Single-Channel Blind Source Separation of Audio Signals

Yevgeni Litvin

Single-Channel Blind Source Separation of Audio Signals

Research Thesis

In Partial Fulfillment of The Requirements for the Degree of

Master of Science in Electrical Engineering

Yevgeni Litvin

Submitted to the Senate of the Technion - Israel Institute of Technology

Tichrey 5770 Haifa October 2009

This Research Thesis Was Done Under The Supervision of Prof. Israel Cohen and Dr. Dan Chazan in the Electrical Engineering Department

It was supported by the Israel Science Foundation under Grant 1085/05 and by the European Commission under project Memories FP6-IST-035300.

Acknowledgment

I am grateful to Prof. Israel Cohen and Dr. Dan Chazan for their guidance throughout all stages of this research. I'm also grateful to Prof. Jacob Benesty for his guidance throughout my work on the subject of spectral kurtosis.

Special thanks to my beloved wife Paulina for her love and support.

The generous financial help of The Technion is gratefully acknowledged.

Contents

List of Figures

List of Tables

List of Papers

- Litvin, Y.; Cohen, I., "Single-Channel Source Separation Of Audio Signals Using Bark Scale Wavelet Packet Decomposition," *Proceedings of the 2009 19th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2009.
- Litvin, Y.; Cohen, I.; Chazan, D., "Single-Channel Source Separation Using Discrete Energy Separation Algorithm", submitted for publication to *Signal Processing.*

Abstract

In this thesis we address the problem of audio source separation from a single audio source. Blind source separation (BSS) of audio signals has been an active area of research in recent years. BSS from a single audio channel is a special case of general BSS problem where data from only one source is available to the algorithm. The problem becomes easier if separated audio signals belong to different signal classes that can be classified based upon prior knowledge using existing statistical learning techniques.

Single audio channel BSS is an under-determined problem with arbitrarily many solutions so some assumptions or prior knowledge are required to perform the separation. Statistical independence, signal sparsity, psycho-acoustical properties, statistical models of spectral shapes and its time trajectories are among properties used to distinguish between sources in a mix. Although many existing solutions produce satisfactory results in special cases, the general problem of single audio channel BSS remains unsolved.

We define and study three different algorithms. We note that for some sets of signal classes, the frequency modulating (FM) component of subbands carries discriminative information. For example, this is true in an important case of speech and music signals. This observation motivates our first algorithm. We use time localized energy of the FM component for the classification of time-frequency bins and create a binary mask that is used for rejecting the undesired signal. The difference in the subband FM signal energy of speech and musical signals, together with sparseness and independence of mixture components make

the separation possible. We show that the proposed algorithm exhibits superior performance when compared to a competitive source separation algorithm.

In the second algorithm we use Bark Scaled (BS) Wavelet Packet Decomposition (WPD) analysis. The BS-WPD analysis was previously used in the speech enhancement task. We introduce a modification of the BS-WPD analysis and combine it with an existing BSS algorithm based on the Gaussian Mixture Modeling (GMM). In the first stage of the algorithm, the signal is analyzed using modified BS-WPD analysis and a Gaussian mixture model is trained. In the second stage a mixed signal is separated using the statistical model. The baseline separation algorithm relies on the differences in statistical model parameters. The proposed psycho-acoustically motivated non-uniform filterbank structure reduces feature vectors dimension. It simplifies training procedure of the statistical model and in some scenarios results in better performance.

Finally, we define short time spectral kurtosis (STSK) as a time localized estimate of spectral kurtosis. Our third algorithm uses a value of STSK as a local time-frequency feature for the classification of time-frequency bins. We create a binary mask based on values of STSK. This algorithm relies on differentiating properties of STSK, sparseness and independence of mixed signals. The mask is capable of rejecting the undesired signal. We present good audio source separation experimental results. In our work we use an ad-hoc definition of STSK. A rigorous definition of the STSK and study of its properties may benefit source separation and other applications. These topics are subjects for future research.

Nomenclature

Abbreveations

AM Amplitude Modulation

AQO Audio Quality Oriented

AR Auto Regressive

ASA Audio Scene Analysis

BASS Blind Audio Source Separation

BS Bark Scaled

BS-WPD Bark Scaled Wavelet Packet Decomposition

BSS Blind Source Separation

CASA Computational Auditory Scene Analysis

CB-WPD Critical Band Wavelet Packet Decomposition

CSR-BS-WPD Constant Sampling Rate Bark Scaled Wavelet Packet Decomposition

CWT Complex Wavelet Transform

DESA Discrete Energy Separation Algorithm

DFT Discrete Fourier Transform

dmey Discrete Meyer

DWT Discrete Wavelet Transform

EFMS Energy of a Frequency Modulating Signal

EM Expectation Maximization algorithm

FM Frequency Modulation

GMM Gaussian Mixture Model

HMM Hidden Markov Model

ICA Independent Component Analysis

ICSR-BS-WPD Inverse Constant Sampling Rate Bark Scaled Wavelet Packet Decomposition

IDWT Inverse Discrete Wavelet Transform

ISTFT Inverse Short Time Fourier Transform

LPC Linear Predictive Coefficients

LSD Log Spectral Distance

MAP Maximum Aposterior

ML Maximum Likelihood

MMSE Minimum Mean Square Error

NMF Non-negative Matrix Factorization

p.d.f. probability density function

PM Posterior Mean

PSR Preserved-Signal Ratio

SAR Signal to Artifact Ratio

SBSS Semi-Blind Source Separation Algorithms

SDR Signal to Distortion Ratio

SIR Signal to Interference Ratio

SK Spectral Kurtosis

SNR Signal to Noise Ratio

SO Significance Oriented

STFT Short Time Fourier Transform

STSK Short Time Spectral Kurtosis

TEO Teager Energy Operator

W-DO W-Disjoint Orthogonal

WPD Wavelet Packet Decomposition

Notation

$\angle$ angle of a complex number

$\circ$ elementwise multiplication of vectors or matrices

$\delta(t)$ Kronecker delta function

δ_{Dirac} Dirac delta function

δ_{SK} threshold on STSK used for creation of time-frequency binary mask

δ_E relative energy threshold used for selection of high energy time-frequency bins

$\gamma_{j,k}$ posterior probability of active components j, k in GMM models

$\hat{\Pr}_c(.)$ empirical probability

$\hat{E}_k(m)$ estimate of the EFMS in k-th frequency band

$\hat{}$	an estimate
κ	AM modulation index
λ_{ij}	penalty for assigning a sample to class i when in fact the sample belongs to class j
$\mathcal{F}$	Fourier basis matrix
$\mathcal{K}_x(k)$	spectral kurtosis of x in k-th frequency band
$\mathcal{R}_1, \mathcal{R}_2$	classification decision regions
diag(y)	diagonal matrix with elements of vector y on its diagonal
median $\{S\}$	median of elements in S
μ	vector of expectation values
∇_x	gradient with respect to x
Ω_a	bandwidth of amplitude modulating signal
Ω_c	carrier angular frequency
Ω_f	bandwidth of frequency modulating signal
Ω_i	instantaneous angular frequency
Ω_{if}	intermediate carrier frequency
Ω_s	radial sampling frequency
ϕ_A	variance of centered random variable A
$\Psi[.]$	Teagers Energy Operator (TEO)
$\Re(.)$	real part of a complex number
Σ	covariance matrix
σ^2	vector of variances of i.i.d. random variables

θ initial phase

$\tilde{\Omega}_c$ carrier frequency after frequency shift by STFT filterbank

$\tilde{M}_k^{(1)}(m)$ oracle binary time-frequency mask

$\tilde{w}(n)$ STFT synthesis window

$a(n)$ amplitude modulating signal

$a_{j,k}$ DWT/WPD approximation signal

b_x, b_w bandwidth of x, w

c signal class index

d size of the region used for selection of high energy regions

$d^{(c)}$ amplitude of FM signal

$d_{j,k}$ detail signal in DWT/WPD

$e_{c,\text{artif}}$ algorithm artifacts error signal

$e_{c,\text{interf}}$ interfering signal component error signal

f_c carrier frequency

f_m frequency of FM signal

f_s sampling frequency

$g(n)$ DWT/WPD approximation analysis filter

$h(n)$ DWT/WPD detail analysis filter

h^+ L^2 to Softy-space mapping filter

h_r high-pass filter used to remove DC component in FM signal estimate

J number of DWT decomposition levels

K number of components in GMM model

k frequency index in time-frequency domain

L number of samples in training audio sequence

L_K length of STSK estimation window

M number of overlapping samples in STFT transform

$M_k^{(c)}(m)$ binary time-frequency mask

N support length of STFT analysis window $w(n)$

n time index

$N(\mu, \Sigma)$ Normal distribution with mean μ and covariance Σ

q index of active GMM component

Q_c set of high energy time-frequency bins

$r(i)$ frequency modulating signal

$S_{1,k}(m)$, $S_{2,k}(m)$ time-frequency representation coefficients of mixture components

$s_1(n)$, $s_2(n)$ mixture components

T sampling period

$U(a, b)$ uniform distribution

$w(n)$ discrete STFT analysis window

$w_c(t)$ continuous STFT analysis window

$w_{\mathcal{K}}(m)$ STSK estimator averaging window

$w_a(n)$ analytic bandpass filter generated by shifting analysis window w in frequency

$x(n)$ time domain signal

x^+ Softy-space image of x

$X_k(m)$ time-frequency representation coefficient

$x_c(t)$ continuous time signal

$X_{l,n}(m)$ m-th sample of signal in node (l,n) of wavelet packet decomposition

$x_l(n)$ l-th partial of harmonic signal x

Chapter 1

Introduction

Blind source separation (BSS) is the task of recovering a set of signals from a set of observed signal mixtures. The problem of BSS is common for different signal processing tasks. It is also at the heart of numerous applications in audio signal processing. BSS algorithms that operate on audio signals are sometimes called Blind Audio Source Separation (BASS) algorithms [1].

Cherry [3] coined the ability of the human hearing system to concentrate on a single speaker in the presence of other interfering signals such as other speakers, music or noise as "cocktail party effect". Although, human audio segregation abilities are fascinating, not necessarily a full audio separation is performed in the inner ear or somewhere in the auditory cortex. It is possible that the human hearing system is only cable of recognizing semantic objects in one of several audio streams the listener is exposed to.

Different settings for the BSS task arise in different applications. In different settings the prior and the posterior information available to a source separation algorithm may differ, such as: number of sources and number of observed channels; mixing model (instantaneous, echoic, convolutive, linear, non-linear); prior information on signal statistical properties of signals; presence of noise.

One of the crucial factors in the definition of the BSS problem is the ratio of the number of observed channels to the number of audio sources in the mixture.

If the number of observed channels is equal to the number of extracted sources then it is usually called an even-determined or a determined case. In an over-determined case the number of channels is greater than the number of sources and in an under-determined case the number of channels is smaller than the number of sources. The under-determined case is the most difficult to handle and requires stronger assumptions on the mixture component properties.

Another important factor that differentiates between BSS problem setups is the mixing model. Instantaneous mixing model implies that several instantaneous mixtures are observed, each having source components mixed in a different proportion. Echoic mixing model allows different delays for each component in each channel. The convolutive mixing model allows different linear filtering of sources at each channel. Naturally, instantaneous is a degenerate case of echoic mixing model and echoic is a degenerate case of convolutive mixing model. The convolutive mixing model is the best in describing a real room recording of several audio sources, but is also the most difficult to handle. In more recent works, non linear mixing models are also studied.

Most source separation algorithms assume that mixture components are statistically independent. Although, this is a reasonable assumption in many cases, it is not necessarily true for all applications. For example, one of the source separation applications is separation of individual musical instrument from polyphonic musical excerpt. In this case, the assumption of statistical independence is inaccurate because if musical instruments play statistically independent parts, then there would be no harmonic nor temporal structure and the musical piece would be a cacophony.

In this work we study the most extreme under-determined case when only a single mixture is observed. We assume presence of two mixture components, instantaneous linear mixing model and absence of noise.

1.1 Monaural source separation

Vincent et al. [1] classify BASS tasks into two groups, according to the desired output: Audio Quality Oriented (AQO) and Significance Oriented (SO). Figure 1.1 depicts the taxonomy of BASS applications.

The purpose of AQO applications is to generate an audio signal that can be listened to directly or after some post-processing. The AQO applications are also divided into two groups: "one versus all" and "audio scene modification" applications. The aim of "one versus all" applications is to separate a single audio source from the mixture. The "audio scene modification" applications aim to change mixing proportions without removing mixed components completely. The "one versus all" problem is more difficult since it requires separation of all mixing components. If we have acquired all mixture components then the solution of the "audio scene modification" would be a simple remixing of these components. On the other hand, solution of the "audio scene modification" problem does not provide a solution to the "one versus all" problem.

Some examples of "one versus all" applications are: separation of individual musical instrument tracks from a polyphonic mixture; speech enhancement and de-reverberation; restoration of old recordings [4]; object-based audio coding. Some examples of "audio scene modification" are: remixing of existing audio recordings and signal enhancement in hearing aids.

Significance oriented applications usually do not aim at extracting audio sources but to extract features that are necessary for some cognitive functionality. Some examples of SO applications are speech recognition in the presence of other sources [5]; speaker identification in presence of other sources [6]; polyphonic music transcription [7]; musical instrument identification in polyphonic music [8].

