the GMM-model is designed to describe every single measurement vector as being part of a Gaussian distribution. The ambiguouity is that the overfitting of the Gaussian mixture model can be exploited for the initialization of the EM-algorithm, which is also the case for the algorithm of MMDL [30]. By starting with k, where k is much larger than the true/optimal number of mixture components, the adopted algorithm is robust with respect to initialization of the EM-algorithm. The MMDL algorithm applies component annihilation, by adopting a Dirichlet prior on the mixing probabilities [44, 30], and selects the number of components by annihilating the weakest component in the M-step of an iterative component-wise EM (CEM^2) [43]. This procedure leads to a smaller model order and still describes every measurement vector as being part of a Gaussian distribution. It is important to notice that every true parameter is then described by at least one or more of the clusters.

3.2.2 Component Metrics for Model Pruning

After model order selection has been applied using the Mixture-minimum description length algorithm, each true panning parameter vector is described by one or more components. Therefore, we have applied a post-processing step to select the true number of clusters from prior spatial knowledge of the covariance in each conditional distribution. In the following this procedure is described, starting with the practical view and lastly we interpret the model pruning as a Bayesian posterior.

The model pruning post-processing step selects clusters from an analysis on each cluster covariance compared to the number of estimated points embedded in each cluster. We know from [18] that due to the non-disjoint spectral overlap of sources, the variance increases in the amplitude panning direction. Therefore, we propose to select clusters with largest amount of estimated points, relative to the size of their respective embedding covariances and their rotational angle in the parameter space. We describe this for the kth covariance $\mathbf{C}_k$ in the following and in the following we will interpret it as a Bayesian posterior. Because the panning parameters are two dimensional, we define $\mathbf{C}_k$ geometrically as an ellipsoid by applying the singular value decomposition as,

$$\mathbf{C}_k = \mathbf{U\Sigma V}^T \tag{3.19}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal rotation matrices and the diagonal of $\mathbf{\Sigma}$ contains the principal axes a^2 and b^2. We compute the angle θ of the principal axes a to the x-axis,

$$\theta = \tan\left(\mathbf{u_2}/\mathbf{u_1}\right)^{-1} \tag{3.20}$$

we center $\mathbf{C}_k$ by subtracting the mean $\boldsymbol{\mu}_k$ as $\mathbf{d} = \mathbf{C}_k - \boldsymbol{\mu}_k$. The x and y coordinates of each estimate is given as,

$$(x, y) = (\mathbf{d}_1 \cos\theta + \mathbf{d}_2 \sin\theta, -\mathbf{d}_1 \sin\theta + \mathbf{d}_2 \cos\theta) \tag{3.21}$$

We count the number of points inside cluster k. The specific point (x, y) is inside the ellipse k if,

$$\frac{x^2}{a_k^2} + \frac{y^2}{b_k^2} < 1 \tag{3.22}$$

Lastly we compute the size of $\mathbf{C}_k$ as the determinant of $\mathbf{C}_k$ and we compute the "shadow" of the covariance on the x-axis as

$$s_k = a_k \cos\theta_k + b_k \sin\theta_k \tag{3.23}$$

3.2.3 Annihilation Steps

The following component annihilation steps uses the source parameters after the MMDL has been applied and by comparing these to the measurements, the true clusters are selected and the rest of the measurement vectors will be removed. In this initial implementation, implemented as a post processing model pruning algorithm, we have two rules applied which is:

1. If a cluster shares an estimated point with a smaller cluster, all points that is only part of the bigger cluster is removed. The overlapping bigger clusters are refered to as sticky clusters.

2. A geometric threshold is applied based on $\det(\mathbf{C}_k)$, the number of points embedded in $\mathbf{C}_k$ and the shadow s on the x-axis as described by (3.2.4) and (3.2.4).

The two steps described as rule 1 and rule 2, is always carried out with step 1 first. Step 1 removes every "sticky" cluster. Once the sticky clusters have been removed, each remaining clusters is measured with the ratio described by (3.2.4) and (3.2.4). An example of these two steps are shown in Figure 3.3. It is noticeable that only by removing the sticky clusters we have reduced the number of clusters from 14 to 5, when the true numbers of clusters is 4. All clusters shown as ellipsoids which are embedding estimates in blue are the sticky clusters which are now ignored. It is easy to see that one of the orange clusters has a lower density than the remaining 4 clusters. In order to remove clusters with relative low density and high correlation between parameters, we apply step 2. Another point to notice is that the one of the low density clusters also differs from the remaining in the angle of its principal component, which shows a relative higher correlation between the two given features, thus it has a greater variance in both directions since it spans a larger region, but especially the variance in the amplitude panning direction is interpretted a sign of non-disjoint orthogonality in the source mixture [18]. Figure 3.5 shows the ratio function of (3.2.4). We note that between $k = 4$ and $k = 5$ there is a ratio difference on the order of 10^{20}, which often is sufficient for a fixed thresholding. Figure 3.5 shows the parameter space after step 2 has been applied. It can be seen that the 4 true clusters now has been found after model pruning. Lastly, the full Gaussian mixture is shown in Figure 3.6. From this figure it can be seen that the problem is not a convex one, and there is at least 6 local minimum. It is worth noticing that one of the true clusters has a lower peak

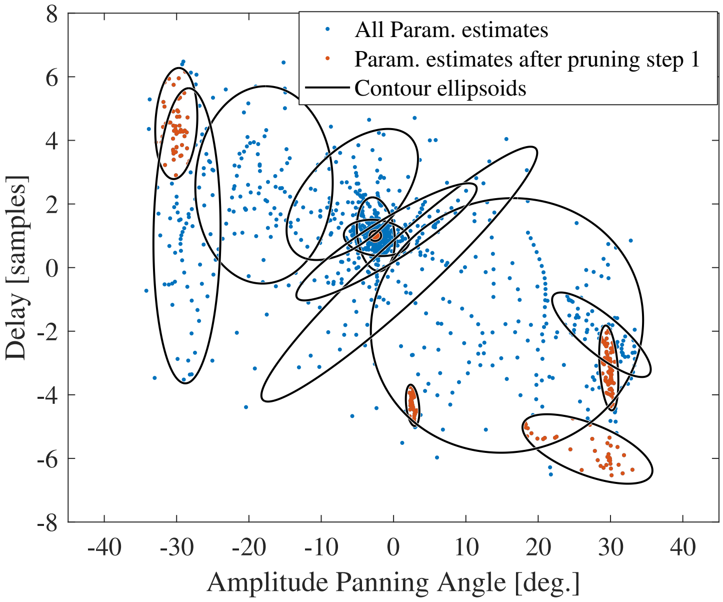


Figure 3.3: Component annihilation step 1 has been applied to the parameter estimates of a mixture of 4 sources. All ellipsoids represent a cluster. The ellipsoids containing orange dots are not sticky and are kept. All blue clusters will be ignored following step 1.

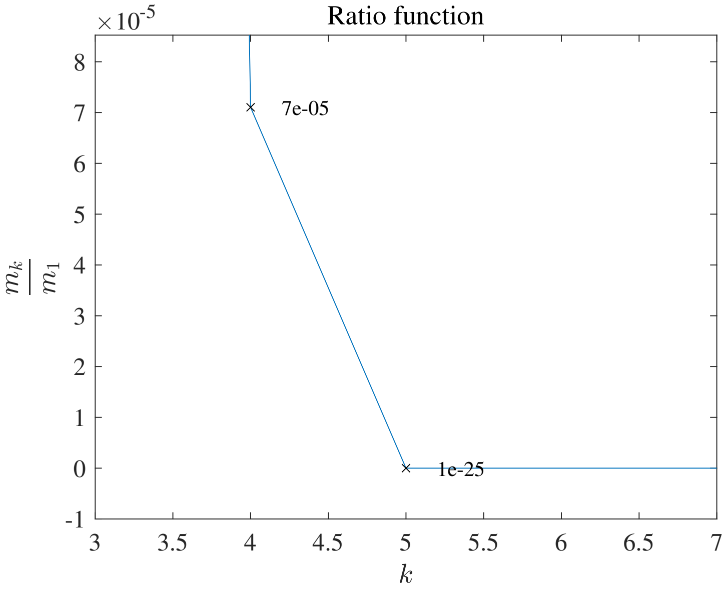


Figure 3.4: The ratio as a function of number of estimated clusters (3.2.4). It has been applied to the mixture of 4 sources.

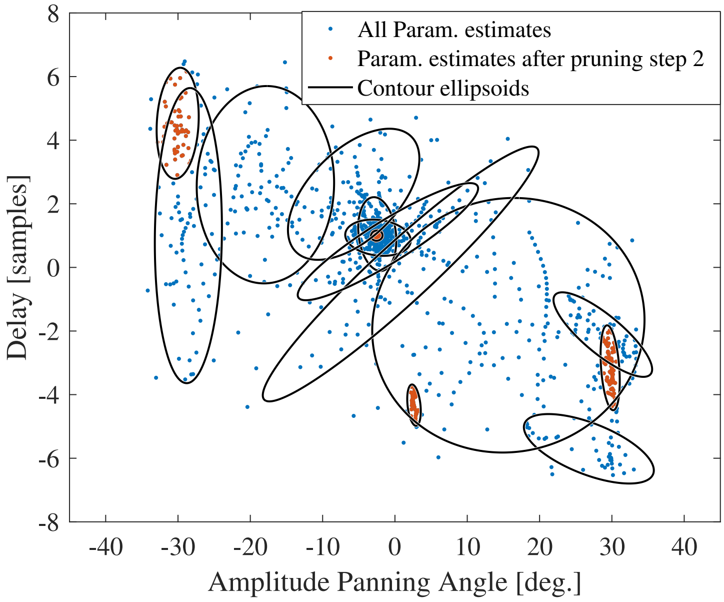


Figure 3.5: Component annihilation step 1 and step 2 has been applied to the parameter estimates of a mixture of 4 sources. All ellipsoids represent a cluster. The ellipsoids containing orange dots are kept after model pruning.

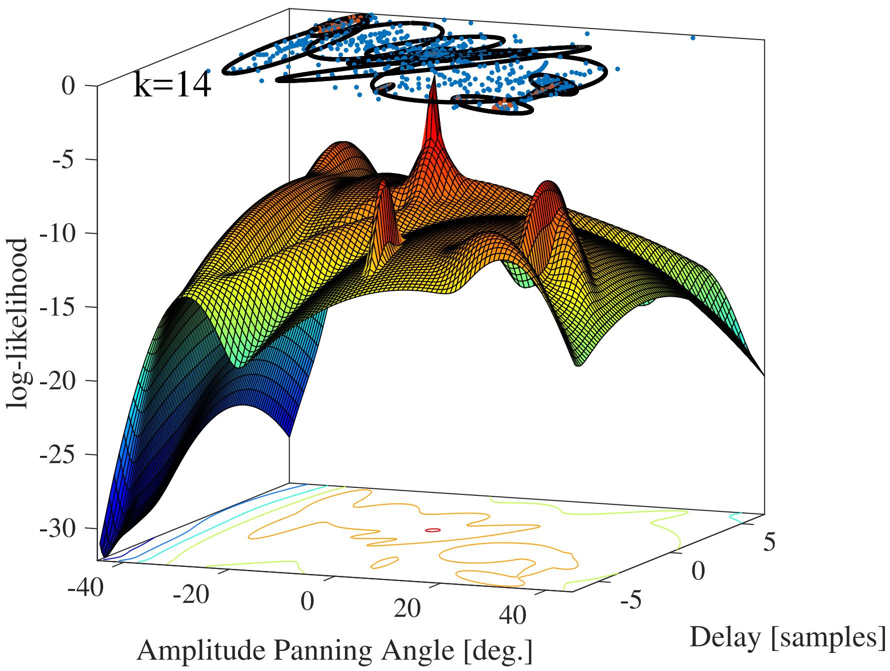


Figure 3.6: The estimated Gaussian mixture. This mixture contains measurement vectors from 4 sources from the SQAM database.

likelihood value. Thus, it has relative low mixing probability and spans a larger region than the other true cluster estimates. The procedure of model pruning by component annihilation has kept this specific cluster because it has a low variance in the amplitude direction when comparing to the other clusters, by using (3.2.4) and it is still embedds many meaurement vectors since it has been kept after the ratio function of (3.2.4) has been applied. In the following is a description of the metrics that has been defined for this component annihilation criteria.

3.2.4 Thresholding on the Cluster Angle and size

The spectral overlap of mixture sources magnifies the variance in the amplitude panning direction [18] . In the case of disjoint orthogonality, the covariance would be very small and diagonal or have greatest variance in the delay-direction. Therefore we can apply a threshold from the rotation angle θ and the size of the region relative to the given number of estimates in the region. We define a variable $0 \leq p_k \leq 1$ which is the percentage of points that is inside the kth cluster. We notice that the mixing parameter **ff** is proportional to $\frac{p}{\det(\mathbf{C})}$. We define a metric of peakiness m for the kth source as,

$$m_k = \frac{p_k}{\det(\mathbf{C}_k)s_k} \tag{3.24}$$

where s_k is the amplitude shadow, $s_k = a_k \cos\theta_k + b_k \sin\theta_k$. The metric m_k carries implicit information of both the size and angle of the kth cluster region, that includes a percentage of all the estimated points (after sticky clusters have been removed). From all metrics $\mathbf{m}$ we define a threshold where m_k is relative to the smallest $m = m_1$. The metric ratio is,

$$ratio_k = \frac{m_k}{m_1} \tag{3.25}$$

Through experiments, we have found that this method of component annihilation has good performance for precisely estimating the number of sources in the mixture and the panning parameters. In the following we will interpret this with Bayesian terminology as a posterior probability by comparing it to the K-nearest neighbour classifier.

3.2.5 Bayesian Interpretation of the Metrics for Annihilation

In the following we describe the model pruning method as a posterior probability, by comparing it to the K-nearest neighbour technique. To do this, we make use of Bayes theorem and apply the K-nearest neighbour method for classification to each cluster separately. Let us suppose that we have a data set with N samples, where N_k points belongs to class $\mathbb{C}_k$, so that $\sum_k N_k = N$. If we wish to classify a point $\mathbf{x}$ with the K-nearest neghbour method, we draw a hypersphere that is centered on $\mathbf{x}$, containing K points irrespective of their class. Suppose this sphere has volume $V(\mathbf{x})$ and contains K_k points belonging to class $\mathbb{C}_k$. An estimate of the density associated with each class is then [33, 36],

$$\hat{p}(\mathbf{x}|\mathbb{C}_k) = \frac{K_k}{N_k V(\mathbf{x})} \tag{3.26}$$

Similarly, the unconditional density is given by,

$$\hat{p}(\mathbf{x}) = \frac{K}{N V(\mathbf{x})} \tag{3.27}$$

and the class priors are given by,

$$\hat{p}(\mathbb{C}_k) = \frac{N_k}{N} \tag{3.28}$$

We can combine these three equations using Bayes' theorem to obtain the posterior probability of class membership

$$\hat{p}(\mathbb{C}_k|\mathbf{x}) = \frac{\hat{p}(\mathbf{x}|\mathbb{C}_k)\hat{p}(\mathbb{C}_k)}{\hat{p}(\mathbf{x})} = \frac{K_k}{K} \tag{3.29}$$

Which means that we can minimize the risk of misclassification, by assigning the point $\mathbf{x}$ to the class having the largest posterior probability corresponding to K_k/K. Such a classification can be expressed as:

$$\hat{\mathbb{C}}_k(\mathbf{x}) = \mathbb{C}_k \quad \text{with} \quad k = \underset{i=1,\dots,K}{\arg\max}\left\{\hat{p}(\mathbf{x}|\mathbb{C}_k)\hat{p}(\mathbb{C}_k)\right\}$$
$$= \underset{i=1,\dots,K}{\arg\max}\left\{\frac{K_i}{N_i V(\mathbf{x})}\frac{N_i}{N_{\text{total}}}\right\} = \underset{i=1,\dots,K}{\arg\max}\left\{\frac{K_i}{N_{\text{total}} V(\mathbf{x})}\right\} \tag{3.30}$$

We can compare this expression to the threshold in (3.2.4) of the model pruning method, where $m_k = \frac{p_k}{\det(\mathbf{C}_k)s_k}$ with p_k the percentage of points inside cluster k and the volume of the hypersphere determined by $\det(\mathbf{C}_k)$ which is weighted relative to the shadow s_k, to prefer the clusters that have lowest variance in the amplitude direction. In Figure 3.2.5 showing the parameter space it is clear that the covariance assumption described above in Section 3.2.3 holds and all 7 true clusters have been estimated correctly. The selected clusters are shown in orange and the unselected clusters are shown in blue. The true cluster covariances are either very small and close to diagonal or they are larger and have dominant variance mainly in the delay direction (upwards). We notice that the MMDL algorithm in this case chose a $k = 18$. Notice that the 11 "wrong clusters" are large with random covariance structure and rotation angle. Figure 3.9 shows the ratio function in (3.2.4). We note that between $k = 7$ and $k = 8$ there is a ratio difference on the order of 10^{18}, which often is sufficient for a fixed threshold.

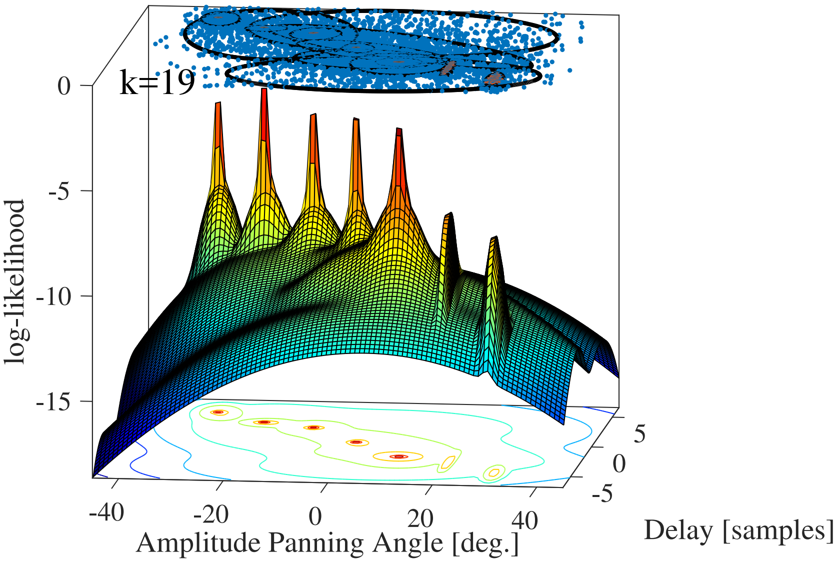


Figure 3.7: Gaussian mixture of 7 sources. From this figure it can be noticed that the two right most components has a low mixing probability.

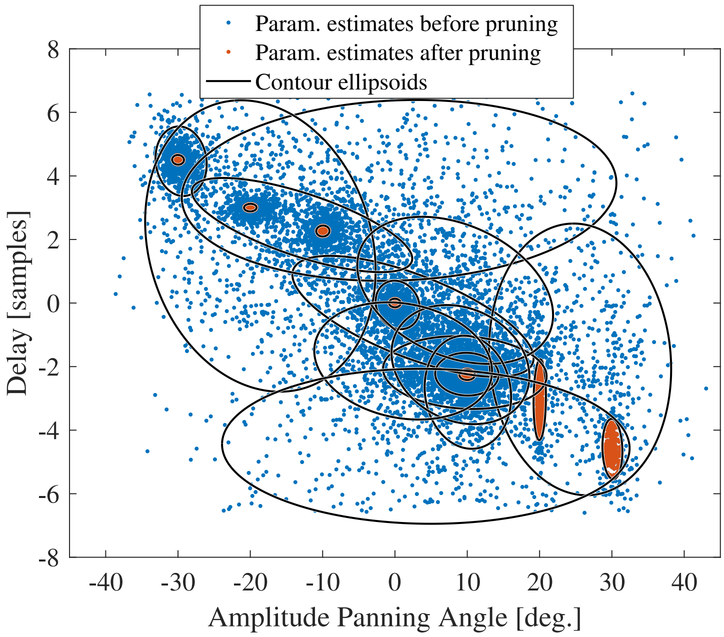


Figure 3.8: Before and after pruning.

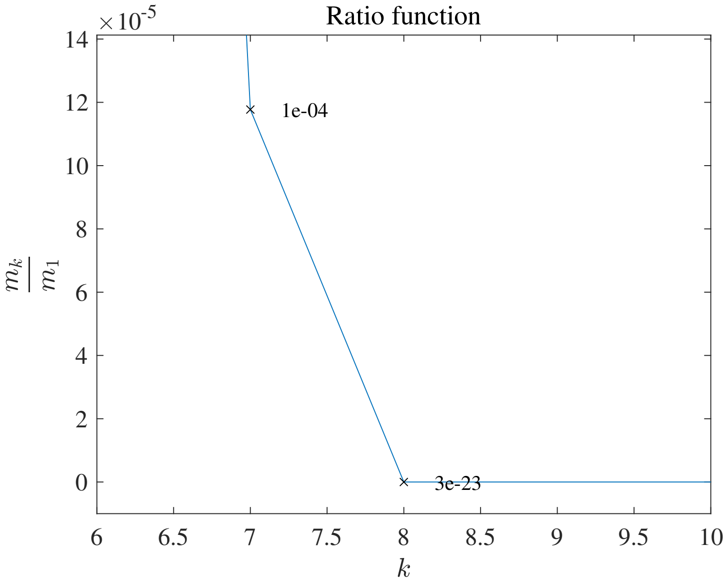


Figure 3.9: Ratio function of 7 sources shown in Figure 3.7

Chapter 4

Segmentation of the Stereophonic Signal

4.1 Optimal Time Segmentation for Signal Modelling

In the following we describe the optimal time segmentation scheme which we propose for source parameter estimation based on the MAP-model clustering algorithm. The goal is to achieve better segmentations of time, entailing a better local model. Once the chosen signal modelling technique can be quantified as a cost function, that is additive over distinct segments, a time segmentation based on [45] guarantees the global optimality of the scheme.

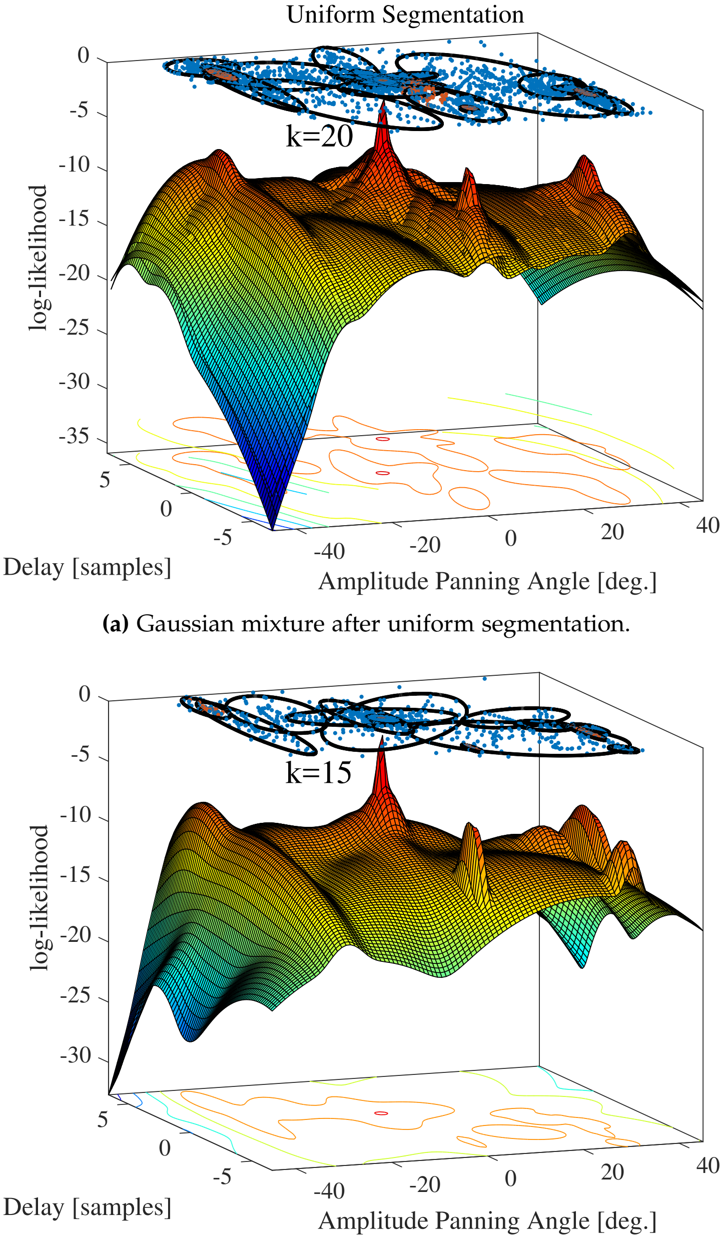
4.1.1 Segmentation of the Stereophonic Signal

The characteristics of the observed stereophonic signal are varying over time with different durations. Consequently, a fixed segment length is not optimal for the MAP-clustering model. Using the MAP criterion, the cost associated with the different outcomes from the set of segment lengths is additive and can be compared, and the optimal can be chosen as the one that minimizes the cost (3.1). The implementation of the segmentation scheme is based on a dynamic programming in [46, 47, 48]. The implemented algorithm is outlined in the algorithm showed in Figure 4.1. A minimal segment length, N_{min} generating a block of samples and dividing the signal into M blocks. Since this will give 2^{M-1} ways of segmenting the signal into M blocks a maximum number of blocks K_{max} is defined to ease on computational complexity, since very high segment length is assumed to be generating noise in the distribution. The maximum number of samples in one segment is $N_{\text{max}} = K_{\text{max}} N_{\text{min}}$. A dynamic programming algorithm, computes the optimal segment length k_{opt} for all blocks, $m = 1, \ldots, M$, starting at $m = 1$ moving continously to $m = M$. For every block, the cost of all new block combinations are reused from earlier blocks. When the end of the signal is reached, the optimal segmentation of the signal is found, starting with the last block and continuing through the signal to the beginning. Starting at $m = M$, setting the number of blocks in the last segment to $k_{\text{opt}}(M)$. The next segment ends at block $m = M - k_{\text{opt}}(M)$ and includes $k_{\text{opt}}(M - k_{\text{opt}}(M))$ blocks. This is continued untill $m = 0$.

```
while $m \times N_{\text{MIN}} \leq$length(signal) do
  Initialize $K = \min([m, K_{\max}])$.
  for $k = 1$ to $K$ do
    block of signal to use is $m - k + 1, \ldots, m$
    estimate $(\hat{\gamma}, \hat{\delta})$ from (1.4.1) and (1.4.1)
    compute $\mathcal{L}(\boldsymbol{\theta}|\mathbf{X})_{(m-k+1)m}$ from (3.1)
    if $m = 1$ then
      $\mathcal{L}(\boldsymbol{\theta}|\mathbf{X})_{(k)} = \mathcal{L}(\boldsymbol{\theta}|\mathbf{X})_{(m-k+1)m} + \mathcal{L}(\boldsymbol{\theta}|\mathbf{X})_{1(m-k)}$
    else
      $\mathcal{L}(\boldsymbol{\theta}|\mathbf{X})_{(k)} = \mathcal{L}(\boldsymbol{\theta}|\mathbf{X})_{(m-k+1)m}$
    end if
  end for
end while
$m = M$
while $m > 0$ do
  number of blocks in segment is $k_{\text{opt}}(m)$
  $m = m - k_{\text{opt}}(m)$
end while
```

Figure 4.1: Optimal segmentation algorithm, based on the MAP-cost function.

Figure 4.2 - Figure 4.4 shows the Gaussian mixture that has been evaluated at each point in the MAP log-likelood function of (3.1). Each of these figures is a representation of the clusters that we have modelled and what is shown in these is firstly the 3D representation of the mixture with a contour on the "ground" plane. In the top of each of these figures, the measurement vectors are shown in 2D, along with the ellipsoid that we use in the algorithm for model pruning and also for the Gaussian moddeling in general. What is noteceable from Figure 4.2 is to the left a more scatterd mixture form than to the right. The reason for this is that Figure 4.2a is the modelled mixture with uniform segmentation, while Figure 4.2b is modelled with the optimal segmentation scheme. Figure 4.3a and Figure 4.3b is representing a mixture of 2 sources, which is indeed clear to see from the figures after optimal segmentation. Lastly, the Gaussian mixtures visualized in Figure 4.4a and Figure 4.4b shows an example of a mixture of 5 sources, where the optimal segmentation scheme has removed one of the true clusters. In these figures it is noticeable that the right most cluster at approximately

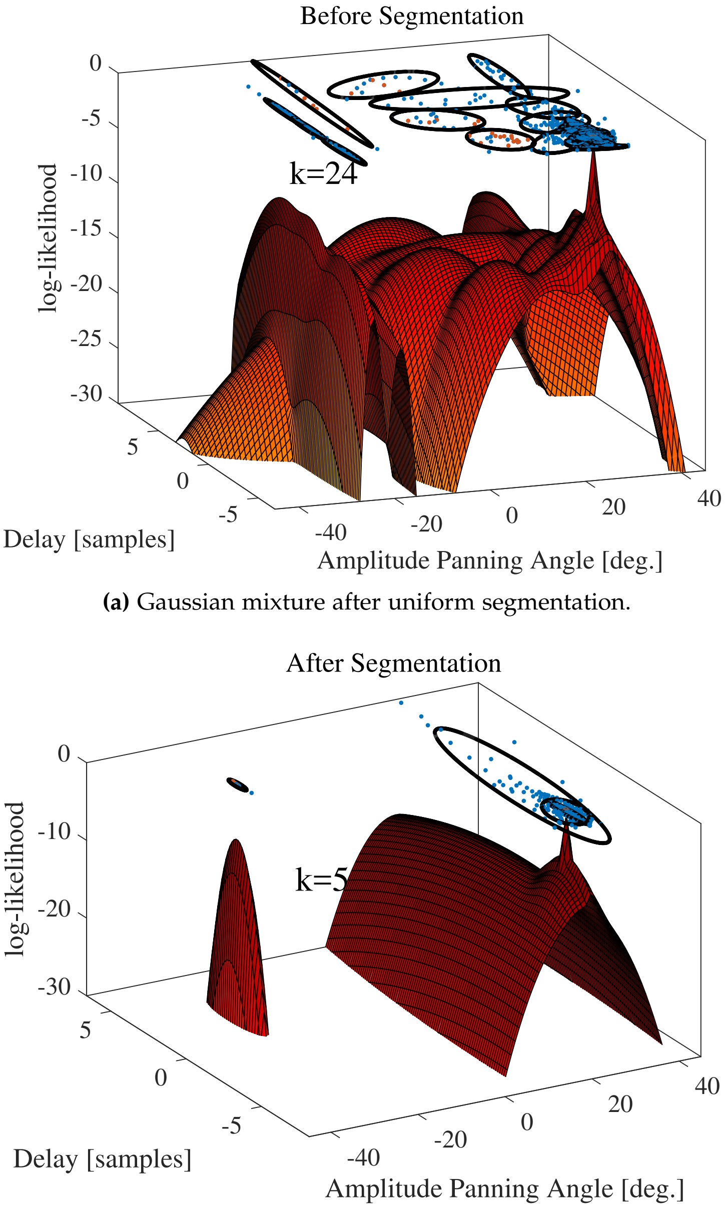


(a) Gaussian mixture after uniform segmentation.

(b) Gaussian mixture after optimal segmentation.

Figure 4.2: Comparison of Gaussian mixtures with uniform and optimal segmentation on a mixture of 4 sources.