Different approaches rely on one or more properties of mixture components to perform separation such as statistical independence, sparseness, certain spectral and temporal structure. Following is a short description of several single channel

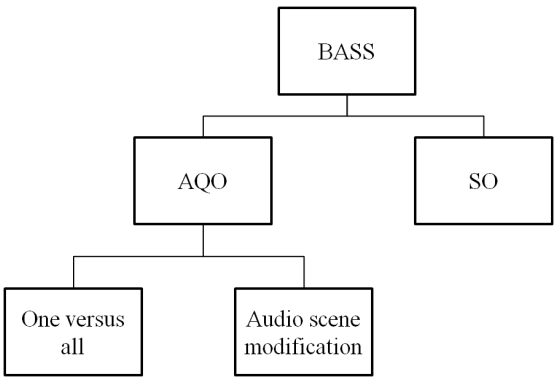


Figure 1.1: BASS tasks taxonomy [1]

source separation approaches.

Co-channel speaker separations

Some single channel source separation algorithms assume that both sources contain speech signals. Early attempts to solve this problem stem form speech enhancement algorithms that are designed to separate speech from a background noise using pitch information [9]. Hanson [10] implemented a co-channel speaker separation system that first estimated the pitch of one of the talkers and then used harmonic information and spectral subtraction technique to separate two speech signals. In order to be effective in the presence of loud noises or interfering speakers this kind of approach usually requires more complex algorithms which take into consideration the possibility of noise or the presence of two pitches. These techniques aim at separating harmonic parts of speech, hence they are also applicable to musical instrument separation from polyphonic musical signals.

Independent component analysis and sparse decomposition

The blind source separation problem was first formulated in a statistical framework by Herault et al. [11] in 1984. Comon [12] introduced the Independent Component Analysis (ICA) in 1994 and numerous theoretical and practical works followed. A basic ICA algorithm assumes even-determined BSS case and

instantaneous mixing model. Under these assumptions, a demixing matrix has to be found. In order to find such matrix the ICA algorithm minimizes statistical dependency between unmixed channels. Various methods may be used in order to reduce statistical dependency, such as maximization of non-Gaussianity between channels or minimization of mutual information [13]. The search is usually done using gradient descent or fixed point algorithms. Unfortunately, most algorithms in the ICA family require several mixtures to be observed in order to perform the separation.

If a signal at hand is known to be sparse in some domain, then a sum of two sparse signals will be less sparse than its components. The BSS algorithms that use this property looks for a matrix that will produce sparsest signals after demixing [14]. Unfortunately, like in the ICA case, such algorithms require several observed channels to be available.

Computational auditory scene analysis

Many algorithms that deal with source separation of audio signals are based on results acquired in psychoacoustical studies. Bregman's book "Auditory scene analysis: The perceptual organization of sound" [15] contains various psychoacoustical studies and provides a basis for the computational implementation of algorithms that mimic behavior of human auditory apparatus. Computational implementations of psychoacoustic rules are known as Computational Auditory Scene Analysis (CASA). A particularly interesting aspect of ASA is the segregation of audio signal into separate audio streams using segregation cues. For example, if two different frequency sine tones have the same onset in time, then according to the Audio Scene Analysis (ASA) principles they belong to the same audio stream, i.e. produced by the same source.

An example of an algorithm that uses CASA approach is presented by F.R. Bach and M.I. Jordan in [16]. They use spectral clustering algorithm to assign time-frequency bins of the STFT to two different audio sources. Distances between every two time-frequency bins are defined using ASA motivated seg-

regation cues. For example, two time-frequency bins likely belong to the same source if they are adjacent in time or frequency. After a similarity matrix is created a clustering algorithm assigns each time-frequency point to one of two classes. The interfering source is removed using a binary mask created using time-frequency bin assignments and the demixed component is recovered in the time domain.

T. Virtanen and A. Klapuri use sinusoidal modeling to separate several harmonic sounds [17]. First, they use peak tracking to model the entire mixture as a set of sinusoid trajectories. Amplitude, frequency modulations and the probability of two trajectories to have same fundamental frequency are combined into a single similarity measure that is used later for clustering.

Semi-blind source separation

In some cases a database of audio samples is available and statistical signal models can be trained in a supervised manner before the separation process. In this case various techniques from statistical learning can be used. Algorithms that rely on these kind of statistical models are sometimes called Semi-Blind Source Separation Algorithms (SBSS) [18, 19].

In [2], Benaroya introduced a source separation algorithm based on GMM and Hidden Markov Model (HMM) statistical modeling of source signal classes. First GMM or HMM models are trained for each signal class using spectral shapes acquired from the STFT analysis. During the separation stage, these models are used to estimate mixture components using Maximum A-posterior (MAP) or Posterior Mean (PM) estimates. Authors also showed that using more complicated HMM model does not improve the separation performance significantly when compared to the GMM model. Some extensions to that work were presented in [19]. For example, Gaussian Scaled Mixture Model which takes into account variations in amplitude of sounds with similar spectral shapes.

Ozerov et al. [18] proposed to use GMM model adaptation. Model adaptation is successfully used in speaker recognition applications [20]. In the experi-

mental results, the authors demonstrated their method by separating a singing voice from the accompanying musical instruments. Model adaptation was performed during signal excerpts when no vocal was present.

Another signal modeling technique that was found useful in single channel source separation is Auto Regressive (AR) modeling. Srinivasan et al. [21] proposed a codebook of Linear Predictive Coefficients (LPC) trained on speech and interfering signal. Their approach suggests using maximum likelihood estimation to find the most probable pair of codebook members. Wiener filter is used later to suppress the interfering signal. In [22] LPC coefficients are treated as random variables. In these works both algorithms are described and tested in the setup of speech enhancement in the presence of non stationary noise. Nevertheless, they are also applicable to the source separation scenario by modeling one of the sources as speech and the other as noise.

All algorithms presented above in the context of SBSS have common signal codebook modeling approach. They all have a deterministic, stochastic or even adaptive codebook at the core of the algorithm. A detailed description of codebook methods for source separation can be found in [23].

Non-negative matrix factorization

Non-negative Matrix Factorization (NMF) [24] can be applied to mixture magnitude spectrogram $\mathbf{X} \approx \mathbf{AS}$. After the factorization, columns of $\mathbf{A}$ contain frequency basis functions and $\mathbf{S}$ contains representation coefficients. Assuming different audio sources have different spectral characteristics and by assigning frequency basis function (columns of $\mathbf{A}$) to different audio sources we can reconstruct individual sources [25]. An example of NMF based algorithm which incorporates inter-frame temporal continuity prior and uses sparse prior for the NMF proposed by Hoyer [26] can be found in [27].

When using NMF for source separation, each audio source is represented by several frequency basis functions. In order to reconstruct source signals some sort of clustering must be performed on the columns of $\mathbf{A}$. In case of

musical instruments separation, different notes have different basis function, even when played using the same instrument, although their frequency shapes are similar up to some scaling in frequency. FitzGerald et al. [28, 29] presents a modification of the NMF based approach by using positive tensor factorization. Their approach makes it possible to use a single frequency basis function for a wide range of notes played by the same instrument, hence eliminating the need for clustering and reducing redundancy in matrix $\mathbf{A}$. A different approach that aims to solve frequency basis function redundancy in the harmonic musical instrument separation that also adds sparsity optimization goal, can be found in [30].

1.2 Overview of the thesis

In this work we focus on blind and semi-blind source separation of signals. We use audio signals in our experiments, but some of the proposed methods can be extended to other types of signals, such as neural signals, seismic, financial, images and others. In this section we briefly describe the original contribution of this thesis.

An AM-FM decomposition (the FM component in particular) of real signal classes (e.g. speech and music) subbands carries discriminative information about signal class. First we define new time-frequency signal space where each coefficient defines the subband frequency modulating signal energy. In the training stage, we learn a simple statistical model of the coefficients for each signal class. We create a binary mask in the STFT domain and use it to recover mixture components. FM signal energy was found to be a good differentiating factor when speech and musical signals are concerned. Sparseness of audio signals in the STFT domain together with the statistical independence make it possible to use binary mask to suppress an interfering signal. We compare the performance of this method to an existing GMM based source separation algorithm. The experimental results obtained using the proposed method are

significantly better compared to those obtained using the best known current approach where the discrimination is based on differences in spectral properties.

We present a source separation algorithm that uses novel time-frequency analysis method based on a Bark Scaled Wavelet Packet Decomposition (BS-WPD)[31]. The original BS-WPD was modified in order to adapt it for the source separation task: we introduced shiftability to the BS-WPD transform using mapping based complex wavelet transform (Fernandes et al., 2003 [32]) and modified a BS-WPD subsampling scheme to achieve similar sampling frequencies at all subbands at the expense of representation redundancy. We used an existing GMM based source separation algorithm [2] with the new signal analysis. The proposed analysis method results in spectral vectors of reduced dimension, hence allows simpler statistical modeling. Psycho-acoustically motivated filterbank structure also results in better perceptual quality of the separated signals. Experimental results showed improved performance compared to an existing GMM algorithm that uses Short Time Fourier Transform (STFT) in some scenarios and comparable performance in other scenarios. The complexity of the separation algorithm was reduced because of smaller dimension of the vector space achieved by coarser frequency resolution in high frequencies.

Finally, we study a possible application of spectral kurtosis to the task of source separation from a single sensor. We define short time spectral kurtosis (STSK) and its ad-hoc estimator. We use it to create an STFT domain binary mask capable of rejecting interfering signal. The STSK provides means for local time-frequency bins classification. Sparseness and independence of audio sources make it possible to use binary masks for separation. Although, the separation algorithm introduced is extremely simple, the experimental results show very good performance compared to a GMM based source separation algorithm. The use of spectral kurtosis is relatively new in the field of audio signal processing. Our experiments suggest that it can be used successfully for source separation from a single channel.

1.3 Organization

The organization of this thesis is as follows. In Chapter 2 we define a source separation problem from a single channel. We present GMM based Bayesian approach together with GMM based source separation algorithm. This separation algorithm is used extensively throughout the thesis as a baseline for comparison and a prototype to the algorithm presented in one of the following chapters. In Chapter 3 we show a way to estimate energy of the frequency modulating signal and present a separation algorithm based on the subband AM-FM decomposition of the mixture signal. In Chapter 4 we present a constant sampling rate Bark scaled wavelet packet decomposition (CSR-BS-WPD) and a source separation algorithm. In Chapter 5 we define Short Time Spectral Kurtosis (STSK), study some properties of spectral kurtosis and present simple source separation algorithm that uses STSK together with experimental study of the proposed algorithm. Finally, in Chapter 6 we conclude our work and propose some directions for future research.

Chapter 2

Bayesian Source Separation

2.1 Introduction

Benaroya et al. [19] presented Bayesian formalism for the source separation problem. They showed that in the case of two signals and only one observed mixture the probabilistic formalism leads to multiple solutions, hence the problem is underdetermined. On the other hand, when Bayesian formalism is used, prior assumptions on the distribution of sources can be incorporated into the problem. Prior assumptions resolve the ambiguity of the probabilistic formalism. Benaroya et al. studied a case of Gaussian, generalized Gaussian distribution and Gaussian Mixture Models.

Gribonval et al. [33] proposed a number of performance measures for BSS applications. These performance measures take into account special properties of BSS algorithms such as ability to recover a source up to a multiplicative constant in multichannel BSS algorithms; distinguish different types of distortions: those caused by an interfering signal; and others caused by artifacts introduced by an algorithm.

The remainder of this chapter is organized as follows. In Section 2.2 we formulate a monaural source separation problem. In Section 2.3 we shortly present Benaroya's approach to source separation previously published in [19].

We present the GMM based source separation algorithm used as a baseline for comparison in the following chapters. Section 2.4 presents performance evaluation measures used in our experimental results.

2.2 Problem formulation

In this section we formally define the problem of single channel source separation under the assumption of instantaneous mixing and noise absence.

Let $s_1(n)$ and $s_2(n)$ be time domain signals that belong to different signal classes. Let $x(n)$ be a mixture of $s_1(n)$ and $s_2(n)$

$$x(n) = s_1(n) + s_2(n) \tag{2.1}$$

The problem of source separation is defined as finding estimates $\hat{s}_1$, $\hat{s}_2$.

The same problem can also be defined in the frequency domain using STFT transform (A.3)

$$X_k(m) = S_{1,k}(m) + S_{2,k}(m)$$

where k is the frequency band index and m is the time index. Once we find estimates for $\hat{S}_1$ and $\hat{S}_2$ ISTFT transform (A.4) can be used to obtain time estimates of the components.

There are several benefits in solving the separation problem in the STFT domain. The STFT analysis window length is usually selected in a way that results in almost stationary and circular signals (i.e., has a Toeplitz covariance matrix) in each analysis window. As we will show in Section 2.3 this simplifies evaluation of the Maximum posterior (MAP) estimator. Besides, the coefficients of speech and music signals in the STFT domain are sparse and for two independent sources, most of the signal energy is located in the non-overlapping coefficients. This property was studied by Yilmaz et al. in [34] and was coined there as approximate W-disjoint orthogonal (W-DO). More details about W-

DO can be found in Appendix B. This property justifies separation of signal mixtures using simple binary masks in the STFT domain.

2.3 Mixture components estimation in Bayesian framework

In [19] Benaroya et al. develop Bayesian formulation of source estimation under the assumption that both sources are Gaussian and then extend the estimation to the generalized Gaussian and Gaussian mixture distribution case. In this section we shortly present the results that are relevant to our work.