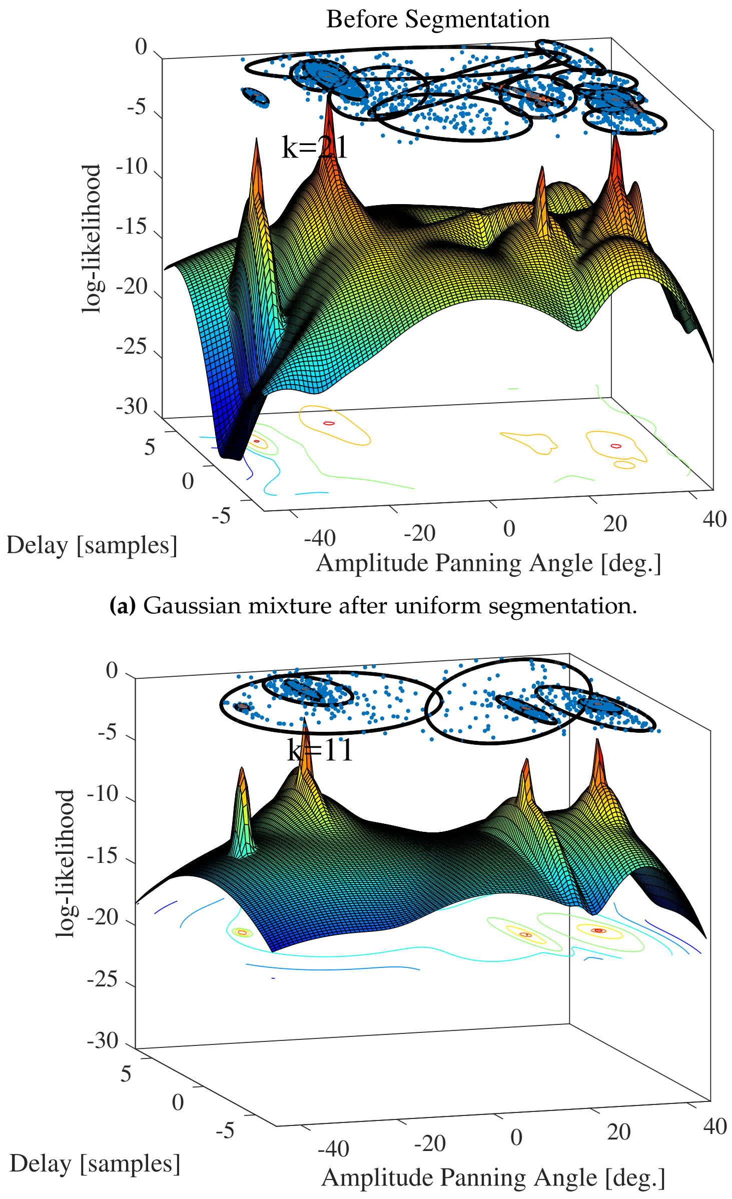
42° amplitude panning angle is removed, which could be caused by the proximity the closest cluster along with the low mixing probability of this specific cluster. The most important issue to conside in this respect, is that although the segmentation scheme will reach the global optimality based on given criteria of the model used for segmentation, the model that is applied might not be the optimal for the given purpose. This was a brief, but sufficient description of the optimal segmentation scheme that use the MAP criteria. In this chapter we have shown the practical implementation along with some visual results on the Gaussian mixture model.



(a) Gaussian mixture after uniform segmentation.

(b) Gaussian mixture after optimal segmentation.

Figure 4.3: Comparison of Gaussian mixture with uniform and optimal segmentation on a mixture of 2 sources.



(a) Gaussian mixture after uniform segmentation.

(b) Gaussian mixture after optimal segmentation.

Figure 4.4: Comparison of Gaussian mixture with uniform and optimal segmentation on a mixture of 5 sources.

Chapter 5

Experiments

5.1 Experiments on Proposed Methods

In the following the different proposed methods are tested through simulations on synthetic signals and real audio from the SQAM database [49]. To represent real music the synthetic signal are based on guitar recordings from which the amplitudes and phases have been extracted, by using an inharmonic approximate non-linear least square (ANLS) pitch estimator [9]. By testing the segmentation with synthetic signals, we can create a ground truth to when each source is active. The following proposed methods are being tested:

- Signal segmentation using the MAP log-likelihood criteria.
    - Segmentation of known synthetic guitar mixture of 2 sources.
    - Segmentation of real audio containing 2 sources from the SQAM database.
- Source Parameter Estimation for uniform segmentation with model pruning.
    - Precision measure on synthetic guitar signals of various duration.
- Source Parameter Estimation for applied optimal segmentation.
    - Precision measure on synthetic guitar signals of various duration, using the MAP log-likelihood criteria.
    - Precision measure on synthetic guitar signals of various duration, using the Calinski Harabasz criteria.
    - Segmentation criteria performance in 50 iterations on synthetic guitar signals of various duration, using the MAP log-likelihood criteria.
    - Segmentation criteria performance in 50 iterations on synthetic guitar signals of various duration, using the Calinski Harabasz criteria.

All synthetic signals were generated with 20 harmonic amplitudes and phases. The fundamental frequencies are representing notes that can be played on a guitar in the range $f_0 \in [80, 1700]$Hz, randomly applied. f_s was set to 44100 Hz.

5.1.1 Segmentation of Known Segments on Synthetic Guitar

The segmentation is tested on a synthetic signal with a source ground truth. The synthetic signal has a duration of 15 seconds and white Gaussian noise has been applied to the signal with an SNR of 50 dB. The synthetic signal is consisting of two sources with a minimum active signal duration of 300 ms and note duaration as multiples of 300 ms. The signal is segmented according to the MAP MMDL criteria of (3.1), where the minimum segment length $N_{\text{Min}} = 150$ ms and the maximum number of blocks $K_{\text{Max}} = 20$ meaning that the maximum length of a segment is 3 s. A representative example of the chosen segment length as a function of time is shown in Figure 5.1.1 with white vertical lines. In the top the two active sources are shown time domain along with a black horizontal line indicating which source is active at which time (the input segment ground truth). In the background the signal frequency content is shown to give a detailed view of the signal content. Generally the chosen segments are long if

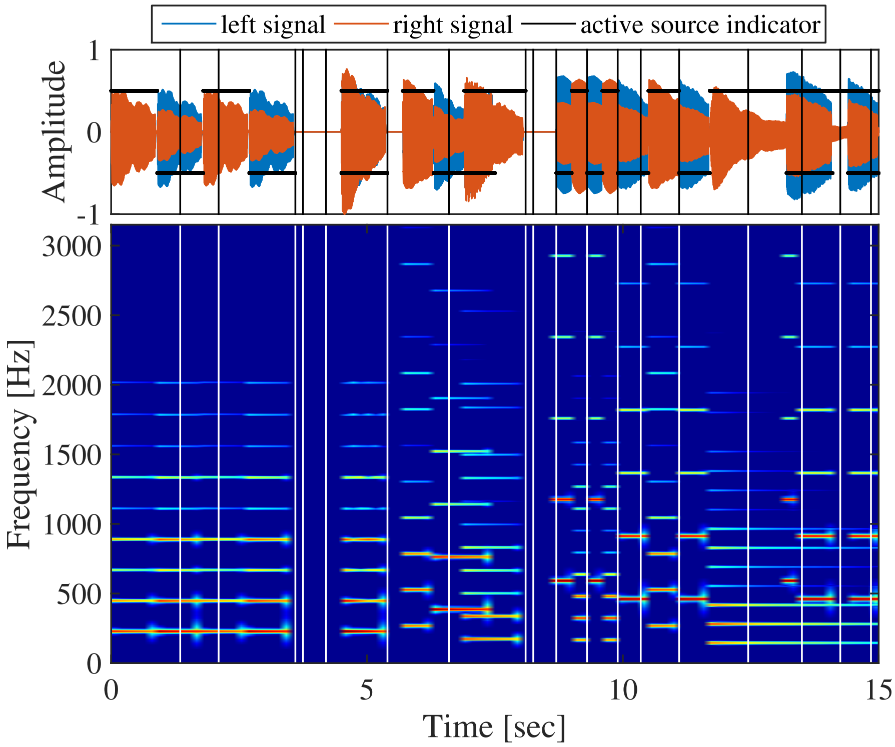


Figure 5.1: Optimal segmentation on known synthetic guitar signal, using the MAP log-likelihood criteria.

the content is not changing. Each segment contains some valuable information and seperates active input segments of the two sources. The four notes played from 0 to 4 sec (with same f_0) will consistenlty produce two underlying clusters, and we would expect the segments to be long but random. When the silent period starts a shorter segment length is chosen in all three silent periods starting at [3.6, 5.4, 8] sec. The note at 5 s. is clearly chosen, and the next three notes has an overlap that is segmented in to two parts, where only the second part has two active sources. The following notes after the silence at 8 s is chosen precisely in 300 ms segments each. Lastly, the long note from 12-15 s. is chosen in longer segments of 600 ms, but in order to separate the underlying clusters, the two overlapping notes in the end is chosen in segments in their respective note duration, even though they both overlap with the longer note. This indicates that the panning model, describes the signal in a precise way considering the source panning parameters, independently of the pitch information. The resulting distribution of parameter estimates can be seen in Figure 5.1.1. It is clear that $k = 2$, after segmentation and thresholding.

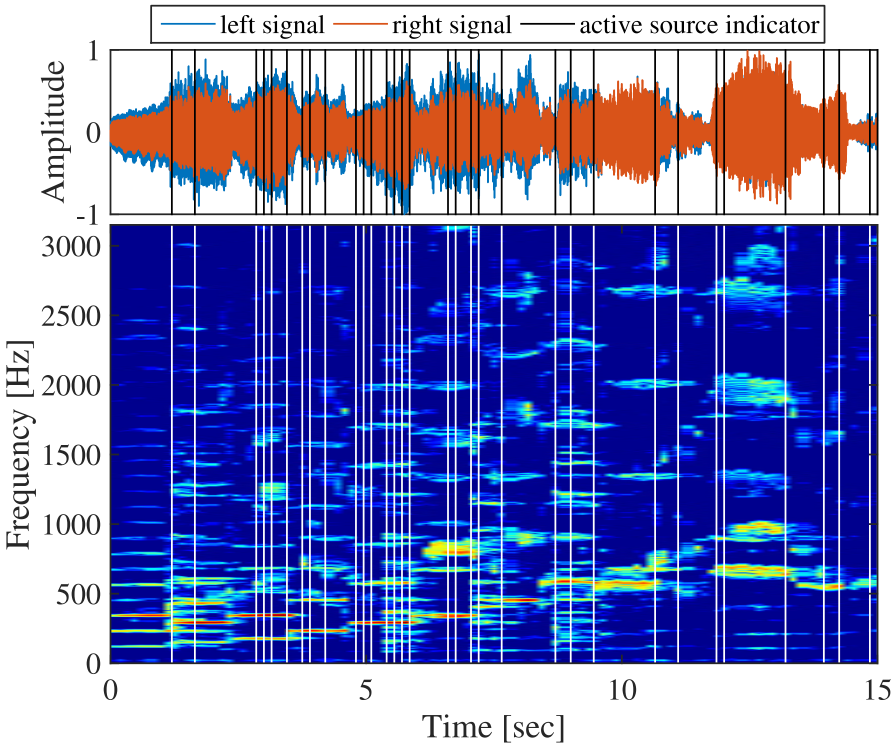


Figure 5.2: This is a visual example of signal segmentation on two sources from the SQAM database. This figure is left for the reader to analyse visually (a bit like interpretation of art). There is more segmentation figures like this in Appendix C

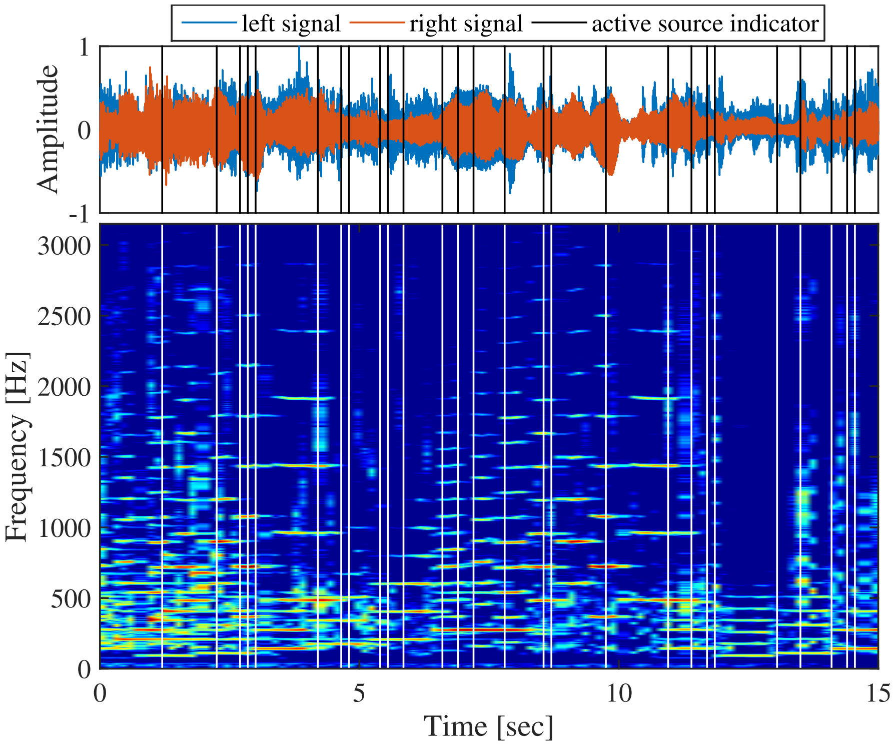


Figure 5.3: Signal segmentation on two sources from the SQAM database. There is more segmentation figures like this in Appendix C

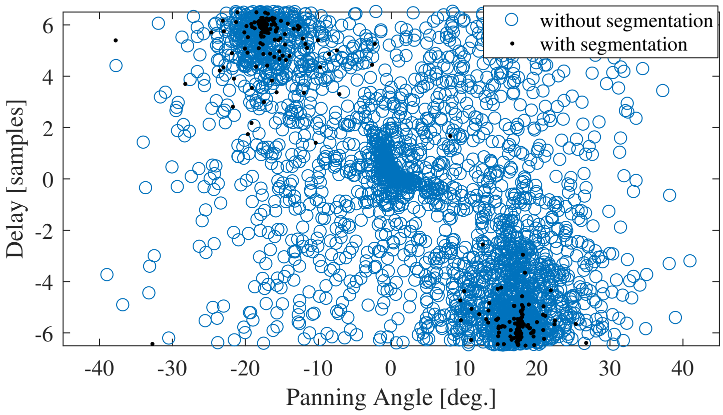


Figure 5.4: Parameter space with and without optimal segmentation and thresholding for the synthetic signal in Figure 5.1.1

5.1.2 Parameter Estimation Performance on Synthetic Guitar

The following test is testing the performance of the proposed estimator and precision of the estimates for both uniform segmentation and optimal segmentation for the MAP log-likelihood criteria and in this test we compare to the Calinski-Harabasz criteria as well. The uniform segmentation has a segment size of 300 ms and the optimal segmentation scheme can choose segment sizes in the range of 100 ms to 2000 ms in 100 ms intervals. The performance measures in this test are:

- Correctly estimated sources. A correct cluster is defined as an error below one half degree amplitude panning angle.
- RMSE of the amplitude angle estimates.
- RMSE of delay estimates.
- Estimated model order.

The test is based on 50 iteration for various durations as seen form the Figures 5.5-5.10. The synthetic signals used in the test have a segmentation ranging from 300 ms to 3000 ms in steps of 300 ms. They all contain 70% silence divided in to the same the segment durations. They are all consisting of a 2-5 source mixture. From Figure 5.5 it can be seen that the correct amount of cluster estimates is directly proportional to the duration of signal under test. Which means that the more data we gather, the more correctly we can estimate the sources in the mixture. From the Figure 5.5, it seems that in the duration range of 2-10 seconds, the performance is getting exponentially better, however this should be always be seen relative to the synthetic signal under test. From these specific 50 iterations it seems that the performance peaks at 99% correctly estimated clusters.

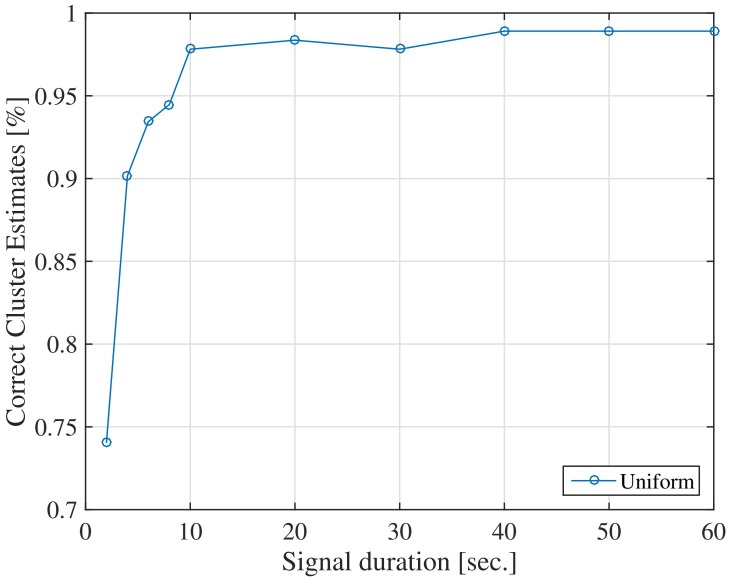


Figure 5.5: Correct cluster estimates on synthetic guitar signals as a function of signal duration. This is the proposed estimator with model order selection by model pruning and uniform segmentation.