2.3.1 Normally distributed mixture components

Let s_1 and s_2 be two vectors in $\mathbb{R}^N$. Assume s_1, s_2 are Normally distributed independent random variables with zero mean and Σ_1, Σ_2 covariance matrices. The p.d.f. function in this case is given by

$$p_c(s) = \frac{1}{(2\pi)^{N/2}|\Sigma_c|^{1/2}} \exp\left(-\frac{1}{2}s^T\Sigma_c^{-1}s\right) \qquad c \in \{1,2\}$$

Assume the linear mixture of s_1, s_2 is observed

$$x = s_1 + s_2 \tag{2.2}$$

Our goal is to find estimates $\hat{s}_1, \hat{s}_2$ of s_1, s_2.

Likelihood of the observation is given by

$$p(x|s_1, s_2) = \delta_{\text{Dirac}}(x - (s_1 + s_2)) \tag{2.3}$$

It is clear that Maximum Likelihood (ML) estimator

$$(\hat{s}_1, \hat{s}_2)_{\text{ML}} = \arg\max_{s_1,s_2} p(x|s_1, s_2)$$

will not produce any meaningful results since any $\hat{s}_1 = x - \hat{s}_2$ maximizes the likelihood. The prior knowledge can be introduced by using MAP estimation.

Due to source independence, the a-priori probability can be factored

$$p(s_{1,}s_2) = p(s_1)\, p(s_2)$$

and MAP estimator is given by

$$\begin{aligned}(\hat{s}_1, \hat{s}_2)_{\text{MAP}} &= \arg\max_{s_1,s_2} p(s_{1,}s_2|x) \\ &= \arg\max_{s_1,s_2} p(x|s_{1,}s_2)\, p_1(s_1)\, p_2(s_2)\end{aligned}$$

Likelihood function (2.3) imposes a constraint $x = s_1 + s_2$ under which estimation of MAP estimator $\hat{s}_1$ reduces to

$$\begin{aligned}\hat{s}_1 &= \arg\max_{s_1} p_1(s_1)\, p_2(x - s_1) \\ &= \arg\min_{s_1} \left(-\log p_1(s_1) - \log p_2(x - s_1)\right) \\ &= \arg\min_{s_1} \left(\frac{1}{2} s_1^T \Sigma_1^{-1} s_1 + \frac{1}{2}(x - s_1)^T \Sigma_2^{-1}(x - s_1)\right) \\ &\triangleq \arg\min_{s_1} J(s_1)\end{aligned}$$

J is a quadratic function in s_1 and since Σ_1 and Σ_2 are positive semi definite, J has a single minimum in s_1.

$$\nabla_{s_1} J(s_1) = \Sigma_1^{-1} s_1 - \Sigma_2^{-1}(x - s_1)$$

The minimum is found by solving $\nabla J = 0$ in respect to s_1

$$0 = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right) s_1 - \Sigma_2^{-1} x$$

which results in

$$\hat{s}_1 = (\Sigma_1 + \Sigma_2)^{-1} \Sigma_1 x$$

MAP estimator $\hat{s}_2$ can be found in the same way:

$$\hat{s}_2 = (\Sigma_1 + \Sigma_2)^{-1} \Sigma_2 x$$

If we assume that s_1 and s_2 are stationary and approximately circular processes, the covariance matrices Σ_1, Σ_2 are Toeplitz and diagonalized by Fourier basis vectors. Let $\mathcal{F}$ be discrete Fourier transform operator. Define $S_c \triangleq \mathcal{F} s_c, X \triangleq \mathcal{F} x$. The distribution of S_1, S_2, X are also Gaussian and given by

$$\begin{aligned} S_c &\sim N\left(0, \operatorname{diag}\left(\sigma_{s_c}^2\right)\right) \quad c \in \{1, 2\} \\ X &\sim N\left(0, \operatorname{diag}\left(\sigma_{s_1}^2 + \sigma_{s_2}^2\right)\right) \end{aligned}$$

MAP estimator of $\hat{S}_1$ is given by

$$\hat{S}_1(i) = \frac{\sigma_1^2(m)}{\sigma_1^2(m) + \sigma_2^2(m)} X(m) \tag{2.4}$$

We see that MAP estimator (2.4) coincides with Posterior Mean (PM) estimator (Wiener filter) in case of Gaussian prior on mixture components.

2.3.2 Gaussian mixture distribution of the mixture

Simple Gaussian prior assumption on the signal distribution does not hold for most real signals such as speech or music. One solution is to assume Gaussian Mixture prior densities (GMM prior) [19].

GMM model describes signal distribution as an outcome of a two stage

process: first an active component k is selected out of K Gaussian distributions in the mixture; then we draw an observation sample using the selected model parameters $\left\{\mu^{(k)}, \Sigma^{(k)}\right\}$ where $\mu^{(k)}$ and $\Sigma^{(k)}$ are the expectation value and the covariance matrix of the k-th component. The probability of selecting k-th component is given by w_k (k-th element of probability vector w). The GMM model is defined by $\left(\left\{\mu^{(k)}\right\}_{k=1}^{K}, \left\{\Sigma^{(k)}\right\}_{k=1}^{K}, w\right)$.

We introduce two hidden variables $q_1(m)$ and $q_2(m)$ in order to estimate mixture components using GMM prior. In the rest of this subsection we omit time index m to simplify the notation. q_1 and q_2 are associated with active component of GMM models of both signals at time m. Mixture component estimation reduces to simple Gaussian case described in the previous section when conditioned on values of q_1 and q_2. We denote $\gamma_{j,k} = p(q_1 = j, q_2 = k|x)$. The MMSE estimator for the mixture component $\hat{S}_1$ is given by

$$\hat{S}_1 = \sum_{j,k=1}^{K} \gamma_{j,k} \Sigma_1^{(j)} \left(\Sigma_1^{(j)} + \Sigma_2^{(k)}\right)^{-1} X \tag{2.5}$$

The estimator for $\hat{S}_2$ is derived in the same manner.

The MAP estimator is acquired by first evaluating $(j^*, k^*) = \arg\max_{j,k} \gamma_{j,k}$ and

$$\hat{S}_1 = \Sigma_1^{(j^*)} \left(\Sigma_1^{(j^*)} + \Sigma_2^{(k^*)}\right)^{-1} X \tag{2.6}$$

The value of $\gamma_{j,k}$ is given by

$$\begin{aligned} \gamma_{j,k} &\propto p(X|q_1 = j, q_2 = k)\, p(q_1 = j)\, p(q_2 = k) \\ &= g\left(X; \Sigma_1^{(j)} + \Sigma_2^{(k)}\right) w_1^{(j)} w_2^{(k)} \end{aligned} \tag{2.7}$$

Algorithm 2.1 GMM based source separation algorithm

Training

1. Compute time frequency representation $S_{1,k}(m), S_{2,k}(m)$ of training signals $s_1(n), s_2(n)$ (A.3)
2. Train Λ_1, Λ_2 GMM models using data vectors $|S_{1,k}(m)|, |S_{2,k}(m)|$ and EM algorithm.

Separation

1. Compute time frequency representation $X_k(m)$ of mixed signal $x(n)$ (A.3).
2. For all time indexes m
   (a) For all pairs $(j,k) \in \{(j,k) | j \in \{1,\ldots,K\}, k \in \{1,\ldots,K\}\}$
      i. Compute $\gamma_{j,k}(m)$ using (2.7)
   (b) Estimate $\hat{S}_1, \hat{S}_2$ using (2.5) or (2.6)
3. Compute estimates of mixture components in time domain using (A.4)

2.3.3 GMM based source separation algorithm

Now we combine results from the previous section and repeat the definition of the GMM based source separation algorithm (Algorithm 2.1) published by Benaroya and Bimbot in [2].

The training stage is performed offline, given a database of signals from two different classes. For each signal class, a time-frequency representation of signals is obtained using the STFT transform (A.3). The GMM model of spectral magnitude vectors for each time frame is trained using EM algorithm [35] with the assumption of diagonal covariance matrix.

In the separation stage, the STFT transform is applied to the mixture. Then, the value of γ is calculated for all possible active GMM component combinations and estimates of both source signals are obtained using PM estimator (2.5) or MAP estimator (2.6) in the time-frequency domain. The separated component signal estimators are recovered using ISTFT (A.4).

This algorithm is used as a baseline for comparison of novel algorithms proposed in this thesis. In Chapter 3.1 we compare this algorithm to an algorithm

based on locate time-frequency FM properties of signals. In Chapter 4 we study the effect Bark-Scaled signal analysis has on the separation quality and in Chapter 5 we compare it to a novel algorithm that uses Spectral Kurtosis values for classification of time-frequency bins and signal separation.

2.4 Evaluation criteria

In this section we define evaluation criteria used in experiments to evaluate the performance of the proposed algorithms. We use common distortion measures described in [33] and BSS_EVAL toolbox [36]. Mixture components s_1, s_2 are assumed to be uncorrelated. Let $\hat{s}_c$ be an estimate of s_c. The estimator will have the following decomposition:

$$\begin{aligned}
\hat{s}_c &= y_c + e_{c,\text{interf}} + e_{c,\text{artif}} \\
y_c &\triangleq \langle \hat{s}_c, s_c \rangle \, s_c \\
e_{c,\text{interf}} &\triangleq \langle \hat{s}_c, s_{c'} \rangle \, s_{c'} \\
e_{c,\text{artif}} &\triangleq \hat{s}_c - (y_c + \langle \hat{s}_c, s_{c'} \rangle \, s_{c'})
\end{aligned}$$

where c is the target class and c' is the interfering class. Now the following criteria are defined:

$$\begin{aligned}
\text{SDR} &\triangleq 10 \log_{10} \frac{\|y_c\|^2}{\|e_{c,\text{interf}} + e_{c,\text{artif}}\|^2} \\
\text{SIR} &\triangleq 10 \log_{10} \frac{\|y_c\|^2}{\|e_{c,\text{interf}}\|^2} \\
\text{SAR} &\triangleq 10 \log_{10} \frac{\|y_c + e_{c,\text{interf}}\|^2}{\|e_{c,\text{artif}}\|^2}
\end{aligned}$$

Signal to Distortion Ratio (SDR) measures the total amount of distortion introduced to the original signal, both due to the interfering signal and artifacts introduced by the algorithm. Signal to Interference Ratio (SIR) measures the amount of distortion introduced to the original signal by the interfering signal.

Signal to Artifact Ratio (SAR) measures the amount of artifacts introduced to the original signal by the separation algorithm that do not originate in the interfering signal.

Usually some algorithm working point can be chosen to tune the trade-off between interfering signal leakage (SIR) and the distortion to the desired signal (SAR). For example it is possible to reduce SIR to $-\infty$ simply by zeroing source estimation. However, the SAR measure will become very high in this case. SDR is a kind of cumulative measure for both SIR and SAR, hence it is convenient to compare algorithm performance based on SDR.

Two additional measures used are Log Spectral Distance (LSD)

$$\mathrm{LSD}\left(X,Y\right) := \sqrt{\sum_{k=1}^{K}\sum_{m=1}^{N}\left(20\log_{10}\frac{\left|X_k\left(m\right)\right|}{\left|Y_k\left(m\right)\right|}\right)^2}$$

where $X_k\left(m\right)$, $Y_k\left(m\right)$ are compared signals in the STFT domain and Signal to Noise Ratio (SNR)

$$\mathrm{SNR}\left(\hat{s},s\right) \quad = \quad 10\log_{10}\frac{\sum_{n=1}^{|s|}s\left(n\right)^2}{\sum_{n=1}^{|s|}\left(\hat{s}\left(n\right)-x\left(n\right)\right)^2}$$

where $|s|$ denotes number of samples in s.

2.5 Summary

We have formulated the source separation problem from a single channel. We have demonstrated how a mixture can be separated using MAP estimation when Gaussian or Gaussian mixture distribution are assumed on the mixture components. We also presented a source separation algorithm that relies on the assumption of Gaussian mixture distribution of mixture components. This algorithms is used as a comparison baseline in the following chapters and also inspires the algorithm presented in Chapter 4. Finally, we described the evaluation measures used in the following chapters to compare the performance of

some novel source separation techniques. We use these performance measures throughout the thesis.

Chapter 3

Subband Frequency Modulating Signal Modeling

3.1 Introduction

In [37, 38] H. M. Teager and S.M. Teager studied airflow and fluid dynamics of human speech apparatus. They described several nonlinear phenomena as well as their sources. Later, Kaiser formulated Teager Energy Operator (TEO) [39, 40]. In [41, 42, 40] the TEO was used to derive a discrete energy separation algorithm (DESA) that separates a signal into its amplitude (AM) and frequency modulating (FM) components. Applications of the AM-FM decomposition of audio signals include formant tracking [43], enhancement of speech recognition and speaker recognition features [44, 45, 46, 47], speech coding [48], analysis and re-synthesis of musical instruments sound [49].

Sinusoidal modeling was previously used for BSS by Virtanen and Klapuri [17]. Their approach requires peak tracking in the spectral domain to establish sinusoidal trajectories followed by grouping of detected trajectories into different audio streams. Although, our approach can also be seen as a kind of sinusoidal modeling, it does not require peak tracking or grouping which may improve the

robustness of the separation algorithm.

In [34], Yilmaz et al. define an approximate W-disjoint orthogonality (W-DO) as an approximate "disjointness" of several signals in the short-time Fourier transform (STFT) domain. They suggest a quantitative W-DO measure and provide evidence of the high level of the W-DO for two speech signals. Their work provides a theoretical basis for speech signal separation using time-frequency bins binary making. Refer to Appendix B of this work for additional details.