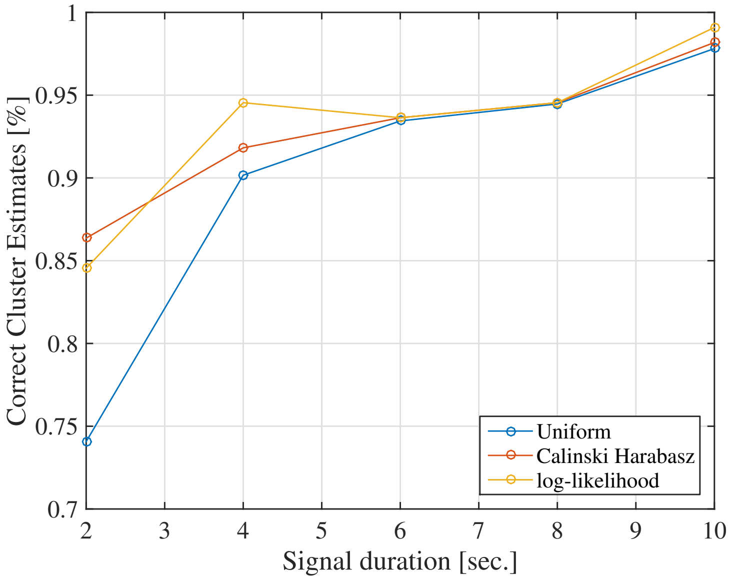


Figure 5.6: Correct cluster estimates on synthetic guitar signals as a function of signal duration. This is the proposed estimator with model order selection by model pruning with the applied optimal segmentation, using two different criterias.

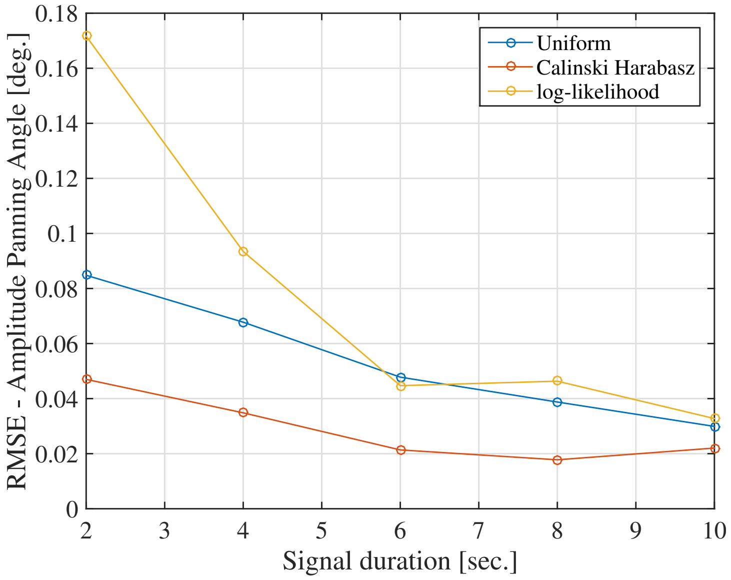


Figure 5.7: RMSE of the amplitude angle estimates on synthetic guitar signals as a function of signal duration. This is the proposed estimator with model order selection by model pruning with the applied optimal segmentation, using two different criterias.

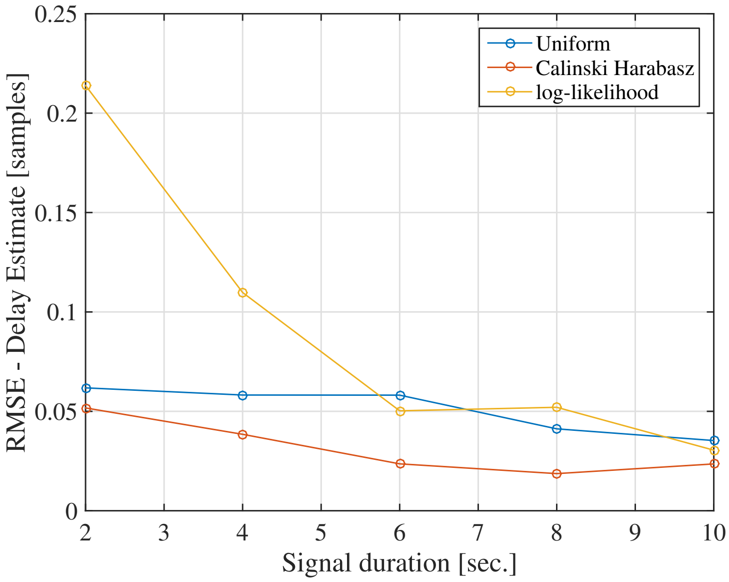


Figure 5.8: RMSE of the delay estimates on synthetic guitar signals as a function of signal duration. This is the proposed estimator with model order selection by model pruning with the applied optimal segmentation, using two different criterias.

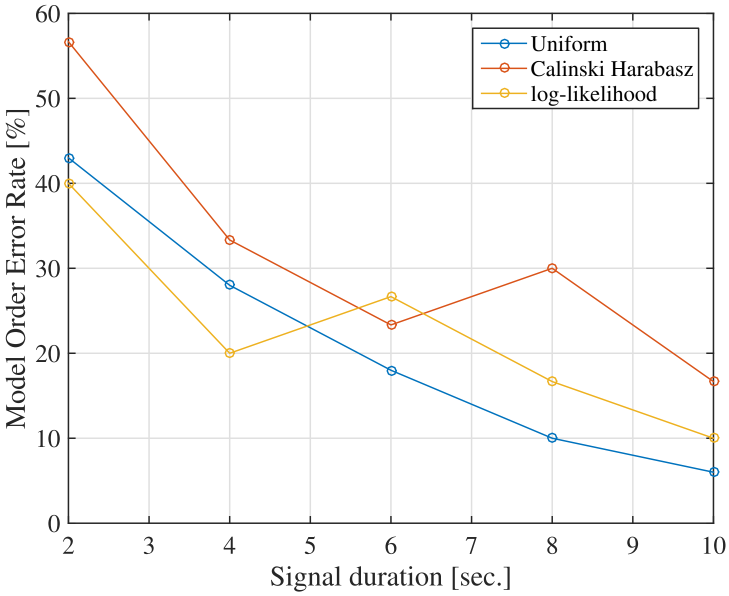


Figure 5.9: Model order estimates on synthetic guitar signals as a function of signal duration. This is the proposed estimator with model order selection by model pruning with the applied optimal segmentation, using two different criterias.

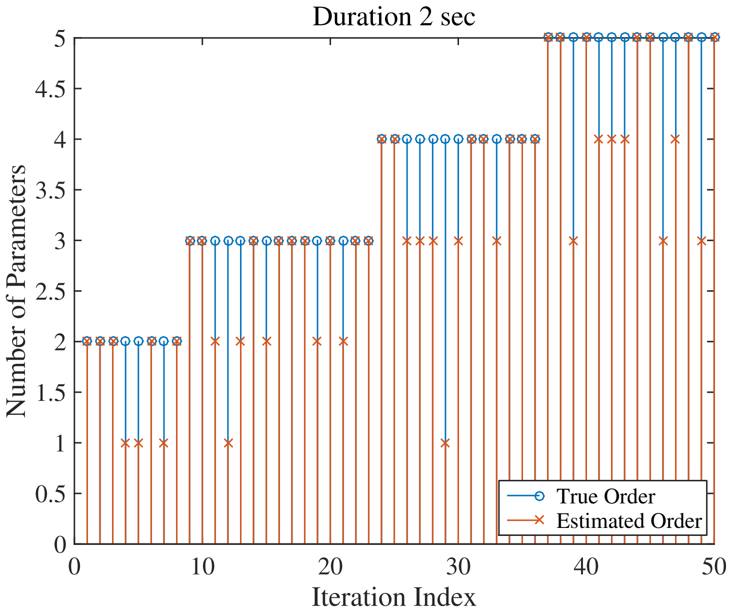


Figure 5.10: Model order estimates on synthetic guitar signals for all 50 iterations. More of these examples can be found in Appendix A.

Figure 5.9 shows the performance comparison of the uniform segmentation versus the optimal segmentation for shorter duration on the range 2-10 seconds. From this specific comparison it seems that by applying the optimal segmentation scheme, it is possible to achieve an improvement in the number of correct cluster estimates, especially for the shorter durations. The case of optimal segmentation based on MAP log-likelihood criteria, the improvement is approximately 5-10% correctly estimated clusters for the shorter durations. Similarly, by using the the Calinski-Harabasz criteria, an improvement is also possible in the same duration range.

It is implied from the test that both of the criterias under test offers improvement and it seems as if the MAP log-likelihood has the best performance for the shorter durations in terms of correct cluster estimates. However, when considering the precision results shown in Figure 5.7 and Figure 5.8 the two criterias under test, differs from this result. For this specific measure the Calinski-Harabasz criteria performs best for all durations in the short term range from 2-10 seconds. Where we have concluded that the performance in terms of correct number of cluster estimates was best for the MAP log-likelihood criteria, in this case it performs the worst when considering the precision of the correct estimates. It it even more unprecise than the uniform segmentation. It should be noted here, that because the global optimality is reached using the optimal segmentation, based on a given model, the model can still be inappropriate.

Lastly, Figure 5.9 shows that the model order estimates improves with signal duration. It is again noticeable that for the shorter duration range, the MAP log-likelihood

Estimates	MAP opt. seg.	uniform seg.
Correct Parameters (err. $\angle < 0.5°$)	94.6%	84.5%
Correct Model Order	89.1%	58.4%
Amplitude Angle (RMSE)	0.1°	0.07
Delay (RMSE)	0.33 samples	0.03

Table 5.1: Source parameter test results for real audio from the SQAM database in 100 iterations. The true number of clusters is in the range 2-5.

performs well, better than the Calinski-Harabasz and the uniform segmentation. However, non of the optimal segmentation schemes performs as well one could expect for this measure. The uniform worst case model order estimates would be for the uniform segmentation of 2 seconds signal duration. This case is shown in Figure 5.10 where it can be seen that the estimated model order is this specific case always underdetermined, meaning that the estimates are below the true value, but only with an error of one or two number of clusters, which is way better than what we saw from the MAP directly without model pruning as described in in section 3.2. More of these model order plots can be found in Appendix A.

5.1.3 Parameter Estimation Performance on real audio

The estimation of source parameters are tested on the SQAM-CD signals in 100 iterations. In this part, the estimation of panning parameters are tested for optimal segments and fixed segments. Panning Parameters implicitly has the model order, as dimensionality. For each iteration, a mixture consists of minimum 2 and maximum 5 randomly picked source components, mixed according to (1.4). Each source signal is normalized to have an absolute maximum amplitude of 1. The duration of each mixture is varying, and is defined from the audio signal in the mixture with the shortest duration, which is minimum 16 seconds for files on the SQAM-CD. The files containing pink noise has been removed from the test set. The fixed segment size is set to 600 ms. and all mixtures are passed through the thresholding of (1.4.1). All applied panning parameters to sources are stereo simulations, which means that every pair of consecutive sources, will be panned equally to each side. Other than that all panning parameters applied are random. The applied optimal segmentation scheme in this test is only the MAP log-likelihood. The results are shown in Table 5.1. We measure the estimated model order, number of correct parameter estimates, and the root mean square error for both source parameters. Clearly, there is an improvement by applying the optimal segmentation scheme, with a correct number of parameter estimates of 94.6% compared to 84.5% with uniform segmentation. The model order i.e. the correct estimate of the number of sources is also clearly improved by the time-segmentation scheme. It seems that by applying the segmentation scheme the estimator loses some of the precision in the parameter estimates, both for the amplitude and delay estimates, which was exactly the case for the test on synthetic signals with the segmentation scheme based on the MAP log-likelihoos criteria.

In this chapter we have tested the proposed sterophonic source parameter estimator. The tests have shown promising results for the unsupervised learning algorithm and by modelling the measurement space as a Gaussian mixture the estimation of source parameters are very precise when the proposed MAP estimator with uniform segmentation and it is implied that for relative short signals an improvement in cluster estimates nad model order is possible by applying the proposed time segmentation scheme.

Chapter 6

Conclusion

In this thesis, we have proposed a novel source parameter estimator for stereophonic mixtures, allowing for panning parameter estimation on multi-channel audio, even if the source pitches and harmonic amplitudes are unknown. To the authors knowledge, it has not been established before that the stereophonic panning parameters, have been estimated explicitly without a preceeding pitch estimate. The presented method does not require prior knowledge of the number of sources present in the mixture. The proposed estimator is formulated in an unsupervised learning framework, using Bayesian statistics for the modelling of the parameter space. The Bayesian approach offers some attractive advantages over i.e. the classical approach for blind source separation, as we outlined in the introduction. One of the great advantages is that the Bayesian approach offers the complete and optimal solution in terms of the posterior distribution on which all probabilistic statements about the problem are based. The maximum a posterior model formulation opens the possibility of estimating the number of sources and applying optimal time segmentation of the stereophonic signal.
Initially we defined the stereophonic signal model, and the relations to virtual sound source positioning. This lead to the definition of the stereophonic panning parameter estimates computed from the Fourier transform. In the proposed method, the distribution of stereophonic panning parameters is modelled with a Gaussian mixture model. The model parameters are estimated by using the maximum a posteriori estimation based on the expectation-maximization algorithm. In order to avoid one cluster being modelled by two or more Gaussians, we have utilized the Dirichlet distributions as the prior of the GMM mixture probabilities. This was done by adopting a Mixture-MDL algorithm. The MMDL algorithm was extended by applying a model pruning algorithm, based on the determinants of component covariances and angles between subspaces in the mixture distribution. This extension has been shown to perform better than the original MMDL in terms of model order selection of source components in the Gaussian mixture model. To obtain a better time segmentation of the stereophonic mixtures, we have proposed a segmentation scheme that guarantees the global

optimality, based on the maximum a posteriori formulation.

The developed estimator was evaluated through simulations on synthetic guitar signals as well as on a real audio signal from the SQAM database. These simulations showed that the developed estimator performs well in terms of source parameter estimation and in estimating number or sources in the stereophonic mixture for the uniform segmentation. The estimator improves its performance proportional to the signal duration. The simulations showed that the optimal time segmentation can be successfully applied to stereophonic mixtures to improve the performance of correct source estimates, in particular, on the short signal durations.

Although applicable for stereophonic source parameter estimation on synthetic and real audio signals, the proposed estimator is still subject to unsolved problems and open to further research. For example, the proposed method for time segmentation is based on the MMDL, thus the asymptotic MAP, for modelling of Gaussian distributions of varying sample sizes. Although, an improvement has been shown, it is possible that the effect of this is that small sample time segments is favored as Gaussians, while they are truly noise from the estimates. Through experiments we have compared this criteria to the Calinski-Harabasz, which shows higher precision in the estimates, which is promising for further research. Another limitation of the proposed estimator which still remains unsolved, is the implementation of the model pruning algorithm. The model pruning shows very good results with fewer errors than the clean MMDL, and the error sizes after model pruning is greatly improved. It would be desirable to extend the proposed estimator to handle the model pruning as part of the maximum a posterior criteria, instead of a post processing algorithm. This is also a subject for further research.

Through this thesis, new knowledge from unsupervised learning of stereophonic mixtures has been obtained and this may be important for future research work. Furthermore, the problem of estimating the stereophonic panning parameters, is in general a new research and the proposed solution has no explicit precursor, but shows promising results.

Bibliography

[1] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Græsbøll Christensen. "Estimation of Multiple Pitches in Stereophonic Mixtures using a Codebook-based Approach". In: *I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings* (Mar. 2017). ISSN: 1520-6149.

[2] R. J. Weiss M. I. Mandell and D. P. Ellis. "Model-Based Expectation-Maximization Source Separation and Localization". In: *IEEE Trans. Audio, Speech and Language Process.* 18.2 (2010), pp. 384–394.

[3] J. Benesty J. R. Jensen and M. G. Christensen. "Joint filtering scheme for nonstationary noise reduction". In: *Proc. European Signal Processing Conf.* (2012), pp. 2323–2327.

[4] Anssi Klapuri and Manuel Davy, eds. New York: Springer, 2006. ISBN: 0-387-30667-6.

[5] G. Tzanetakis and P. Cook. "Musical Genre Classification of Audio Signals". In: *IEEE Transactions on Speech and Audio Processing* 10.5 (2002).

[6] Jesper Rindom Jensen, Mads Græsbøll Christensen, and Søren Holdt Jensen. "Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation". In: *IEEE Trans. Audio, Speech & Language Processing* 21.5 (2013), pp. 923–933.