In this chapter, we propose a source separation algorithm capable of segregating several audio sources from a single channel based on differences of FM components statistical properties. We use Discrete Energy Separation Algorithm (DESA) to estimate frequency-modulating (FM) signal energy. We create time varying filter in the time-frequency domain which is capable of rejecting the interfering signal. The estimation of the FM signal energy uses instantaneous signal properties that are localized both in time and frequency. We present experimental results and demonstrate feasibility of our approach both on synthetic and real audio signals and compare our results to a competitive source separation algorithm. Although we demonstrate our algorithm on speech and piano play signals, the proposed algorithm is applicable to other large classes of audio signals as well.

The core idea of the modulation frequency analysis and filtering is analysis and modification of the subband amplitude modulating (AM) signal. If STFT analysis of subband AM signal is performed, then the resulting signal domain is called joint frequency domain. An application of joint frequency analysis and modification include monaural source separation [50, 51] and speech enhancement [52, 53]. Significant amount of attention was payed to different AM demodulation techniques [52, 54, 55] when an emphasis was made to find an appropriate demodulation method that would allow signal modification in the joint frequency domain. A comprehensive survey of this field can be found in Schimmel's work [52]. Our work relates to the joint-frequency analysis in the sense that we are interested in the FM component and not in the AM component

of the subband AM-FM decomposition.

Our algorithm uses an AM-FM analysis. First we filter the input signal by an STFT filterbank. Then we use the DESA algorithm to estimate a frequency modulating signal in each of the subbands for a given instant in time and an energy of the frequency modulating signal (EFMS). In the training stage a statistical model of EFMS values of all frequency bands is learned for each signal class. In the separation stage, time-frequency bins in the STFT domain are classified into one of target signal classes using EFMS values. The interfering signal is suppressed by zeroing time-frequency bins attributed to the interfering signal. Finally, we reconstruct the separated component by inverting the STFT.

The remainder of this chapter is structured as follows. In Section (3.2) we present the TEO operator and DESA algorithm. In Section 3.3 we describe the estimation of EFMS and present some examples of real audio signal. We explain why the proposed method should perform well at the separation task. Section 3.4 describes a simple training procedure used to learn EFMS features of various audio classes and Bayesian risk minimization approach used to create a STFT domain binary mask that filters out the interfering signal. Section 3.5 presents evaluation of the proposed algorithm performance. The summary is given in Section 3.6.

3.2 Energy separation algorithm

In this section we present continuous and discrete forms of a Teager Energy Operator (TEO). We also present AM-FM decomposition algorithms: continuous time Energy Separation Algorithm (ESA) and discrete time Discrete Energy Separation Algorithm (DESA) for the discrete signals [40].

3.2.1 Continuous signals

Let $x_c(t)$ be a continuous time signal. In his work, Teager [37, 38] noted the importance of analyzing speech from the point of view of the energy required

to generate the signal. He used a non-linear energy tracking operator Ψ_c and its discrete counterpart Ψ. These operators were systematically introduced by Kaiser [39, 56].

$$\Psi_c\left[x_c\left(t\right)\right] = \left(\dot{x}_c\left(t\right)\right)^2 - x\left(t\right)\ddot{x}\left(t\right)$$

For a undriven linear undamped oscillator with an amplitude A, e.g. a body of a mass m and a spring of constant k, the instantaneous total energy (kinetic and potential) is given by

$$\begin{aligned} E_{\text{osc}} &= \frac{1}{2}m\dot{x}_c^2 + \frac{1}{2}kx_c^2 \\ &= \frac{1}{2}m\left(A\omega_0\right)^2 \end{aligned} \tag{3.1}$$

where $\omega_o = \sqrt{k/m}$ is an oscillation angular velocity. The position of the body is described by a solution to the equation $m\ddot{x} + kx = 0$ and is given by

$$x_c\left(t\right) = A\cos\left(\omega_0 t + \theta\right)$$

The TEO of $x_c\left(t\right)$ evaluates to

$$\begin{aligned} \Psi_c\left[x_c\left(t\right)\right] &= \left(\frac{d}{dt}A\cos\left(\omega_0 t + \theta\right)\right)^2 + A\cos\left(\omega_0 t + \theta\right)\frac{d^2}{dt^2}A\cos\left(\omega_0 t + \theta\right) \\ &= A^2\omega_0^2\sin^2\left(\omega_0 t + \theta\right) + A^2\omega_0^2\cos^2\left(\omega_0 t + \theta\right) \\ &= 2A^2\omega_0^2 \end{aligned} \tag{3.2}$$

which is proportional to (3.1). If the amplitude $A\left(t\right)$ or the angular velocity $\omega_0\left(t\right)$ vary in time, then under certain conditions described in [42],Ψ_c (and its discrete version Ψ) can track the energy of that signal.

ESA is a simple method which aims to separate amplitude and frequency

modulation components from a continuous time signal

$$\begin{aligned} x_c(t) &= a(t)\cos(\phi(t)) \\ \phi(t) &\triangleq \omega_c t + \omega_m \int_0^t r(\tau)\, d\tau + \theta \end{aligned}$$

The instantaneous angular frequency

$$\begin{aligned} \omega_i &\triangleq \frac{d}{dt}\phi(t) \\ &= \omega_c + \omega_m r(t) \end{aligned}$$

Let $a(t)$ and $r(t)$ be band limited signals with ω_a and ω_f highest non-zero angular frequencies of $a(t)$, $r(t)$ respectively. Let κ be the AM index. Both, the instantaneous frequency and the amplitude components, contribute to the value of Ψ_c as can be seen in (3.2).

Under conditions

$$\omega_a \ll \omega_c \quad \text{and} \quad \kappa \ll 1 \tag{3.3}$$

$$\omega_f \ll \omega_c \quad \text{and} \quad \omega_m/\omega_c \ll 1 \tag{3.4}$$

the following two equations separate these components and define the ESA algorithm [40]:

$$\omega_0 \approx \sqrt{\frac{\Psi_c[\dot{x}_c(t)]}{\Psi_c[x_c(t)]}} \tag{3.5}$$

$$|A| \approx \frac{\Psi_c[x_c(t)]}{\sqrt{\Psi_c[\dot{x}_c(t)]}} \tag{3.6}$$

3.2.2 Discrete signals

Let $x(n) = x_c(nT)$ be a sampled version of $x_c(t)$ where T is the sampling period. A discrete version of TEO (Ψ) defined:

$$\Psi\left[x\left(n\right)\right] \quad = \quad x^{2}\left(n\right)-x\left(n-1\right)x\left(n+1\right)$$

In the discrete signal case we assume the following signal model

$$x\left(n\right) \quad = \quad a\left(n\right)\cos\left(\Omega_{c}n+\sum_{i=0}^{n}r\left(i\right)\frac{1}{T}+\theta\right)$$

where n is a discrete time index, Ω_c is a carrier angular frequency, θ is some constant phase value and $a\left(n\right)$, $r\left(n\right)$ are amplitude and frequency modulating signals respectively.

Similarly to (3.5), (3.6), $\Psi\left[x\left(n\right)\right]$ is used to estimate the instantaneous frequency $\hat{\Omega}_{i}\left(n\right)$ and the instantaneous amplitude $\hat{a}\left(n\right)$:

$$\begin{aligned}\hat{\Omega}_{i}\left(n\right) &\approx \frac{1}{2}\arccos\left(1-\frac{\Psi\left[x\left(n+1\right)-x\left(n-1\right)\right]}{2\Psi\left[x\left(n\right)\right]}\right) \qquad (3.7)\\ &\approx \Omega_{c}+q\left(m\right)\\ \left|\hat{a}\left(n\right)\right| &\approx \frac{2\Psi\left[x\left(n\right)\right]}{\sqrt{\Psi\left[x\left(n+1\right)-x\left(n-1\right)\right]}} \qquad (3.8)\end{aligned}$$

Conditions equivalent to (3.3), (3.4) in the discrete case are:

$$\Omega_{a}\ll\Omega_{c} \quad \text{and} \quad \kappa\ll 1 \qquad (3.9)$$

$$\Omega_{f}\ll\Omega_{c} \quad \text{and} \quad \frac{\sup\left\{r\left(n\right)\right\}}{\Omega_{c}}\ll 1 \qquad (3.10)$$

where Ω_{a},Ω_{f} are the bandwidths of $a\left(n\right)$ and $r\left(n\right)$ respectively and κ is an AM modulation index ($a\left(n\right)$ assumed to be positive). Several versions of DESA algorithm are described in [40] . The difference between different versions of DESA algorithm is the way time derivatives of $x\left(n\right)$ are estimated. Equations (3.7), (3.8) define DESA-2 algorithm.

3.3 Energy of frequency modulating signal

In this section we demonstrate frequency modulation analysis on some examples of speech and piano signals. We define the energy of the frequency modulating signal (EFMS). We show that the EFMS of speech and piano signals can be used as local time-frequency discriminating factor which can be used to reject the interfering source. These examples will motivate formulation of our algorithm.

Partials of voiced phonemes in speech signals have a stronger frequency modulating component than partials of piano signals. In order to define an algorithm that exploits this property we need to formulate a quantitative measure for this phenomenon.

Let $x(n)$ be a time signal. We assume it is an harmonic signal with one or more harmonic partials present. We treat each partial as a separate carrier. Most of the AM-FM demodulation algorithms, including DESA, cannot deal with multiple carriers being present in the analyzed signal. In order to apply the analysis we note that each of the signals produced by filtering the analyzed signal through a narrow band filterbank is likely to contain a single FM modulated carrier. In our work we use STFT filterbank.

The STFT transform of $x(n)$ is given by (A.2). Let us repeat the definition here as well for completeness:

$$X_k(m) = \sum_{n=-\infty}^{\infty} w(mM - n)\, x(n)\, e^{-j\frac{2\pi}{N}kn} \tag{3.11}$$

where $w(n)$ is the analysis window with support of N and bandwidth of b_w radian and N, M define frequency and time resolution of the transform. Equation (3.11) can be rewritten in a filter like form

$$X_k(m) = e^{-j\frac{2\pi}{N}kmM}\,(x * w_a)(mM) \tag{3.12}$$

where $w_a(n)$ is an analytic bandpass filter generated by shifting $w(n)$ in frequency by $2\pi k/N$ radians.

The time series $X_k(m)$ indexed by m, can be treated as a time domain, bandpass version of the analytic signal of $x(n)$ with bandpass center frequency shifted to zero. We assume that only a single partial is present in $X_k(m)$. This allows us to use AM-FM decomposition algorithm. In the AM-FM decomposition, each harmonic will act as a carrier. Instantaneous deviations from the carrier frequency (caused by intonation in speech and speech production nonlinearities) will appear as a frequency-modulating signal.

3.3.1 Energy of frequency modulating signal

Assume the AM-FM model for the l-th harmonic partial

$$x_l(n) = a(n)\cos\left(\Omega_c n + \sum_{i=0}^{n} r(i)\frac{1}{T} + \theta\right) \tag{3.13}$$

Let b_x be $x_l(n)$ bandwidth. Assume that the $x_l(n)$ energy is found almost entirely in the k-th band of the STFT filterbank which results in the approximation.

$$x_a(n) \approx x_l(n) * w_a(n) \tag{3.14}$$

Where $x_a(n)$ is an analytic signal of $x_l(n)$. Equation (3.14) can hold only approximately since the theoretical bandwidth of a frequency modulated signal is infinite. We can also write

$$\Omega_c + \frac{b_x}{2} < \frac{2\pi}{N}k + \frac{b_w}{2} \quad \bigcap \quad \Omega_c - \frac{b_x}{2} > \frac{2\pi}{N}k - \frac{b_w}{2} \tag{3.15}$$

$$\left|\frac{2\pi}{N}k - \Omega_c\right| < \frac{b_w - b_x}{2} \tag{3.16}$$

After modulating $x_a(n)$ by a complex exponent $e^{-j2\pi kn/N}$ and decimation

by a factor of M, the output of the STFT filterbank (3.12) is given by

$$X_k(m) \approx a(mM) \exp j\left(\tilde{\Omega}_c mM + \sum_{i=0}^{mM} r(i)\frac{1}{T} + \theta\right) \tag{3.17}$$

$$\tilde{\Omega}_c = \Omega_c - \frac{2\pi}{N}k \tag{3.18}$$

$\tilde{\Omega}_c$ is close to zero and from (3.16), (3.18) yields $\left|\tilde{\Omega}_c\right| < b_w/2$. Since the bandwidths of $a(n)$ and $r(n)$ remain unchanged the DESA algorithm assumptions (3.9), (3.10) no longer hold. In the notations of this section

$$\begin{aligned} \Omega_f &\ll \tilde{\Omega}_c \\ \Omega_a &\ll \tilde{\Omega}_c \end{aligned}$$

The remedy is to modulate the filterbank output to some intermediate frequency Ω_{if} by multiplying $X_k(m)$ by $e^{j\Omega_{\text{if}}m}$