[7] Ville Pulkki and Matti Karjalainen. "Localization of Amplitude-Panned Virtual Sources". In: *J. Audio Eng. Soc* 49.9 (2001), pp. 739–752.

[8] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Græsbøll Christensen. "Pitch estimation of stereophonic mixtures of delay and amplitude panned signals". In: *23rd European Signal Processing Conference, EUSIPCO 2015, Nice, France, August 31 - September 4, 2015*. 2015, pp. 36–40.

[9] Mads Græsbøll Christensen and Andreas Jakobsson. *Multi-Pitch Estimation*. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2009.

[10] T. Nilsson et al. "Multi-pitch estimation of inharmonic signals". In: *Proc. European Signal Processing Conf.* (2013), pp. 1–5.

[11] I. Barbancho et al. "Inharmonicity-Based Method for the Automatic Generation of Guitar Tablature". In: *IEEE Trans. Audio, Speech and Language Process.* 20.6 (2012), pp. 1857–1868.

[12] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Græsbøll Christensen. "Multi-pitch estimation of audio recordings using a codebook-based approach". In: *24th European Signal Processing Conference, EUSIPCO 2016, Budapest, Hungary, August 29 - September 2, 2016*. 2016, pp. 983–987.

[13] T. Kronvall et al. "Sparse Multi-Pitch and Panning Estimation of Stereophonic Signals". In: *IEEE Trans. Audio, Speech and Language Process.* (Dec. 2016).

[14] Jens Blauert. *The Psycophysics of Human Sound Localization*. MIT Press, 2009.

[15] Y. Huang J. Benesty J. Chen. *Microphone Array Signal Processing*. Springer, 2008.

[16] Vincent Mohammad Tavakoli et al. "A partitioned approach to signal separation with microphone ad hoc arrays". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. 2016, pp. 3221–3225.

[17] S. Rickard and O. Yilmaz. "Blind separation of speech mixtures via time-frequency masking". In: *IEEE Transactions on Signal Processing* 52.7 (2004), pp. 1830–1847.

[18] S. Rickard and O. Yilmaz. "On the Appriximate W-Disjoint Orthogonality of Speech". In: *IEEE Acoustics, Speech, and Signal Processing* (2002).

[19] S. Rickard A. Jourjine and O. Yilmaz. "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing.* 5 (2000), pp. 2985–2988.

[20] Carlos Avendano and Jean-Marc Jot. "Frequency Domain Techniques for Stereo to Multichannel Upmix". In: *AES 22nd international Conference on Virtual, Synthetic and Entertainment Audio* (2002).

[21] Dirk Ormoneit and Volker Tresp. "Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates". In: *IEEE Trans. Neural Networks* 9.4 (1998), pp. 639–650.

[22] Zoran Zivkovic and Ferdinand van der Heijden. "Recursive Unsupervised Learning of Finite Mixture Models." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 26.5 (2004), pp. 651–656.

[23] Duane H. Cooper. "Problems with Shadowless Stereo Theory: Asymptotic Spectral Status". In: *J. Audio Eng. Soc* 35.9 (1987), pp. 629–642.

[24] Robert A. Katz. *Mastering Audio: The Art and the Science*. Butterworth-Heinemann Newton, 2009.

[25] Alan Dower Blumlein. *U.K. Patent 394* (1931); reprinted in *Stereophonic Techniques* (Audio Engineering Society, New York, 1986).

[26] Benjamin B. Bauer. “Phasor Analysis of some Stereophonic Phenomena”. In: *The Journal of The Acoustical Society of America* 33.11 (1961), pp. 1536–1540.

[27] Benjamin Bernfeld. “Attempts for Better Understanding of the Directional Stereophonic Listening Mechanism”. In: *Audio Engineering Society Convention 44*. 1973. URL: `http://www.aes.org/e-lib/browse.cfm?elib=1743`.

[28] Ville Pulkki. “Virtual Sound Source Positioning Using Vector Base Amplitude Panning”. In: *J. Audio Eng. Soc* 45.6 (1997), pp. 456–466.

[29] Stuart P. Lloyd. “Least squares quantization in pcm”. In: *IEEE Transactions on Information Theory* 28 (1982), pp. 129–137.

[30] Mário A. T. Figueiredo and Anil K. Jain. “Unsupervised Learning of Finite Mixture Models”. In: *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 24 (2000), pp. 381–396.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1 (1977), pp. 1–38.

[32] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 1996. ISBN: 9780471123583. URL: `https://books.google.dk/books?id=iRSWQgAACAAJ`.

[33] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.

[34] T. Caliński and J. Harabasz. “A Dendrite Method for Cluster Analysis”. In: *Communications in Statistics* 3.1 (1974), pp. 1–27. DOI: `10.1080/03610927408827101`.

[35] David Arthur and Sergei Vassilvitskii. “K-means++: The Advantages of Careful Seeding”. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA ’07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[36] D. de Ridder F. van der Heijden R.P.W. Duin and D.M.J. Tax. *Classification, Parameter Estimation and State Estimation*. 1. ed. John Wiley and Sons, Ltd., 2004, pp. 17–32.

[37] Robert E. Kass and Adrian E. Raftery. “Bayes Factors”. In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795.

[38] Mads Græsbøll and Andreas Jakobsen. *Multi-Pitch Estimation*. 1. ed. Morgan and Claypool, 2009.

[39] Petar M. Djuric. “Asymptotic MAP criteria for model selection”. In: *IEEE Trans. Signal Processing* 46.10 (1998), pp. 2726–2735.

[40] Chris Fraley and Adrian E. Raftery. "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis." In: *The Computer Journal* 41.8 (1998), pp. 578–588.

[41] Matthew Brand. "Structure Learning in Conditional Probability Models via an Entropic Prior and Parameter Extinction". In: *Neural Computation* 11.5 (1999), pp. 1155–1182.

[42] Andrew Gelman et al. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003. ISBN: 158488388X.

[43] Gilles Celeux et al. "A Component-Wise EM Algorithm for Mixtures". In: *Journal of Computational and Graphical Statistics* 10.4 (2001), pp. 697–712.

[44] Shoko Araki et al. "Stereo Source Separation and Source Counting with MAP Estimation with Dirichlet Prior Considering Spatial Aliasing Problem". In: *Independent Component Analysis and Signal Separation, 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings*. 2009, pp. 742–750.

[45] *Flexible time segmentations for time-varying wavelet packets*. 1994, pp. 9–12.

[46] M. Vetterli P. Prandoni M. M. Goodwin. "Optimal time segmentation for signal modeling and compression". In: *IEEE Acoustics, Speech, and Signal Processing* (1997), pp. 2029–2032.

[47] Paolo Prandoni and Martin Vetterli. "R/D Optimal Linear Prediction". In: *IEEE Trans. Speech and Audio Proc* 8 (2000), pp. 646–655.

[48] Sidsel Marie Nørholm, Jesper Rindom Jensen, and Mads Græsbøll Christensen. "Instantaneous Fundamental Frequency Estimation With Optimal Segmentation for Nonstationary Voiced Speech". In: *IEEE/ACM Trans. Audio, Speech & Language Processing* 24.12 (2016), pp. 2354–2367.

[49] European Broadcasting Union. *Sound quality assessment material recordings for subjective tests: Users handbook for the EBU SQAM CD*. Tech. Rep. EBU - TECH 3253. 2008.

Appendix

Appendix A

Examples of Model Order Estimates

In the following some examples of model order estimates are shown. This is shown for the proposed estimator with MAP model pruning and uniform segmentation. What is noticeable from all figures showing various durations, is that when the model order is estimated wrongly, it is mainly an error of one or two orders below the true value. Which is promising result compared tot he original order estimates without model pruning as seen in section 3.2.

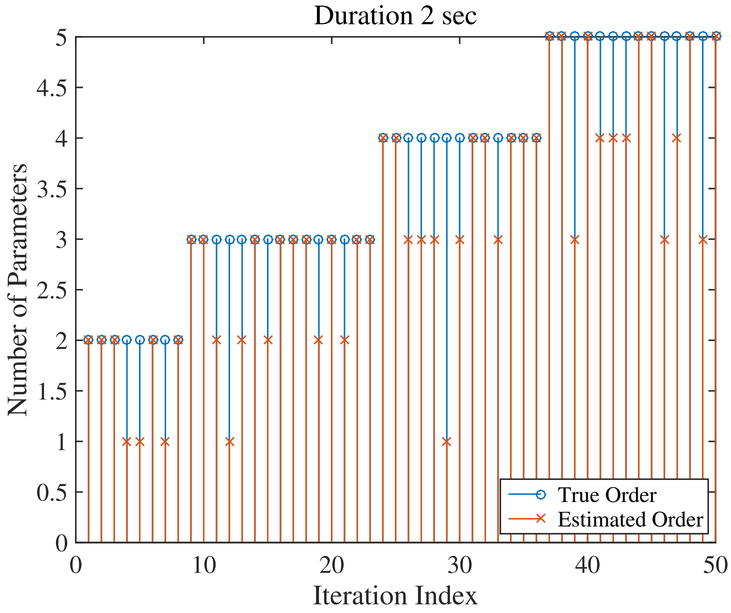


Figure A.1: Model order estimates on synthetic guitar signals for all 50 iterations.

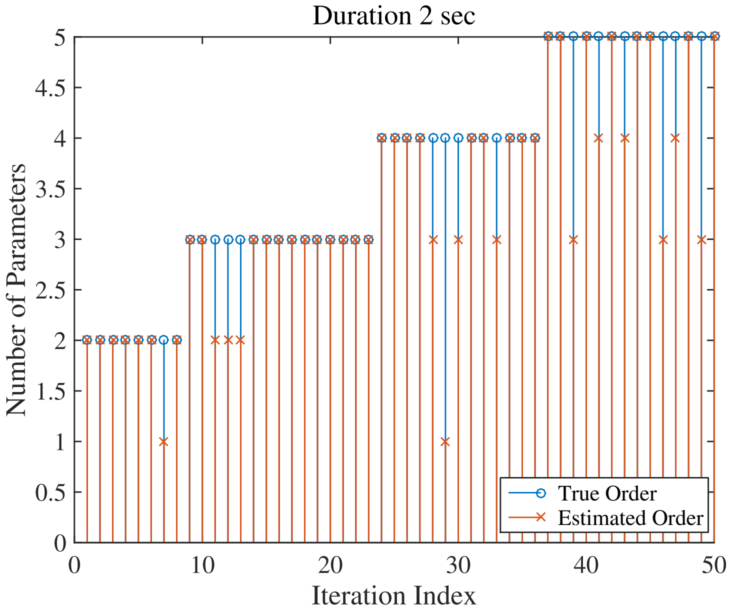


Figure A.2: Model order estimates on synthetic guitar signals for all 50 iterations.

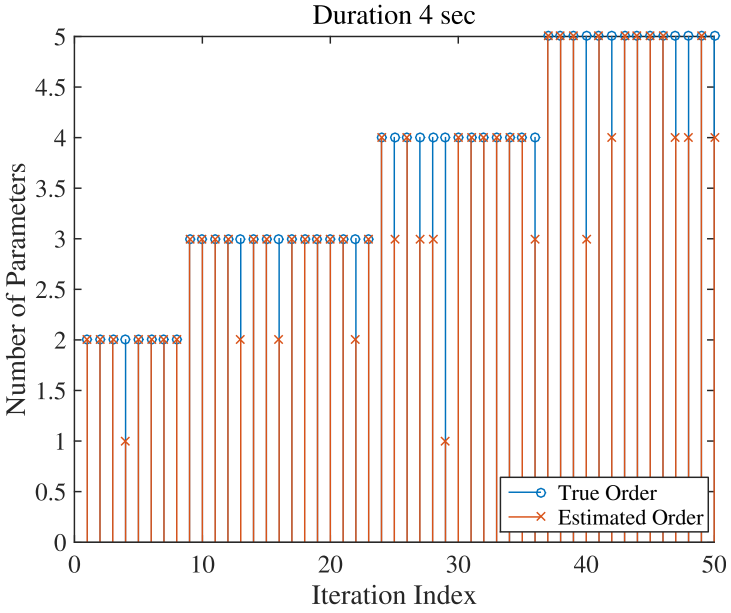


Figure A.3: Model order estimates on synthetic guitar signals for all 50 iterations.

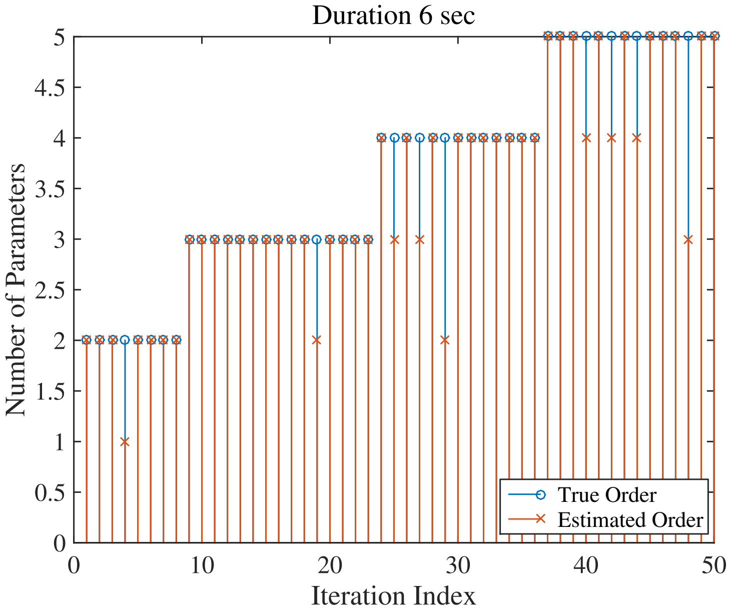


Figure A.4: Model order estimates on synthetic guitar signals for all 50 iterations.

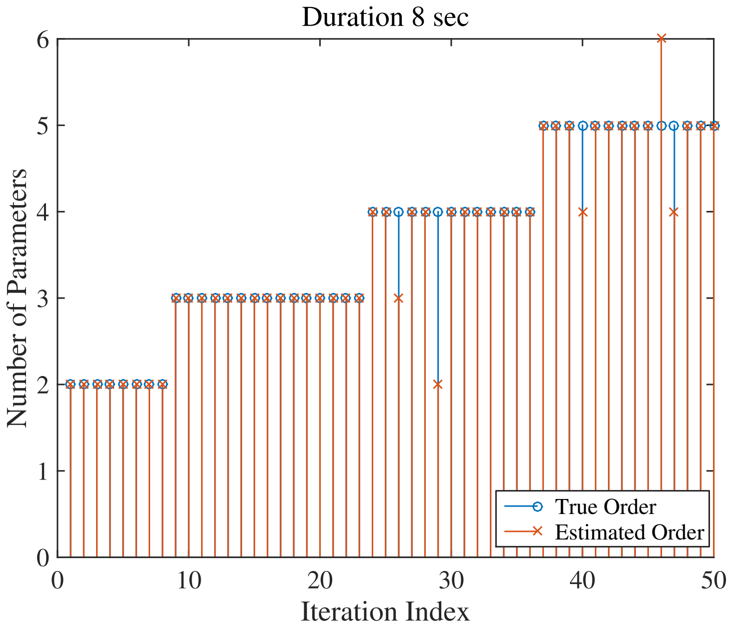


Figure A.5: Model order estimates on synthetic guitar signals for all 50 iterations.

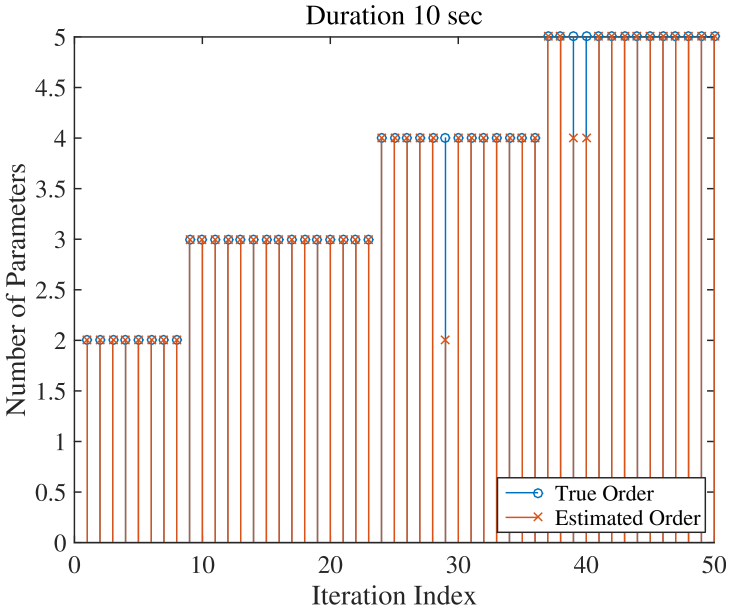


Figure A.6: Model order estimates on synthetic guitar signals for all 50 iterations.

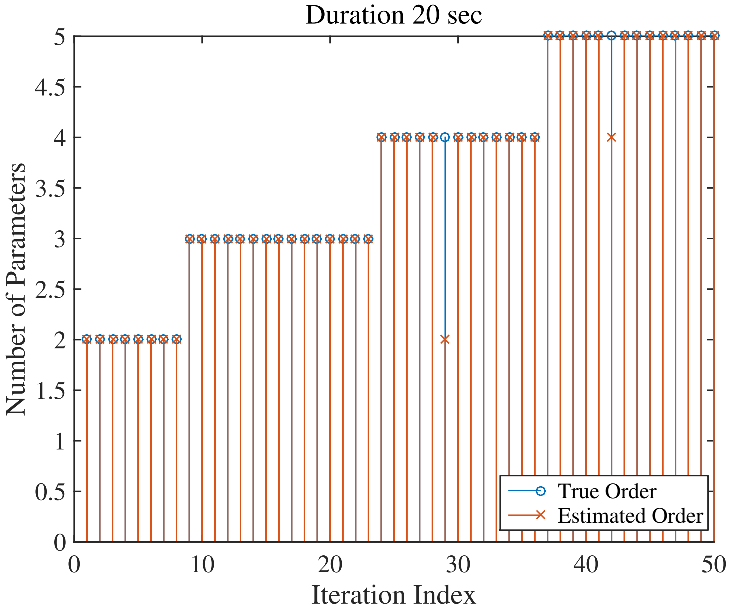


Figure A.7: Model order estimates on synthetic guitar signals for all 50 iterations.

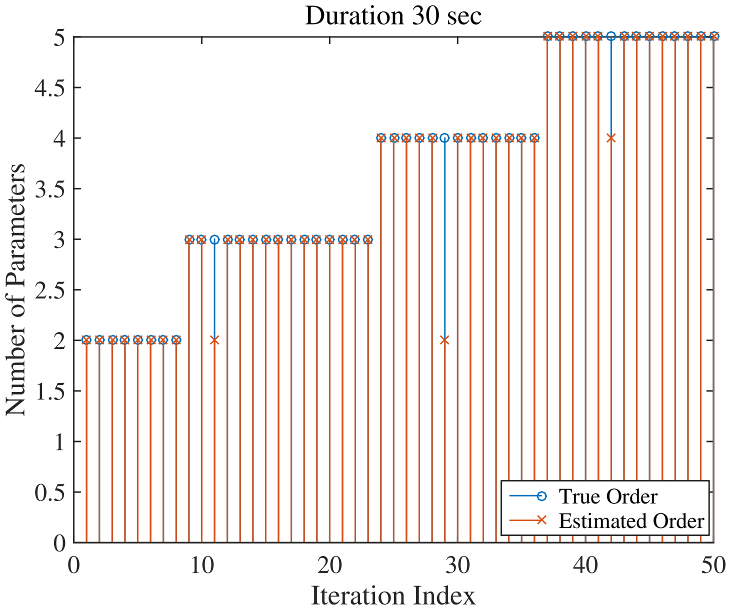


Figure A.8: Model order estimates on synthetic guitar signals for all 50 iterations.

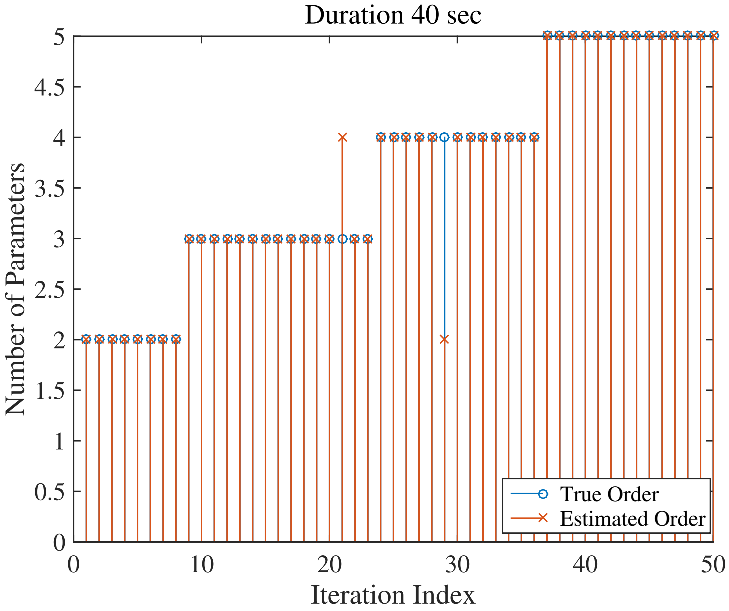


Figure A.9: Model order estimates on synthetic guitar signals for all 50 iterations.

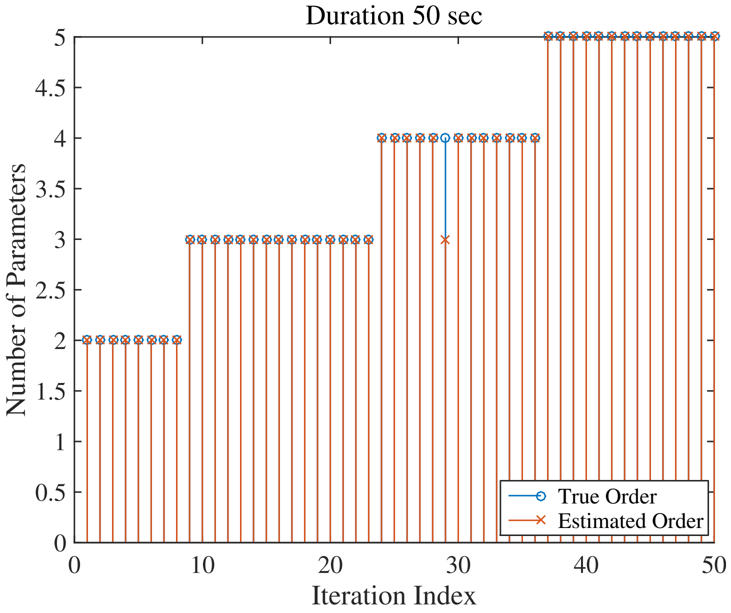


Figure A.10: Model order estimates on synthetic guitar signals for all 50 iterations.

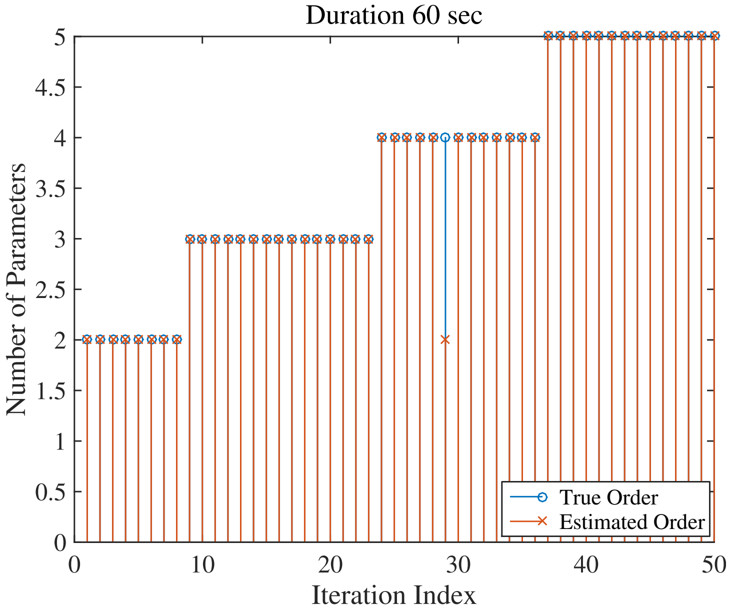


Figure A.11: Model order estimates on synthetic guitar signals for all 50 iterations.

Appendix B

Estimation the Amplitude Panning Angle

It is possible to estimate the panning angle by doing a simple search within the frequency domain. This was an initial estimator of the amplitude panning parameter. What is interesting about the following algorithm is that it requires very little amount of data for a simple estimate and therefore it is interesting to use this simple algorithm to make estimates in time-pan domain as we often see the time-frequency domain refered to as the spectrogram; we refer to time-plot plot as the panogram. The algorithm is introduced in Section 1.4.2. In this section we show the panogram representations of stereo mixture of instruments. All panning parameters have been applied by using the Digital Audio Workstation (DAW) called Logic Pro. Therefore all panning parameters are applied without the use of our model, but only estimated using out model. As it is seen, the mean energy of the amplitude panning parameter estimates are at the correct location. The panning knob interface of a DAW like Logic Pro, is based on a loudspeaker aperture of 60°.

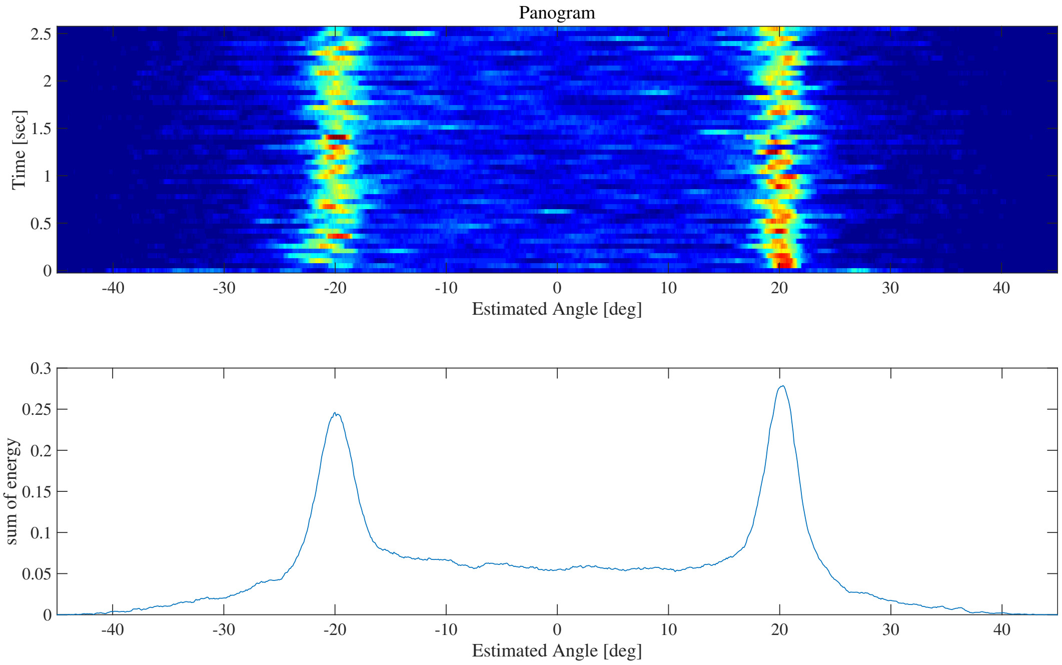


Figure B.1: Panogram of the trumpet mixture refered to in the experiments in [1].

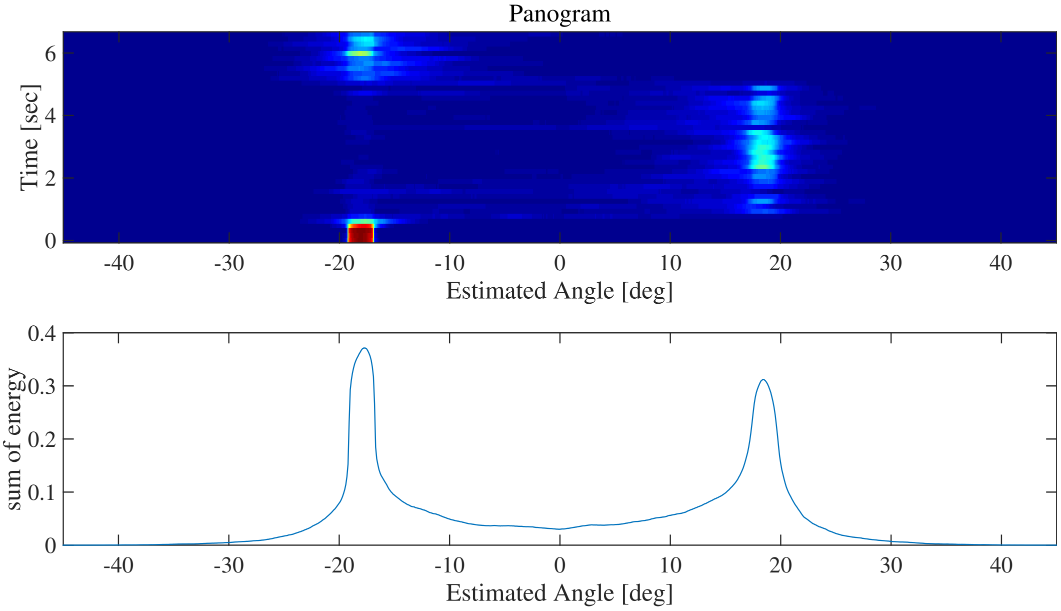


Figure B.2: Panogram of a mix of celloes. In Logic Pro these are panned on the given knob to 28. This fits with a ratio of 60/90 because the estimate is based on a loudspeaker aperture of 90°.

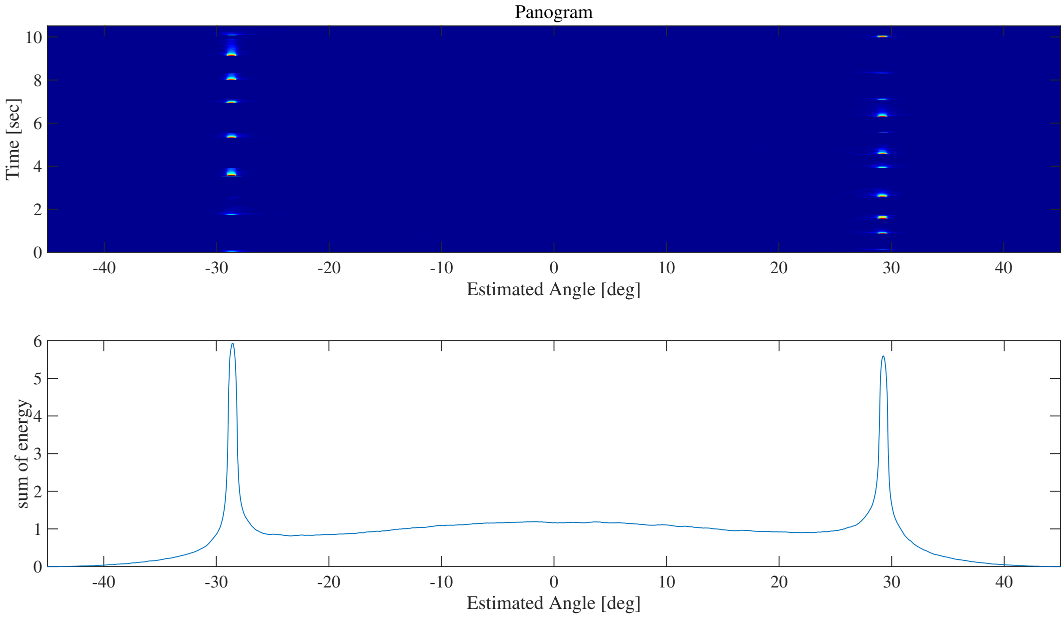


Figure B.3: Panogram of a mix of guitar plucks. In Logic Pro these are panned on the given knob to 43. This fits with a ratio of 60/90 because the estimate is based on a loudspeaker aperture of 90°.