We choose $\Omega_{\text{if}} = \frac{\pi}{3}$ (shift $X_k(m)$ by $\frac{\pi}{3}\left[\frac{rad}{sec}\right]$) i.e. we set a new carrier frequency to be in the lower $\frac{1}{3}$-rd of the frequency axis so as to minimize the risk of aliasing (the choice of $\Omega_{\text{if}} = \frac{\pi}{3}$, (e.g. instead $\Omega_{\text{if}} = \frac{\pi}{2}$) was dictated by better experimental results). DESA operates on the real valued signals, we use only the in-phase component of the modulated filterbank output

$$\tilde{X}_k(m) = \Re\left(X_k(m)\, e^{j\Omega_{\text{if}}m}\right) \tag{3.19}$$

In order to avoid aliasing during modulation and in-phase component extraction the following conditions must hold

$$\Omega_{\text{if}} \geq \frac{b_x M}{2} \tag{3.20}$$

$$\Omega_{\text{if}} \leq \pi - \frac{b_x M}{2} \tag{3.21}$$

Assume that the bandwidth of b_x is equal to the STFT subband bandwidth,

i.e.

$$b_x = 2\pi/N \tag{3.22}$$

and the location of the intermediate-frequency is arbitrary

$$\Omega_{\text{if}} = \alpha\pi \tag{3.23}$$

for some $0 < \alpha < 1$. Substituting (3.22), (3.23) into (3.20), (3.21) results in the following bound on M:

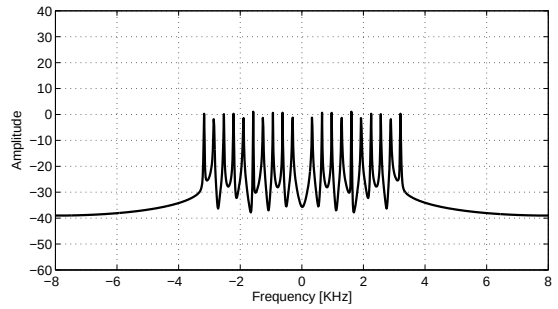
$$M \leq \min\{\alpha N, (1-\alpha)N\} \tag{3.24}$$

Fig. 3.1 shows an example of the processing steps. A synthetic harmonic signal with ten partials is used in this example. First partial is an FM modulated signal. The FM modulating signal is a sinusoid having an amplitude of 2π and a frequency of 10 Hz. The Fourier transform of the signal is shown in Fig. 3.1(a). Most of the energy of the first partial is located in the 21-st band. The Fourier transform of $X_{21}(m)$ is shown in Fig. 3.1(b). $X_{21}(m)$ is a complex signal, hence positive and negative frequencies of the Fourier transform are not complex conjugate. Fig. 3.1(c) shows the Fourier transform of $\tilde{X}_{21}(m)$. $\tilde{X}_{21}(m)$ is a real valued signal modulated to the intermediate frequency. The dashed line shows regions of the spectrum originally filtered out by w_a.

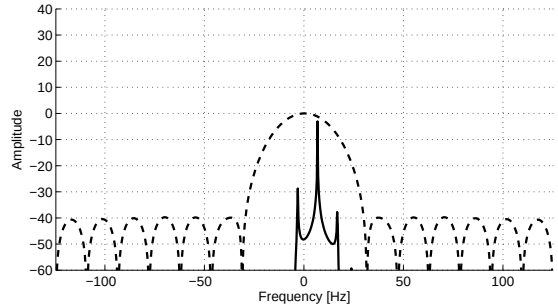
DESA estimator (3.7) can now be used to find the FM component of $\tilde{X}_k(m)$

$$\hat{\Omega}_{i,k}(m) \approx \frac{1}{2}\arccos\left(1 - \frac{\Psi\left[\tilde{X}_k(m+1) - \tilde{X}_k(m-1)\right]}{2\Psi\left[\tilde{X}_k(m)\right]}\right)$$

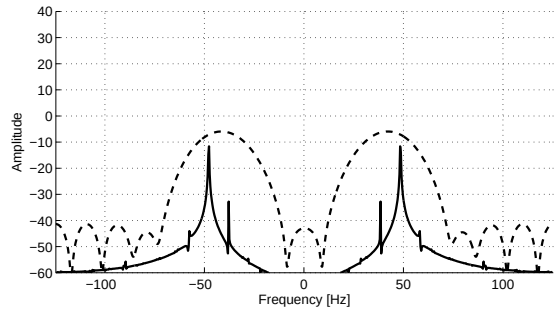
The instantaneous frequency $\hat{\Omega}_i$ also includes a slowly varying $\tilde{\Omega}_c + \Omega_{\text{if}}$ term. To remove it we filter $\hat{\Omega}_i$ with a high pass filter h_r which results in an estimate of $r(n)$. We note that Ω_c is not necessarily constant in time, but assume that



(a)



(b)



(c)

Figure 3.1: Input signal preprocessing for the DESA algorithm. A Dashed line shows which portions of the spectrum were originally filtered out by $w_a(n)$. (a) input signal that contains 10 carriers. (b) frequency domain representation of the signal at the STFT filterbank output ($X_k(m)$). (c) STFT filterbank output modulated to the intermediate frequency ($\tilde{X}_k(n)$).

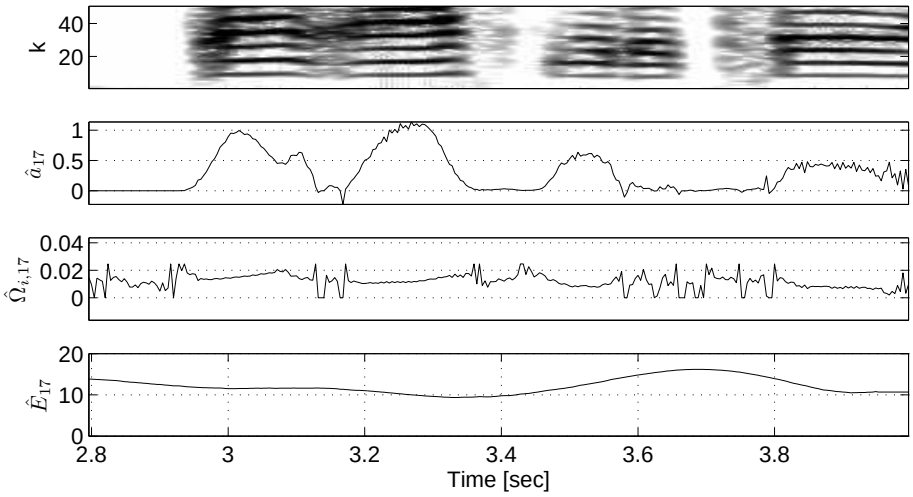


Figure 3.2: Upper pane shows the spectrogram (50 lower frequency bands) of the "she had" utterance. Vertical axis labels show frequency band numbers. Second pane shows the estimated AM component of the 16-th frequency band ($\hat{a}_{16}$). Third pane shows the instantaneous frequency estimation $\hat{\Omega}_{i,16}$ of the 16-th frequency band. Lower pane shows the EFMS ($\hat{E}_{16}(n)$).

it changes slowly compared to $r(n)$.

$$\begin{aligned} \hat{r}(m) &\approx \left(\hat{\Omega}_i * h_r\right)(m) \\ &\approx \left(\left(\tilde{\Omega}_c + \Omega_{\text{if}} + r(n)\right) * h_r\right)(m) \end{aligned}$$

We define the EFMS by

$$\hat{E}_k(m) \triangleq \left(u * \hat{r}_k^2\right)(m) \tag{3.25}$$

where $u(m)$ is an N_u points Hamming window which purpose is to reduce the variance of the energy estimator $\hat{r}_k^2(m)$. In the rest of the paper we denote the EFMS of a time signal $x(n)$ by $\hat{E}\{x\}_k(m)$ and omit x and indices k, m when the meaning is clear from the context.

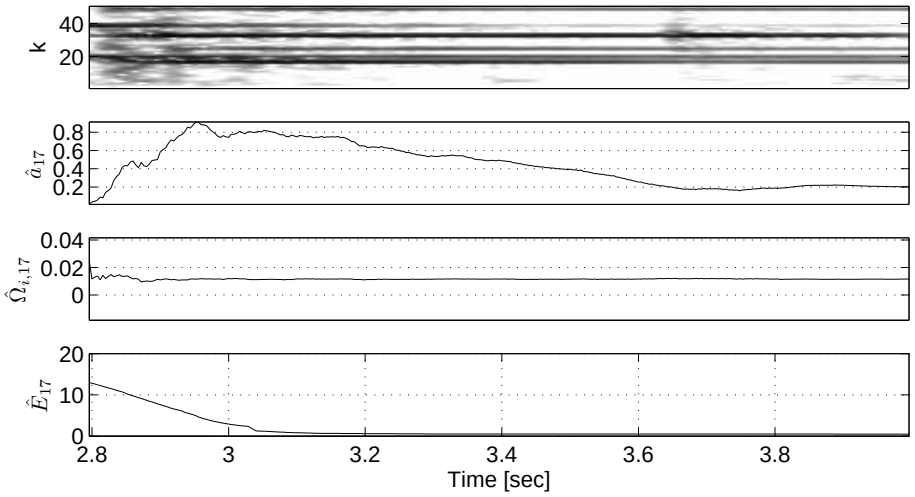


Figure 3.3: Upper pane shows the spectrogram (50 lower frequency bands) of the piano play sample. Vertical axis labels show frequency band numbers. Second pane shows the estimated AM component of the 17-th frequency band ($\hat{a}_{17}$). Third pane shows the instantaneous frequency estimation $\hat{\Omega}_{i,16}$ of the 20-th frequency band. Lower pane shows the EFMS ($\hat{E}_{16}(n)$).

3.3.2 EFMS analysis of real signals

Figure 3.2 shows a speech fragment containing the utterance "don't ask me to carry". The upper pane shows the 50 lower frequency bands of the STFT filterbank output. First six harmonic partials are visible. We manually pick 16-th frequency band which contains the second partial for some period of time. The second pane shows amplitude envelope $\hat{a}_{16}(m)$ of the selected frequency band estimated by the DESA algorithm. There are several amplitude peaks corresponding to voiced phonemes. Third pane shows $\hat{\Omega}_{i,16}$ estimate. The lowest pane shows EFMS $\hat{E}_{16}(m)$ values. In voiced parts of the speech fragment the energy of the FM component is high. Fricative and plosive phonemes are not described well by the AM-FM model and DESA estimate of the instantaneous frequency has high variance at these locations. The result is high values of EFMS at /sh/ and /d/ phoneme locations. This observation is consistent with our claim that the EFMS of speech is higher than the EFMS of piano play.

The piano play fragment depicted in Fig. 3.3 contains several piano notes. As in the previous case, we manually pick a frequency band that contains a single

harmonic partial. We take the 17-th band and perform the same analysis. We observe that $\hat{E}_{17}(m)$ values are low while the note is being played, hence we have the evidence that a piano produces audio signals with low EFMS. We speculate from the examination of Figs. 3.2 and 3.3 that it is harder to discriminate signal classes by the shape of amplitude envelope than by the shape of instantaneous frequency and EFMS $\hat{E}$.

3.3.3 EFMS analysis of synthetic signals

In the next example, we apply EFMS analysis to synthetic signals: a harmonic signal (x_1) and white noise with unit variance (x_2). The harmonic signal has fundamental frequency $f_0 = 250$ Hz and $N_p = 30$ partials. Let p denote the index of a partial. The carrier frequency and the amplitude of the frequency modulating signal of p-th partial are $f_0 \cdot p$ and $A_0 \cdot p$. Both grow linearly with the index of the partial, like in a speech or a musical signal. The frequency f_{FM} of the FM component is fixed $f_{\mathrm{FM}} = 10$ Hz.

$$x_1(n) = \sum_{p=1}^{N_p} x_{1,p}(n)$$

$$x_{1,p}(n) = \cos\left(2\pi f_0 pn + \sum_{i=0}^{n} q_p(n)\frac{1}{T}\right)$$

$$q_p(n) = 2\pi A_0 p \cos(2\pi f_{\mathrm{FM}} n)$$

Fig. 3.4 shows the distribution of EFMS values for every value of frequency (only values of EFMS that are located at time-frequency bins that have high energy participate in this analysis. The exact method for selecting these frequency bins is described in Section 3.4.1). The amplitude of the FM signal grows linearly with the index of the partial. In the case of a sinusoidal signal, the square root of signal energy is proportional to its amplitude. The dashed line in Fig. 3.4 shows theoretically predicted values of $\sqrt{\hat{E}}$. It is given by $A_0 f/\sqrt{2} f_0$. Ac-

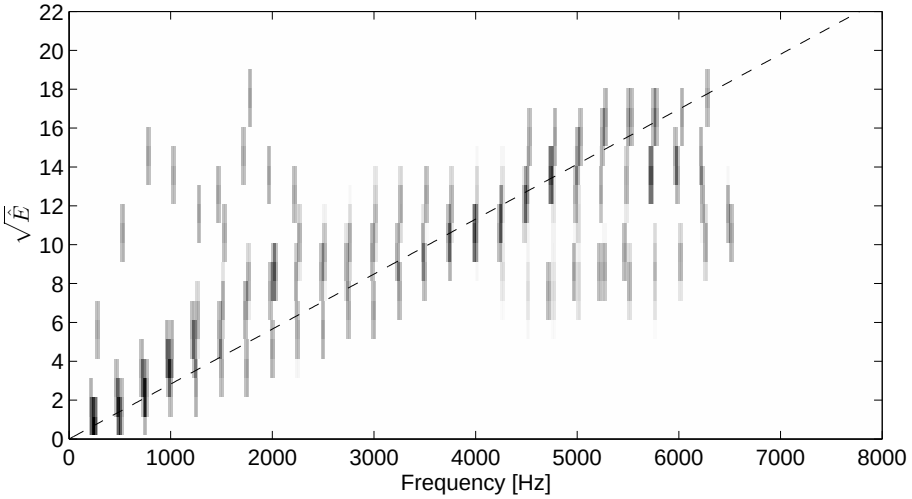


Figure 3.4: Distribution of the EFMS values for the synthetic signal (x_1) having 30 partials with linearly increasing amplitude of frequency modulating component. A dashed line shows theoretically predicted values.