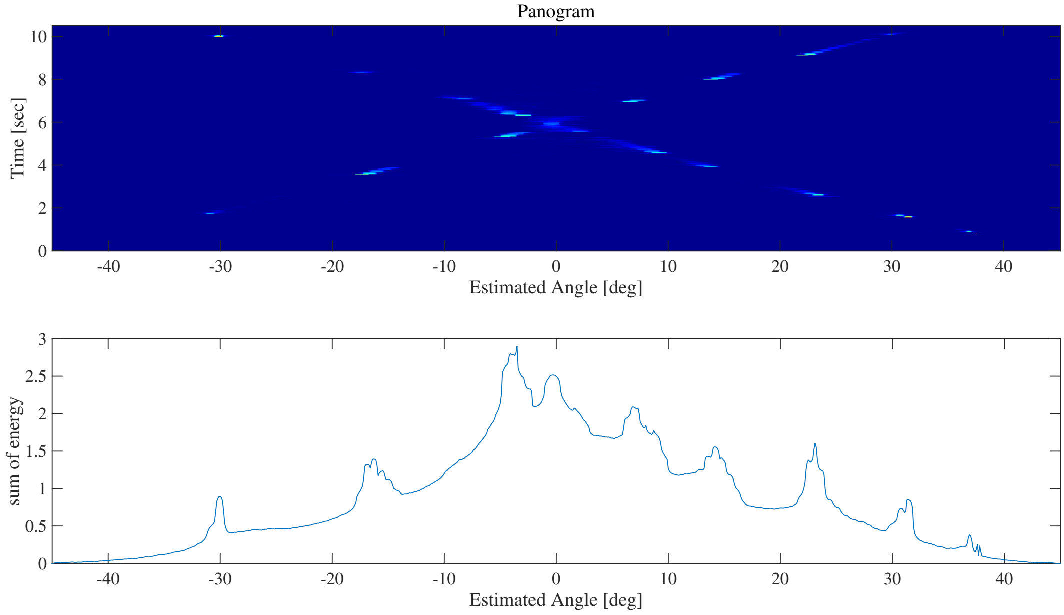


Figure B.4: Panogram of the mix of same guitar plucks from Figure B.3 only now they are changing position over time.

Appendix C

Segmentation Visualized in Time-Frequency Domain

This section shows the segmented signals in time-frequency domain. It is possible that these figures can be visually inspected, to give an idea or an understanding of the optimal segmentation scheme, based on the MAP MMDL criteria. Every figure in the following is either representing a mixture of 2 sources or a mixture of three sources. It would make the most sense to understand the segmentation from two sources. When looking at the spectrograms of three sources, it can make sense sometimes, while other times it does not. There is two main reasons for this. Firstly, the separation of sources is based on the underlying distribution of the clusters using the given criteria, and furthermore, while the segmentation scheme reaches the global optimum, the model might not be optimal. Secondly, we can visualize the left and right channel as we do in these figures, and we can visualize the left or the right spectrum as we do in these figures. However, we have not visualized the seperate sources that the mixtures consists of. It is important to have in mind, while looking at these representations, that the segmentation is based on the clustering of the sources that is not clear in the figure.

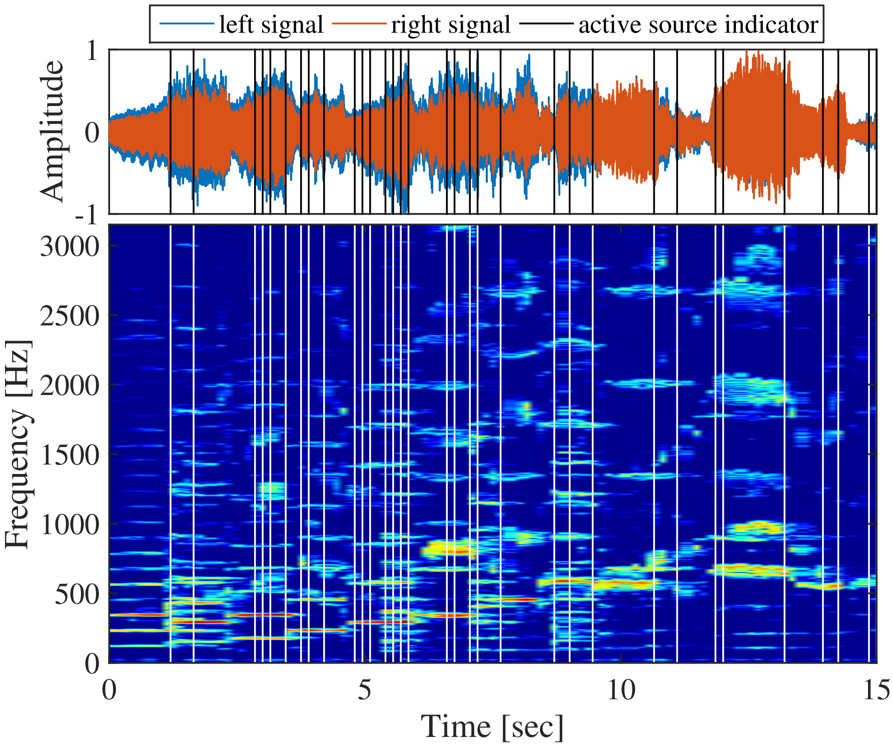


Figure C.1: This is a visual example of signal segmentation on a mixture two sources from the SQAM database. This figure is left for the reader to analyse visually (a bit like interpretation of art).

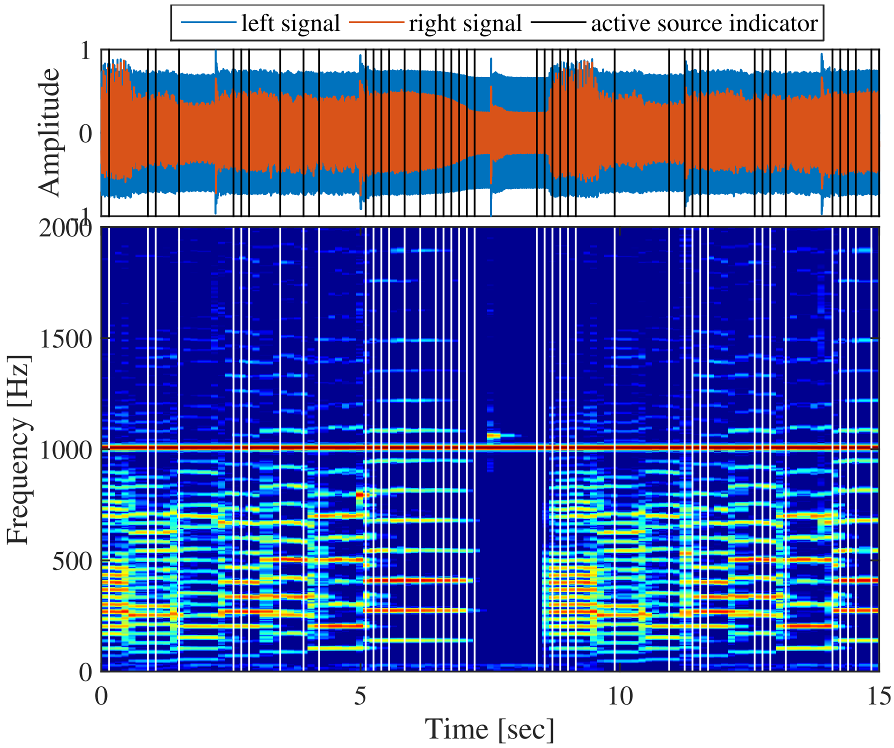


Figure C.2: Signal segmentation a mixture on two sources from the SQAM database.

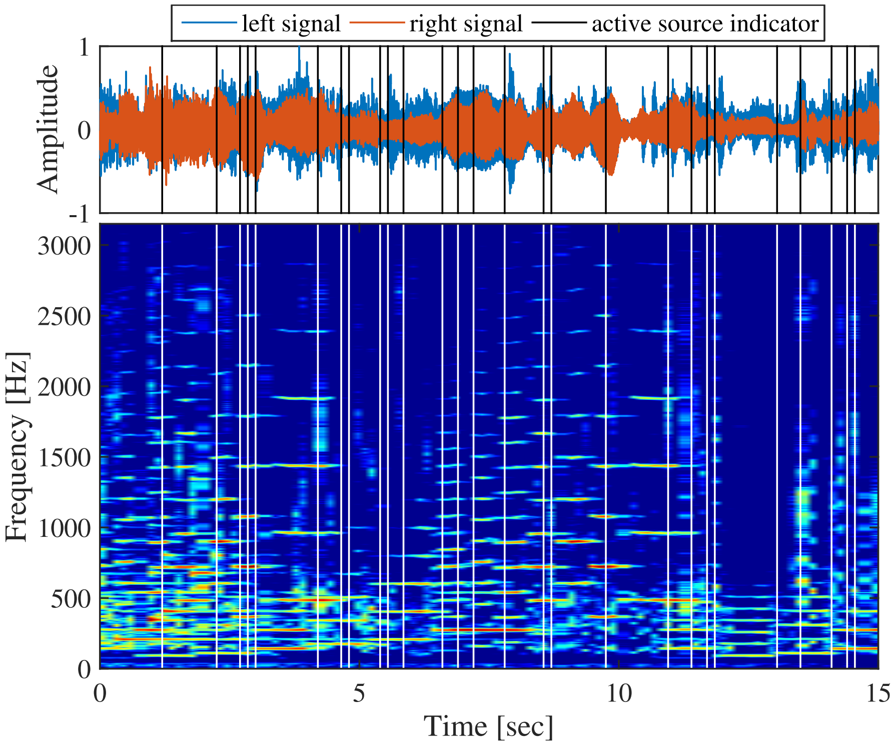


Figure C.3: Signal segmentation on a mixture two sources from the SQAM database.

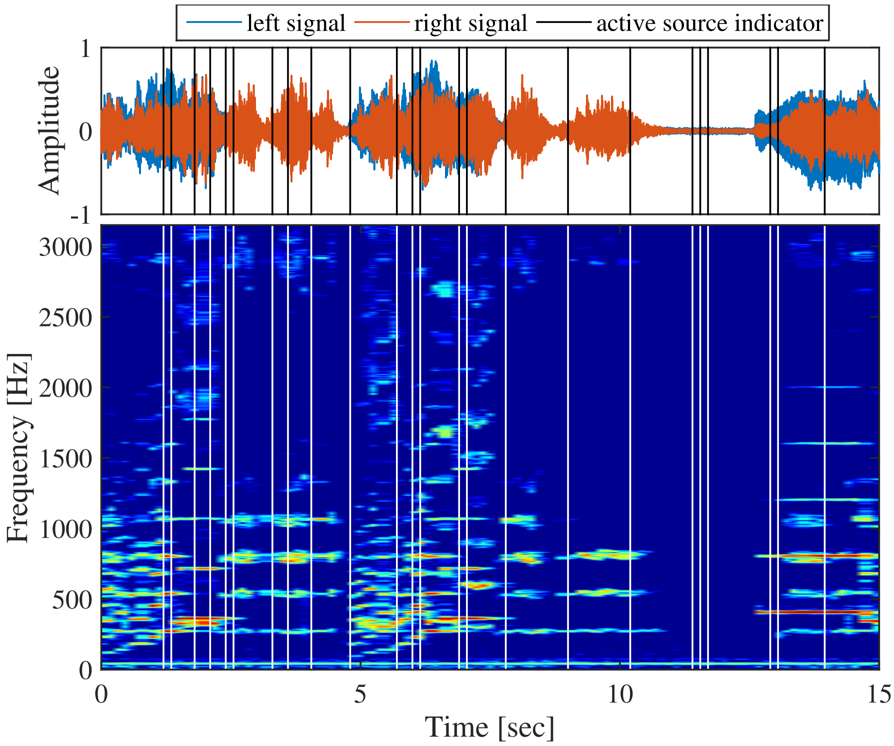


Figure C.4: Signal segmentation on a mixture three sources from the SQAM database.

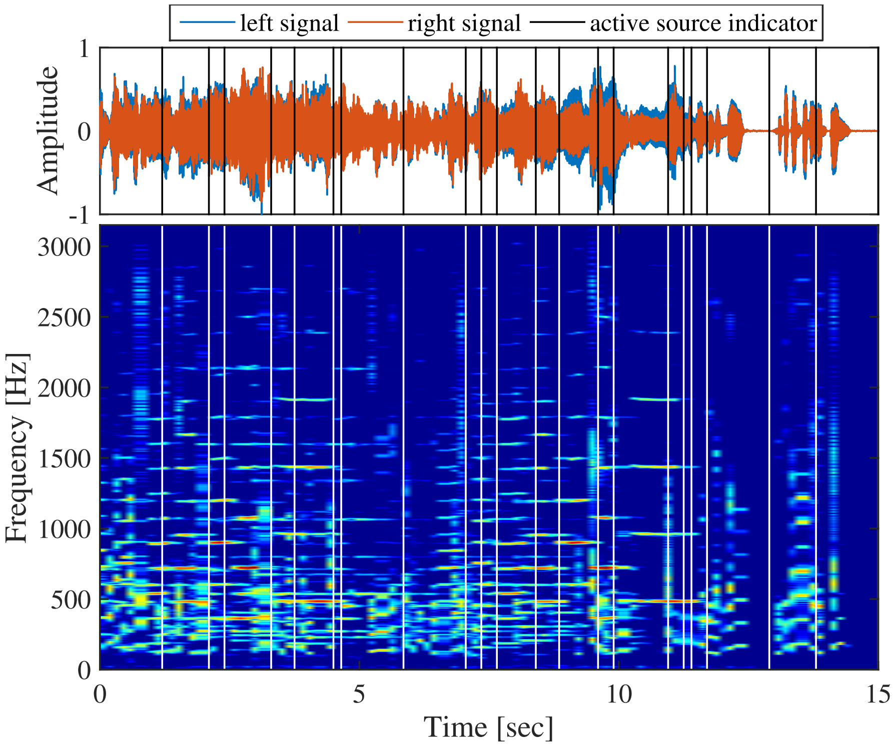


Figure C.5: Signal segmentation on a mixture three sources from the SQAM database.

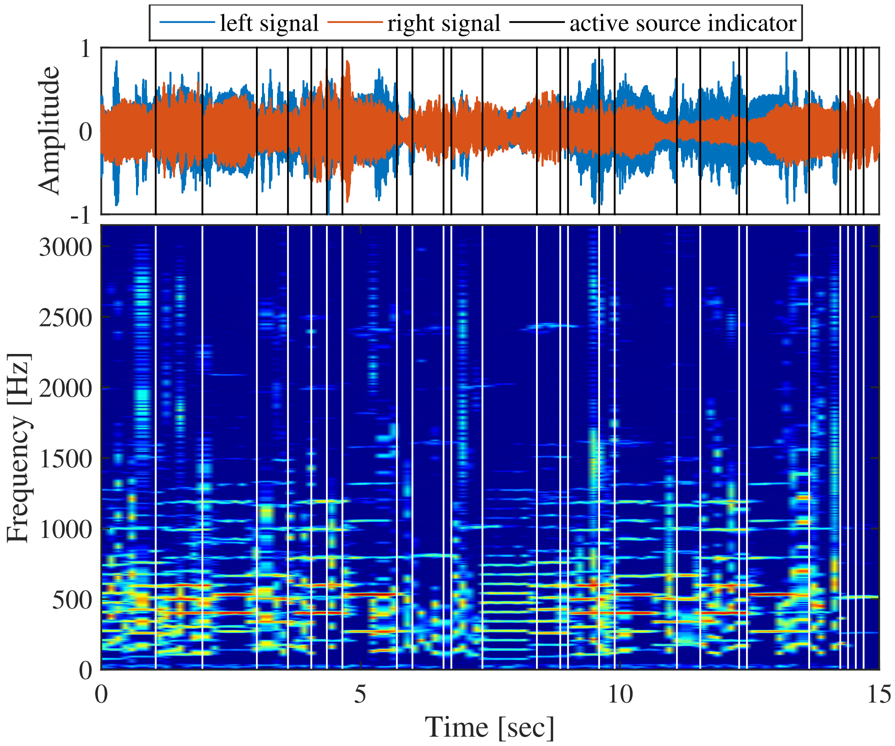


Figure C.6: Signal segmentation on a mixture three sources from the SQAM database.

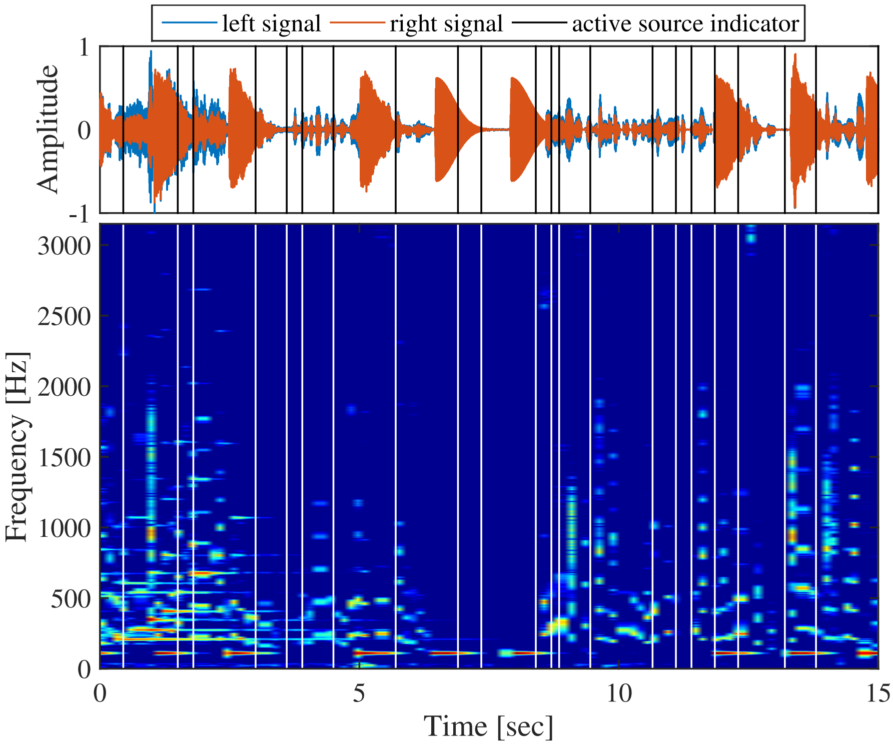


Figure C.7: Signal segmentation on a mixture three sources from the SQAM database.

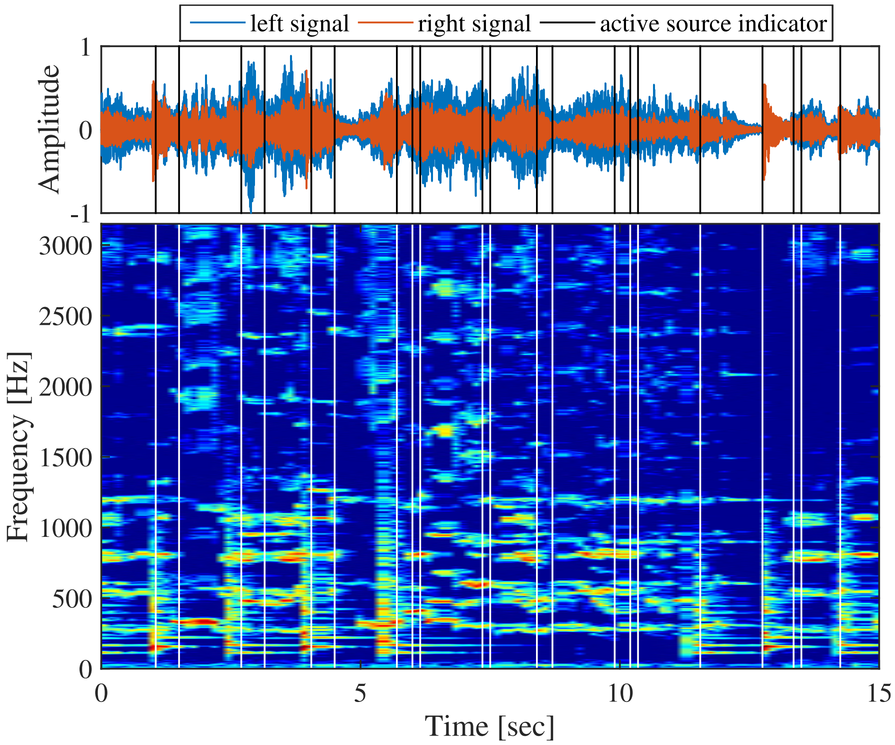


Figure C.8: Signal segmentation on a mixture three sources from the SQAM database.

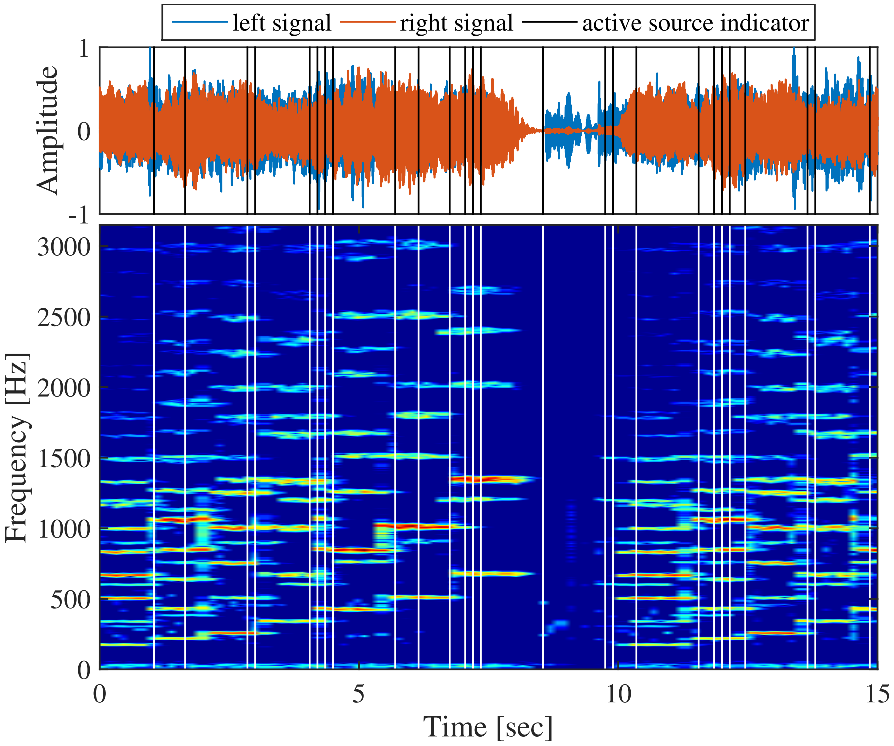


Figure C.9: Signal segmentation on a mixture three sources from the SQAM database.

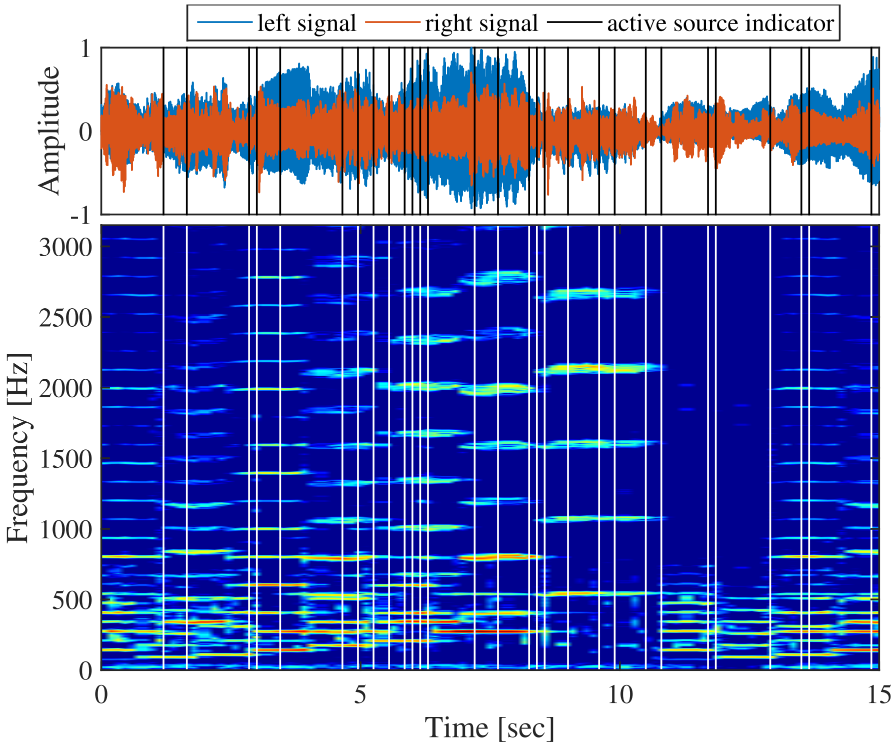


Figure C.10: Signal segmentation on a mixture three sources from the SQAM database.

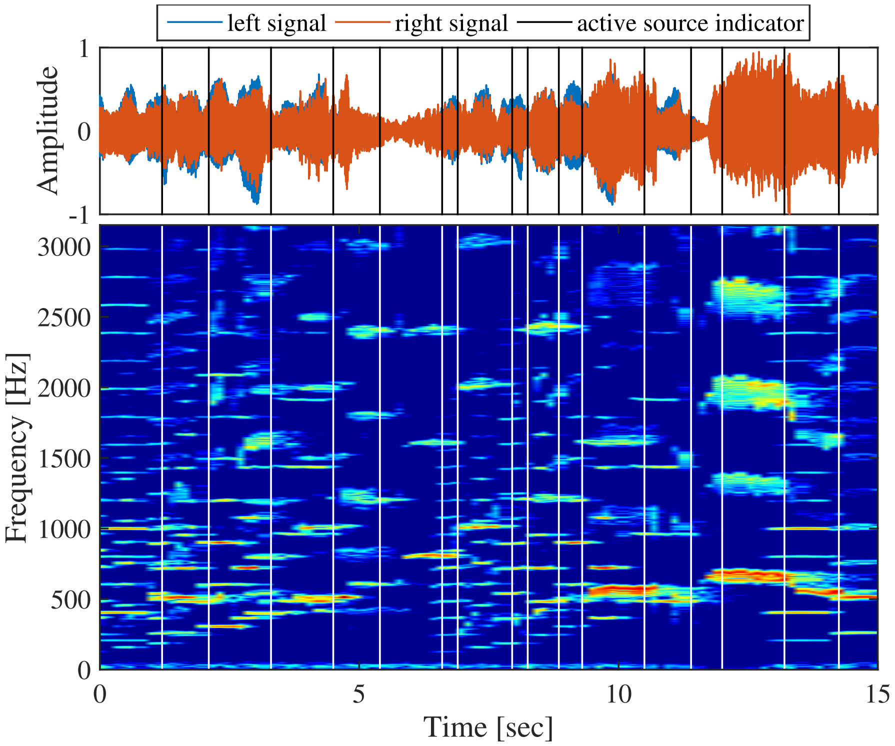


Figure C.11: Signal segmentation on a mixture three sources from the SQAM database.

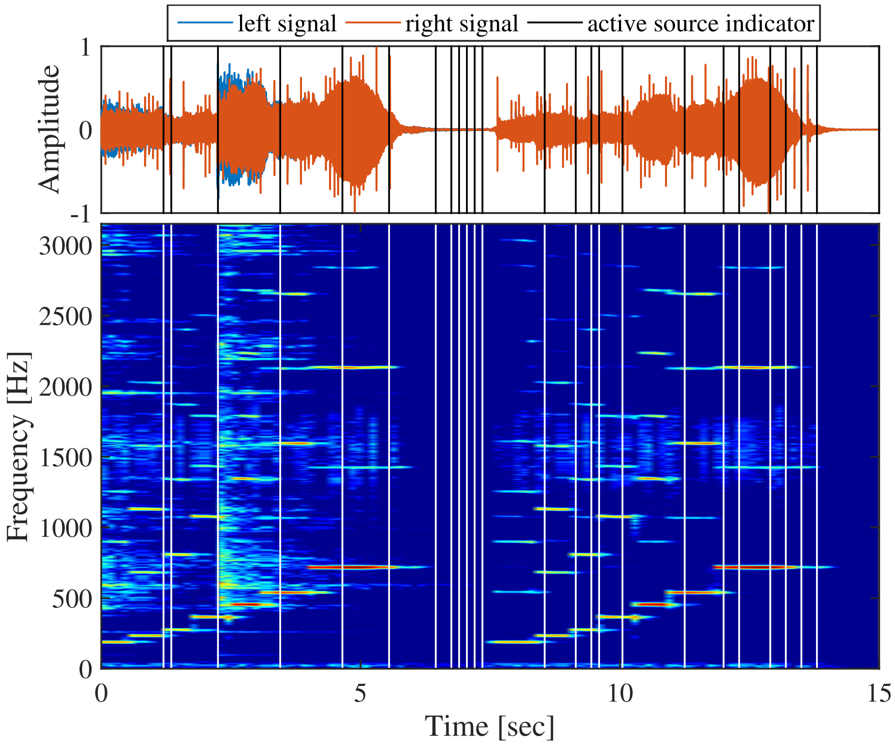


Figure C.12: Signal segmentation on a mixture three sources from the SQAM database.

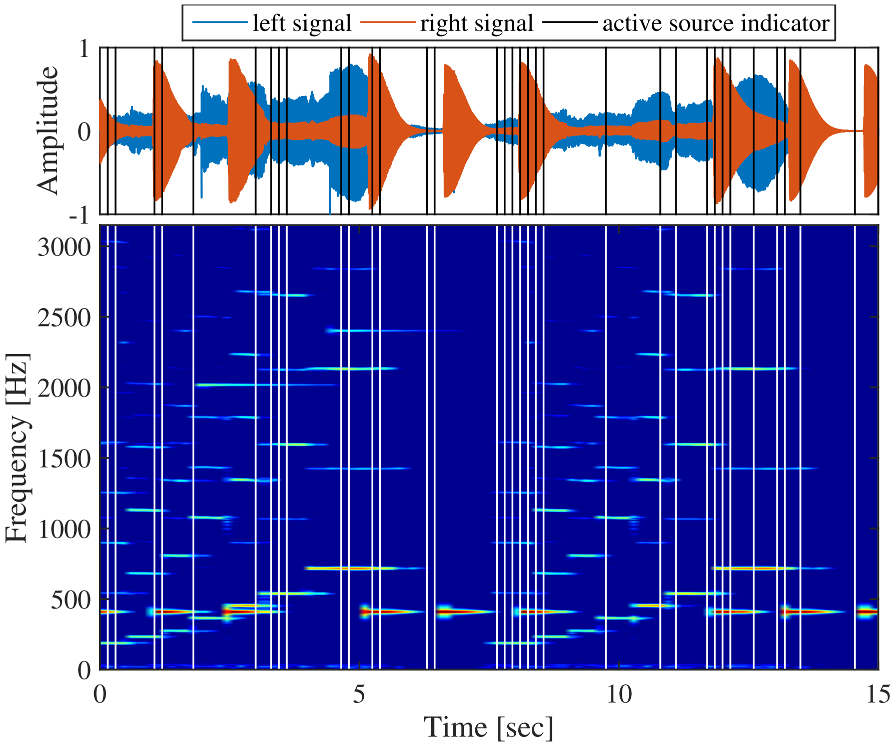


Figure C.13: Signal segmentation on a mixture three sources from the SQAM database.

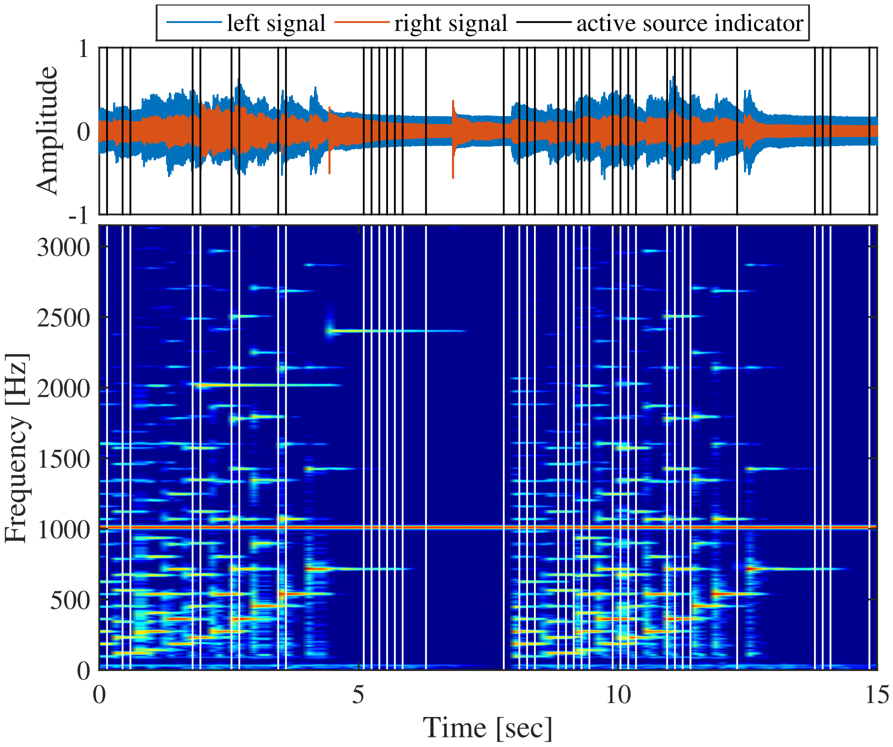


Figure C.14: Signal segmentation on a mixture three sources from the SQAM database.