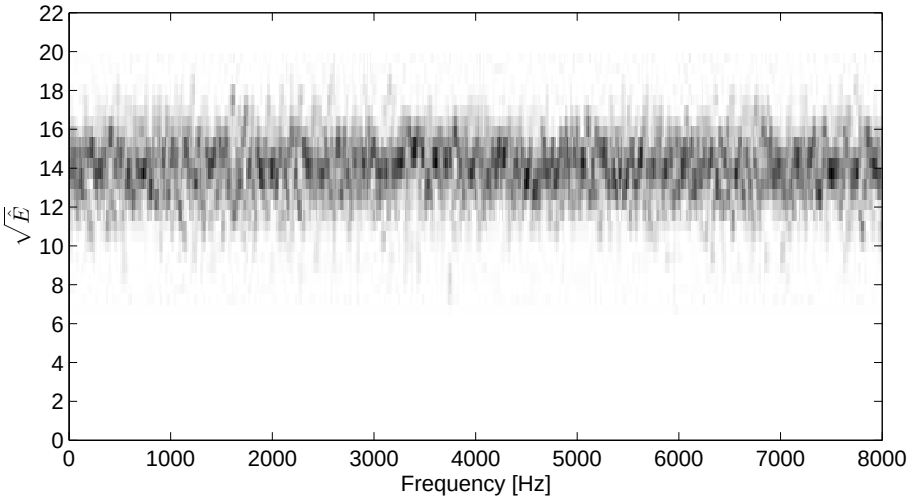


Figure 3.5: Distribution of EFMS values for white noise (x_2).

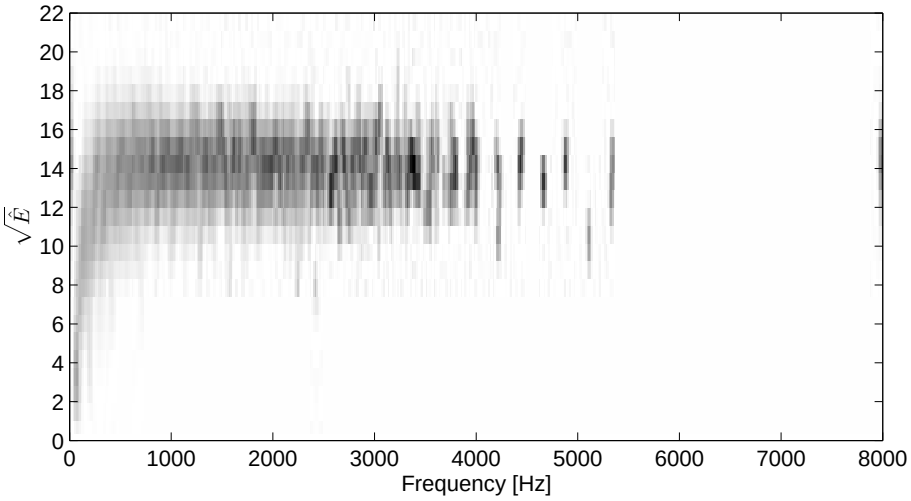


Figure 3.6: Distribution of EFMS values for speech signal.

tual values of $\sqrt{\hat{E}}$ are located in the vicinity of theoretically predicted values, but not exactly on it. There are several reasons for the mismatch:

- Bandpass filtering of a frequency modulated signal alters its sidebands. This results in distortion of the FM modulating signal. This is especially true for high frequency partials: their bandwidth is relatively high due to the high amplitude of the modulating signal.

- Partials that "leak" to neighboring bands have low SNR levels and result in EFMS estimates similar to EFMS of white noise.

White noise signal is not described well be the AM-FM model. The resulting EFMS values for all frequency bands are distributed randomly around some constant value as can be seen in Fig. 3.5.

Figs. 3.6 and 3.7 show EFMS distributions for a speech and a piano signal respectively. The EFMS analysis of speech resembles white noise for frequency greater than 500 Hz. Smaller values of EFMS are present under 500 Hz but nevertheless they are generally higher than EFMS values of a piano play. The piano play has relatively low values of EFMS that grow approximately linearly with frequency, as was predicted by the harmonic signal model.

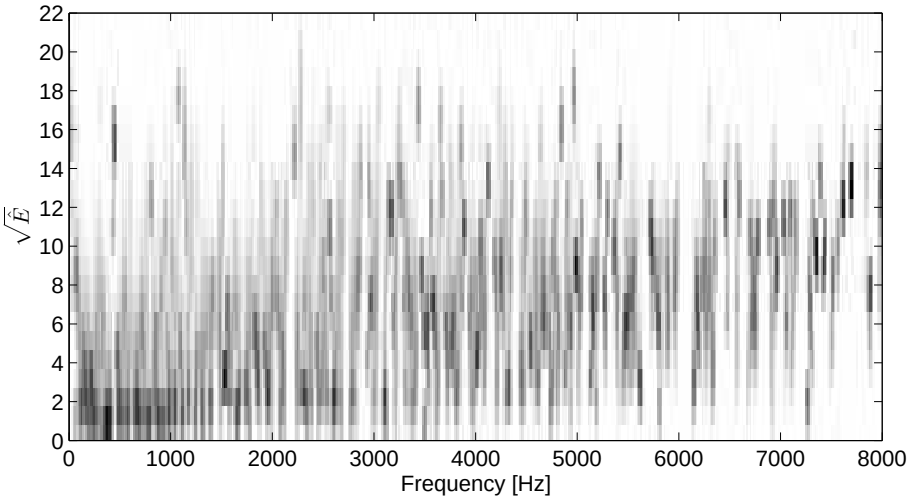


Figure 3.7: Distribution of EFMS values for piano play (x_2).

3.4 Source separation procedure

We assume mixing model as in (2.1). As in previous sections, we denote STFT transform by capital letter, e.g. STFT of $s_c(n)$ is denoted by $S_{c,k}(m)$, where $c \in \{1, 2\}$ denotes the signal class index.

In the training stage we find the empirical probability density function for $\hat{E}\{s_1\}$ and $\hat{E}\{s_2\}$. In the separation stage we use estimated pdf to define a minimum risk decision rule for classification of STFT time-frequency bins based on $\hat{E}\{x\}$.

3.4.1 Training

The empirical probability density function for class c ($\hat{\Pr}_c\left(\hat{E}\right)$) is estimated using a normalized histogram of

$$\left\{\hat{E}\{s_c\}_k(m) \,|\, (k,m) \in Q_c\right\}$$

where Q_c is a set of time-frequency bin indices where the energy is high compared to the neighboring bins. Let $M_{k,m}$ be the median of energy values in the time-frequency vicinity of (k, m) bin

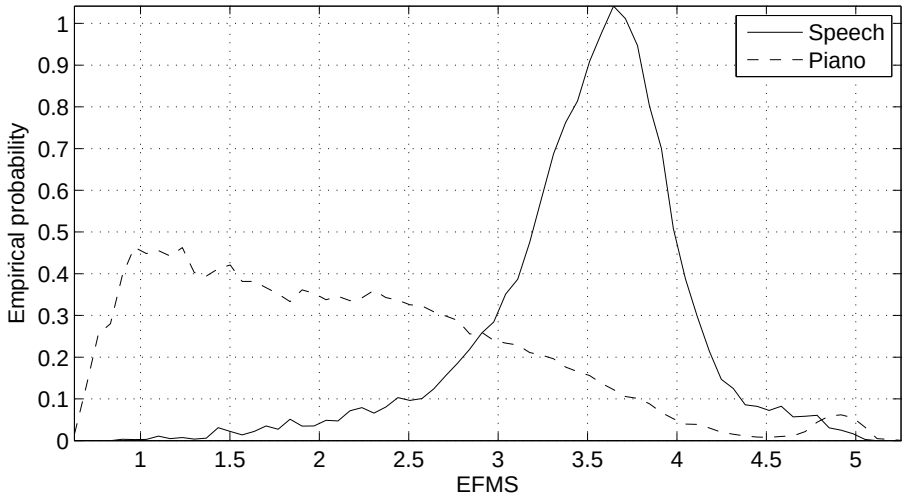


Figure 3.8: Empirical probability density function. EFMS of piano play have higher probability obtaining low values then EFMS of speech.

$$M\left\{S_c\right\}_{k,m} \quad = \quad \text{median}\left\{\left|S_{c,i}\left(j\right)\right|^2 \,|\, |i-k| \leq d, |j-m| \leq d\right\}$$

where d defines the vicinity.

δ_E is a threshold that defines which energy values are considered high. Q_c is given by

$$Q_c \quad \triangleq \quad \left\{(k,m) \left| 20\log_{10}\frac{\left|S_{c,k}\left(m\right)\right|}{M\left\{S_c\right\}_{k,m}} \geq \delta_E \right.\right\} \tag{3.26}$$

Fig. 3.8 shows empirical p.d.f. of $\sqrt{\hat{E}}$ for speech and piano play signals. Large non overlapping areas indicate that the separation of these signals using only $\hat{E}\left\{x\right\}$ values should be possible.

3.4.2 Separation

In this section we explain how we use of Bayes minimum-cost decision rule with reject option ([57] see 2.11.13-14) to construct optimal STFT binary mask, the

filtering process and the recovery of demixed source signal estimate.

Denote by $\xi_k(m) = \hat{E}\{x\}_k(m)$ the estimated value of EFMS. Let $H_k^{(c)}(m)$ be a hypothesis that signal from source c is present in (k,m) time-frequency bin. We will omit indices (k,m) for brevity where possible. Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be the classification decision regions for different classes and $\mathcal{R}_r$ a rejection region, i.e. we prefer not to assign the sample into either class. Let λ_{ij} be a penalty for assigning a sample ξ to class i when in fact the sample belongs to class j and λ_r be a penalty for rejecting a sample. We define a loss function

$$L = \int_{\mathcal{R}_1} \lambda_{12} p\left(H^{(2)}|\xi'\right) p(\xi')\, d\xi' + \int_{\mathcal{R}_2} \lambda_{21} p\left(H^{(1)}|\xi'\right) p(\xi')\, d\xi' + \int_{\mathcal{R}_r} \lambda_r p(\xi')\, d\xi'$$

In order to minimize this loss function, the decision regions should satisfy

$$\begin{aligned}
\xi \in \mathcal{R}_i &\iff \begin{cases} \lambda_{ij} p\left(H^{(j)}|\xi\right) p(\xi) < \lambda_{ji} p\left(H^{(i)}|\xi\right) p(\xi) \\ \lambda_{ij} p\left(H^{(j)}|\xi\right) p(\xi) < \lambda_r p(\xi) \end{cases} \\
\xi \in \mathcal{R}_r &\iff \lambda_r p(\xi) \le \lambda_{ij} p\left(H^{(j)}|\xi\right) p(\xi) \\
& i,j \in \{1,2\}; i \neq j
\end{aligned}$$

Defining $\eta \triangleq \frac{p\left(\xi|H^{(1)}\right)p\left(H^{(1)}\right)}{p\left(\xi|H^{(2)}\right)p\left(H^{(2)}\right)}$ we rewrite the classification decision rule using Bayes formula as

$$\xi \in \mathcal{R}_r \iff \begin{cases} \frac{\lambda_r}{\lambda_{12}} \le \frac{1}{1+\eta} \\ \frac{\lambda_r}{\lambda_{21}} \le \frac{1}{1+1/\eta} \end{cases} \tag{3.27}$$

$$\xi \in \mathcal{R}_1 \iff \begin{cases} \frac{\lambda_{12}}{\lambda_{21}} < \eta \\ \frac{\lambda_r}{\lambda_{12}} > \frac{1}{1+\eta} \end{cases} \tag{3.28}$$

$$\xi \in \mathcal{R}_2 \iff \begin{cases} \frac{\lambda_{12}}{\lambda_{21}} > \eta \\ \frac{\lambda_r}{\lambda_{21}} > \frac{1}{1+1/\eta} \end{cases} \tag{3.29}$$

We can tune the algorithm by changing values of $\lambda_{12}, \lambda_{21}, \lambda_r$.

In order to decrease the number of class 1 time-frequency bins that are classified falsely as class 2 we may increase value of λ_{12}. This will result in higher penalty for this kind of error on the one hand (less false alarm errors), but on the other hand more time-frequency bins that truly belong to class 1 will now be classified as class 2 (more misdetect errors). In other words, the SIR of class 1 decreases and SAR increases.

If we decrease λ_r, more time-frequency bins will be rejected, i.e. not assigned to any of the signal classes. This increases the number of time-frequency bins that cannot be classified reliably and decreases the number of time-frequency bins of the interfering signal in both audio sources simultaneously. In other words, SAR increases and SIR decreases for signals of both classes. In our application, actual values of $\lambda_{12}, \lambda_{21}, \lambda_r$ are tuned manually.

We design a binary mask in the STFT domain by assigning each time-frequency bin to one of the signal classes based on (3.27)-(3.29). Time-frequency bins that are assigned to the interfering source or rejected are zeroed and those assigned to the desired signal are set to 1. In order for the binary mask to be effective, we assume that approximate W-disjoint orthogonality [34] holds. The quantitative measure of the W-DO property is described in Appendix (B). We verify the W-DO assumption in our experimental results in Section 3.5.

Binary masks are defined by

$$M_k^{(c)}(m) = \begin{cases} 1 & \xi_k(m) \in \mathcal{R}_c \\ 0 & \text{otherwise} \end{cases} \qquad c \in \{1, 2\} \tag{3.30}$$

The interfering source is removed by multiplying STFT transform of the mixture by $M^{(c)}$

$$\hat{X}_k^{(c)}(m) = M_k^{(c)}(m) X_k(m) \tag{3.31}$$

Inverse STFT transform (A.4) gives time domain estimate of the demixed source

$$\hat{x}^{(c)}(n) = \text{ISTFT}\left\{\hat{X}_k^{(c)}(m)\right\} \tag{3.32}$$

We conclude this section with the summary of the our algorithm shown in Algorithm (3.1). In the training stage, we learn the distribution of EFMS values for both signal classes. We need to identify time-frequency bins that contain the target signal, otherwise, the distribution of the EFMS values will be effected by irrelevant values of EFMS originate from time-frequency regions containing only noise. In 1.a time-frequency regions with high energy are found, the EFMS values are estimated in 1.b and p.d.f. modeled using normalized histogram.

In the separation stage, the EFMS values are estimated in the entire time-frequency space. Each time-frequency bin is classified into either class or rejected based on the p.d.f. estimated in 1.c. Finally, a binary mask is created using the classification decision in 2.b. and the time domain version of separated signal is obtained by masking and inverse STFT transform.

3.5 Experimental results

In this section we present experimental results. First we verify the feasibility of source separation on synthetic signals and then we separate two real audio recordings of speech and piano play. We compare performance of the proposed algorithm to an existing source separation algorithm (see Section (2.3.3).

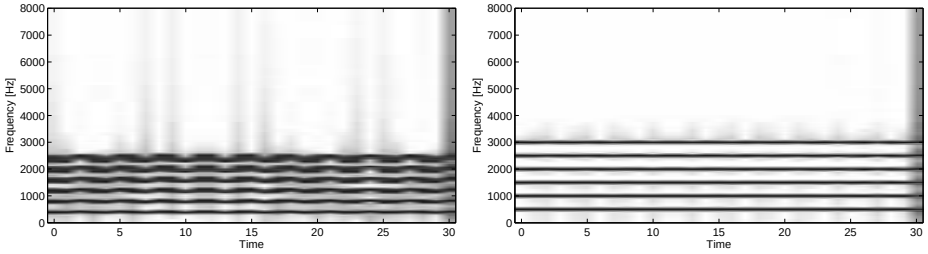
Algorithm 3.1 EFMS based source separation algorithm

Training

1. For each signal class $c \in \{1, 2\}$
   (a) Find a set of high energy time-frequency bins S_c using equation (3.26)
   (b) Estimate EFMS values $\hat{E}$ for all time-frequency bins in S_c using equation (3.25)
   (c) Estimate p.d.f. of $\hat{E}$ using normalized histogram

Separation

1. Estimate EFMS $\hat{E}$ of mixture for all time-frequency bins
2. For all time-frequency bins (k, m) of the mixture
   (a) Estimate EFMS $\hat{E}$.
   (b) Use p.d.f. estimate from the training to either signal class or reject using equations (3.27)-(3.29).
3. For each class $c \in \{1, 2\}$
   (a) Generate time-frequency binary mask using (3.30).
   (b) Estimate time-frequency of a single source using (3.31).
   (c) Obtain time domain estimate using (3.32).



(a) (b)

Figure 3.9: Spectrogram of synthetic signals used for testing: (a) strongly frequency modulated signal (b) weakly frequency modulated signal.

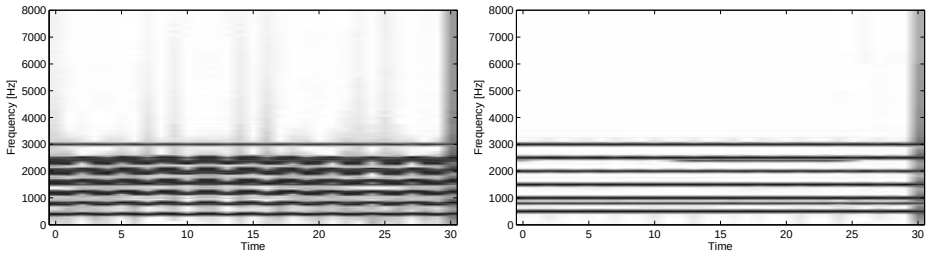
3.5.1 Synthetic signals

First we verify the ability of the proposed algorithm to segregate signals that differ in their subband frequency modulation signal energy. We choose synthetic signals which properties are similar to voiced phonemes and piano play (i.e. several frequency modulated partials). More precisely:

$$\begin{aligned} s_c(n) &= \sum_{l=0}^{N_h} \cos\left(l \cdot 2\pi f_c^{(c)} n/f_s + \sum_{m=0}^{n} q_l^{(c)}(m) \frac{1}{T} \right) \\ q_l^{(c)}(n) &= l \cdot d^{(c)} \cos\left(2\pi f_m^{(c)} n/f_s \right) \end{aligned}$$

where $c \in \{1,2\}$ is class index, N_h is the number of harmonic partials, f_c, f_s and f_m are carrier, sampling and modulation frequencies respectively and d is the modulating signal amplitude. We choose $N_h = 6$, $f_c^{(1)} = 400\,[\text{Hz}]$, $f_m^{(1)} = 10\,[\text{Hz}]$, $d^{(1)} = 20$, $f_c^{(2)} = 500\,[\text{Hz}]$, $f_m^{(2)} = 10\,[\text{Hz}]$, $d^{(2)} = 1$. Note that $d^{(1)} \gg d^{(2)}$ as assumed by our model for speech and piano. We normalize variance of s_c to 1. Fig. 3.9 shows spectrograms of synthetic signals used in this experiment and Fig. 3.10 shows source signal estimates recovered from the mixture. We can see that mixture components and their extracted counterparts looks very much the same.

We perform another experiment that shows that our algorithm is capable of separating white noise from weakly frequency-modulated signal. Signals chosen



(a) (b)

Figure 3.10: Spectrograms of (a) estimate of strongly modulated signal and (b) weakly frequency modulated signal. Both signals are recovered from 0 dB mixture.

	SDR_1	SIR_1	SAR_1	SNR_1	SDR_2	SIR_2	SAR_2	SNR_2
(a)	10.4	15.3	12.3	1.2	10.8	24.5	11.1	1.2
(b)	10.5	12.2	15.8	2.2	13.3	32.9	13.3	2.2

Table 3.1: Synthetic signals separation. (a) Two frequency modulated signals. (b) Noise and frequency modulated signal.

for this experiment have similar properties to fricative phonemes and piano play.

Table 3.1 shows separation performance results and Fig. 3.11 shows the spectrogram of source signals estimate recovered from the mixture. The spectrogram of noise signal is omitted. As in the previous case, the extracted harmonic component has only few artifacts that look like a "musical noise".

We conclude that our method is capable of segregating two audio signals having similar properties to real speech and piano play by visually examining the spectrograms of the signals estimates and noticing high values of objective measures. Nevertheless, looking at Fig. 3.10(b) we notice that some energy of the second partial (800 Hz) of s_1 "leaked" to $\hat{s}_2$. We attribute this leakage to simplified learning of EFMS distribution at different frequency bands. The amplitude of the frequency modulating signals of each partial grows linearly with the amplitude of the frequency modulating signal of the fundamental partial as explained in Section (3.3.3). Nevertheless we do not take this effect into consideration in our algorithm.

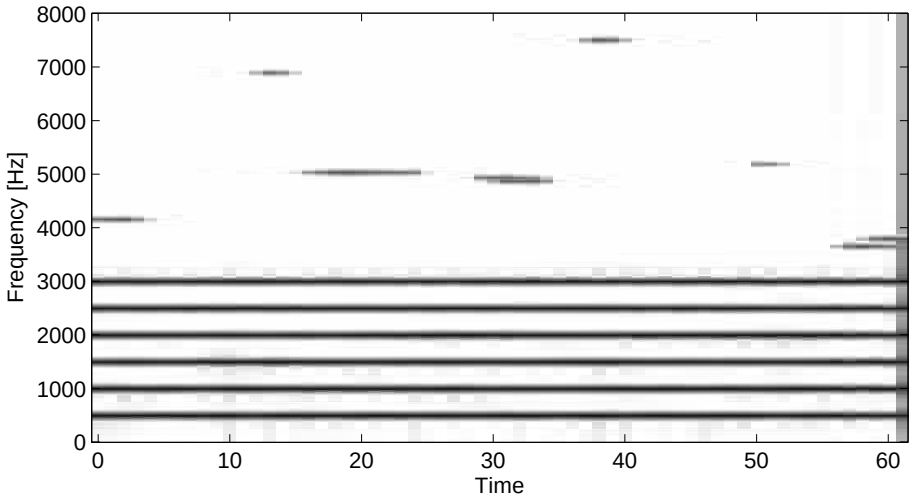


Figure 3.11: Spectrograms of reconstructed weakly frequency modulated signal.

3.5.2 Real signals

Now we describe simulation and informal listening test results of the proposed algorithm and compare its performance to a simple GMM monaural separation algorithm described in Section (2.3.3).

We use 60 seconds of speech (male only) taken from TIMIT database sampled at 16 KHz for GMM training. We use 1024 points STFT transform, Hamming synthesis window, 50% overlap and 12 components GMM.

The parameters used for the proposed algorithm were: $N = 1024$, $M = 64$, $N_u = 121$, $\delta_E = 15$dB, $\lambda_{12} = \lambda_{21} = 1$, $\lambda_r = \infty$, $\alpha = 1/3$. We used a Hamming synthesis window for the STFT transform. The high-pass filter used for the removal of Ω_c component is 122 taps FIR filter with stop angular frequency of 0.01π.

The WDO value (see Appendix (B)) for the pair of signals used in our experiment equals to 0.96 which according to [34] guaranties perceptually perfect

	SDR_1	SIR_1	SAR_1	LSD_1	SDR_2	SIR_2	SAR_2	LSD_2
Oracle mask	18.9	42.6	18.9	0.73	17.9	47.2	18.0	0.8
EFMS	6.1	11.8	7.8	2.4	6.4	19.7	6.6	2.0
GMM	2.4	9.3	3.8	2.9	2.6	7.9	4.8	2.5

Table 3.2: EFMS based separation algorithm performance

separation using the following binary mask in the STFT domain

$$\tilde{M}_k^{(1)}(m) = \begin{cases} 1 & \frac{|S_{1,k}(m)|}{|S_{2,k}(m)|} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.33}$$

$$\tilde{M}_k^{(2)}(m) = \begin{cases} 1 & \frac{|S_{1,k}(m)|}{|S_{2,k}(m)|} \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.34}$$

We will call masks defined in (3.34) "oracle" masks. The performance of the oracle mask in source separation induces an upper bound in the mean square sense on the performance of any separation algorithm that uses binary masking and same analysis filterbank. We present the results of source separation using these masks together with other separation results.

We evaluated the performance of algorithms using another 45 seconds of speech and piano signals. The results are shown in Table 3.2. The "oracle" masks obtain the highest performance according to all measures. However, this is only a theoretical results since creation of "oracle" mask requires a-priori knowledge of mixture components energy in every time-frequency bin. The EFMS based separation method shows better separation performance in all measured parameters compared to the GMM based separation algorithm.

Figures 3.12 and 3.13 show spectrograms of speech and piano play signals used in the mixture together with the signals recovered by the GMM based algorithm and by the proposed algorithm. Smaller amounts of interfering signals can be seen in speech and piano play for signals recovered by the proposed method compared to the GMM based algorithm.

Informal listening to signals separated by the proposed algorithm reveals

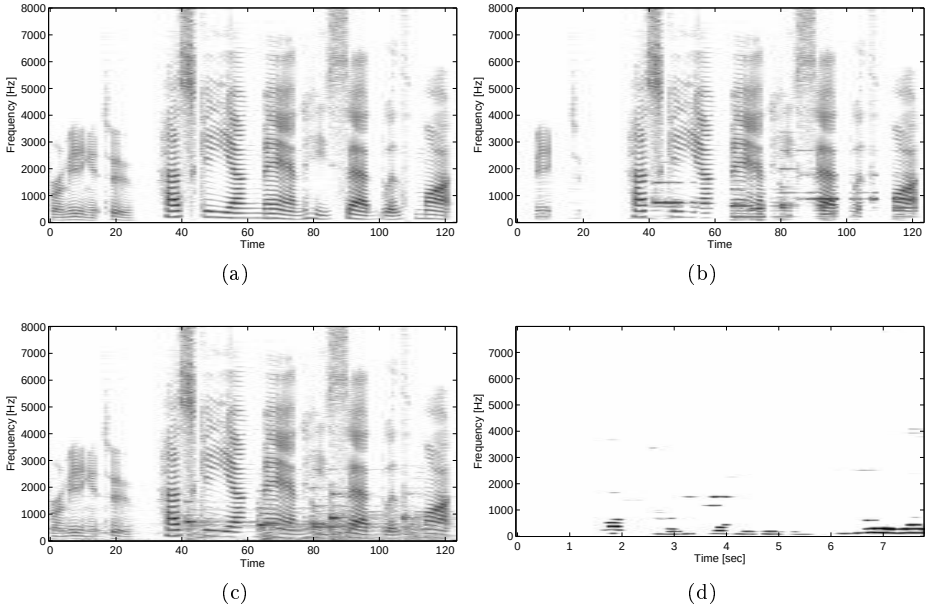


Figure 3.12: Spectrograms of the (a) clean , (b) GMM based algorithm recovered, (c) the proposed algorithm recovered speech signals and (d) residual speech signal after applying the algorithm to clean speech signal.

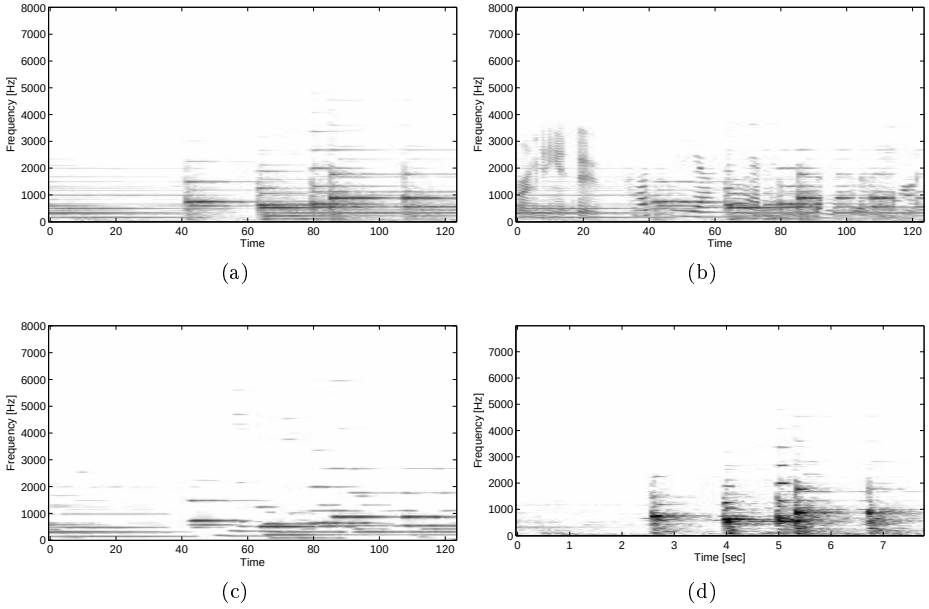


Figure 3.13: Spectrograms of the (a) clean, (b) GMM based algorithm recovered, (c) the proposed algorithm recovered piano signals and (d) residual piano play signal after applying the algorithm to clean piano signal.

that the proposed method produces much smaller amount of artifacts and more pleasant sound than the GMM based algorithm. The mixture separated using oracle mask has perceptually perfect separation.

The proposed method fails detecting onsets of piano notes. The reason is that piano strings are excited by a strike of a felt covered hammer. It results in a strong non-harmonic component near the note onset. Only harmonic components of piano play are detected by our algorithm and the rest of the signal leaks into estimated speech component.

In order to find out which part of speech signal leaks to the piano channel, we applied our algorithm to a clean speech signal instead of speech-piano mixture (i.e. $x(n) = s_1(n)$). Ideal separation algorithm would estimate $\hat{s}_2(n) = 0$. We examine the actual $\hat{s}_2$ signal. Fig. 3.12(d) shows the spectrogram of $\hat{s}_2$. The leaking speech parts are harmonic in their nature and have constant pitch over relatively long periods of time (0.5-1 sec). A certain amount of musical noise is also present. On the other hand, applying the algorithm to a clean piano play signal and examining $\hat{s}_1$, 3.13(d) reveals that most of the leaking signal is the hammer strikes as was observed earlier in the informal listening.

Finally we tuned λ_{12}, λ_{21}, λ_r parameters manually, to achieve the most perceptually plausible separation results. We chose $\lambda_{12} = 4, \lambda_{21} = 1, \lambda_r = 0.4$. The resulting objective measures $\mathrm{SDR}_1 = -0.1$, $\mathrm{SIR}_1 = 18.6$, $\mathrm{SAR}_1 = 0.0$, $\mathrm{SDR}_2 = 5.5$, $\mathrm{SIR}_2 = 21.5$, $\mathrm{SAR}_2 = 5.6$. Although some measures show deteriorated performance, the separated speech is intelligible and contains very low amount of audible interfering signal. Piano audible quality remains similar to the previous experiment with slightly higher rate of interfering signal rejection.

Audio files used for training and performance evaluation as well as separation results can be found at http://sipl.technion.ac.il/~elitvin/EFMS/.

3.6 Summary

We have presented and evaluated a novel technique for single-channel source separation based on the energy of subband frequency modulating signal. The subband frequency modulating signal is intimately connected to the subband instantaneous frequency and phase of the analyzed signal. Signals we worked with could be easily classified just by visually inspecting the subband phase behavior in a time-frequency localized region. This observation inspired us to define the EFMS analysis and the proposed separation algorithm.

An a-parametric probability distribution model of EFMS was learned during the training stage. Such simple modeling technique was found appropriate because we had to model a single-dimensional variable and the amount of training data, even from a several seconds of audio signal, is large. The classification of time-frequency bins was done using minimal Bayesian risk rule. We added a rejection option to this rule in order to improve the perceptual quality. The proposed method requires very simple and computationally efficient training compared to the GMM based algorithm. We confirmed that the proposed algorithm also produces superior results by objective performance measures and informal listening tests.

EFMS is not the only local time-frequency property that can be used for source separation. In Chapter 5 we experimentally study a possibility of using higher order statistics of spectrum magnitude for time-frequency bin classification.

Chapter 4

Bark-Scaled WPD

4.1 Introduction

Traditionally, short time Fourier transform (STFT) is used in many audio and speech processing applications. Bark-Scaled Wavelet Packet Decomposition (BS-WPD) [31] is a time-frequency signal transformation with non uniform frequency resolution. This transformation reflects the critical bands structure of the human auditory system. Mapping based complex wavelet transform (CWT) [58] is based on bijective mapping of a real signal into a complex signal domain followed by standard wavelet analysis performed on the complex signal. Among others, CWT partially mitigates lack of shift invariance of wavelet analysis. Benaroya et al. [19] proposed and analyzed blind source separation using time varying Wiener filter in the STFT domain. First the GMM models for two different signal sources are trained using training samples. Then, the separation is performed by maximizing maximum a posterior (MAP) criterion.

In this chapter we propose a source separation algorithm that follows Benaroya's STFT based algorithm, but operates on non uniform WPD filter-bank. We modify the BS-WPD analysis to equalize sampling rates of different scale-bands, which enables construction of instantaneous spectral shapes that are used in training and separation stages of the separation algorithm. We also use

CWT in order to achieve some level of shift invariance. The non-uniform frequency resolution of the BS-WPD filterbank, reduces the dimension of feature vectors by allocating fewer vector elements to higher frequencies. This behavior is similar to the critical bands structure of human auditory system. In a series of experiments we validate our approach using various types of wavelet families and show that the proposed approach is capable of performing the separation task.

The remainder of this chapter is structured as follows. In Section 4.2 we describe Bark Scaled WPD and the modification designed to equalize sampling frequencies in all sub-bands. Section 4.3 presents our mixing model and MAP estimators for its components. Section 4.4 describes training and separation stages of the algorithm. Section 4.5 presents our experimental results.

4.2 Bark-scaled wavelet packet decomposition

In this section we present the BS-WPD and introduce a modification that has some favorable properties for the frame-by-frame classification used in our algorithm.

Let $E \subset \left\{(l,n) : 0 \leq l < L, 0 \leq n < 2^l\right\}$ be a set of terminal nodes of a WPD tree. The center frequency of a terminal node $(l,n) \in E$ is roughly given by

$$f_{l,n} \quad = \quad 2^{-l}\left(\mathrm{GC}^{-1}(n) + 0.5\right) f_s$$

where GC^{-1} (n) is the inverse Gray code of n and f_s is a sampling frequency. Critical band WPD (CB-WPD)[31] filterbank structure is obtained by selecting a terminal nodes set E in a way that positions center frequencies $f_{l,n}$ approximately 1 Bark apart. Another constraint that must be taken into consideration is that a dyadic interval set $\left\{I_{l,n} : I_{l,n} = 2^{-l}n, 2^{-l}(n+1)\right\}$ must form a disjoint cover of $[0,1)$. Only in this case the set of wavelet packet family functions to be able to span the signal space.

BS-WPD is defined in [31] as CB-WPD with two additional levels of expan-

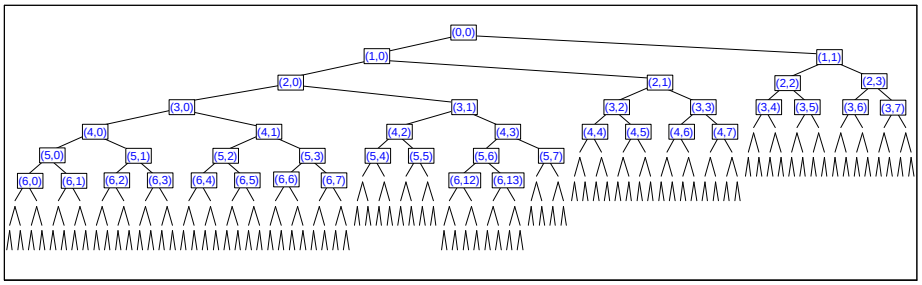


Figure 4.1: CSR-BS-WPD decomposition tree. Nodes having $l > 6$ are not decimated. This way, sampling frequencies of signals in all terminal nodes will be the same. Only few of the node labels are shown due to the space limitations.

sion. Due to higher frequency resolution BS-WPD performed better in the task of speech enhancement (in our experimental setting, adding three additional levels of decomposition proved even more beneficial).

The source separation algorithm used in this chapter requires every time instance to be described by a single feature vector holding the instantaneous spectral information from all sub-bands. We define a version of the BS-WPD transform that has equal sampling frequency in each sub-band. Unfortunately, terminal nodes of BS-WPD are located at various depths and each depth is associated with different sampling frequency. In order to align signals from all sub-bands and equalize the sampling frequency we do not decimate the lowpass and detail signal in nodes with $l > 6$. We call this transform Constant Sampling Rate BS-WPD (CSR-BS-WPD). We note that by canceling decimation in lower levels of the WPD tree we introduce a certain amount of redundancy into CSR-BS-WPD representation. Fig. 4.1 shows CSR-BS-WPD decomposition tree.

The CSR-BS-WPD analysis produces only 168 sub-bands, compared to 513 sub-bands of STFT analysis with approximately the same frequency resolution in low frequencies. We sacrifice frequency resolution at higher frequency range, in accordance with human auditory system which also has a coarser resolution in high frequency range. Reducing the number of sub-bands results in smaller dimension of data used in training and separation stages. Smaller data dimension has a potential to increase accuracy of the GMM estimation because of the

reduced redundancy in feature vectors and to reduce computational burden.

Let $x(n)$ be a time sequence and

$$x^{+}(n) = h^{+}(n) * x(n) \tag{4.1}$$

where h^{+} is a digital mapping filter with an approximately zero magnitude response for negative frequencies and approximately unit response for positive frequencies. Suppression of signal negative frequencies (similar to Hilbert filter) maps an input signal x into a so called Softy-space (for details see Section A.3). The image of x in Softy-space is denoted by x^{+}. We denote the CSR-BS-WPD transform of $x^{+}(n)$ as $X_{l,n}(m)$ where (l,n) are indices of terminal nodes and m is time index. Since all terminal nodes have the same sampling rate we can rearrange the elements of $X_{l,n}(m)$ into a single column complex vector $\bar{X}(m) \in \mathbb{C}^{M}$. The dimension of $\bar{X}(m)$ is given by the number of sub-bands $M = 168$.

The CSR-BS-WPD is a reversible transform. Given CSR-BS-WPD signal $\bar{X}(m)$ we can acquire a time domain signal $x(n)$ first by inverting the WPD and then taken the real part of $x^{+}(n)$ (or filtering $x^{+}(n)$ with a digital filter g^{+} as defined in [32]).

4.3 Mixture components estimation

We assume mixing model as in (2.1). The Softy-space mapping (4.1) is linear:

$$x^{+}(n) = s_1^{+}(n) + s_2^{+}(n)$$

The CSR-BS-WPD expansion is also linear because WPD is linear:

$$\bar{X}(m) = \bar{S}_1(m) + \bar{S}_2(m)$$

$$\bar{S}_1(m), \bar{S}_2(m), \bar{X}(m) \quad \in \quad \mathbb{C}^M$$

If we assume Gaussian model on s_1, s_2 then due to the linearity of Soft-space mapping and the CSR-BS-WPD transform $\bar{X}, \bar{S}_1, \bar{S}_2$ will have complex Gaussian distribution and the reasoning in section (2.3) holds.

In order for equation (2.4) to hold, the covariance matrices of the random vectors ought to be diagonal. In section 4.3 we justified assuming stationarity and approximate circularity of the vectors. In order to justify (2.4) in this case we need to assume that the covariance matrices of $\bar{S}_1, \bar{S}_2$ are diagonal. The diagonality of a covariance matrix means lack of correlation between samples in different frequency bands. We can interpret CSR-BS-WPD transform as a filterbank, and we can expect that as long as WPD filters have good frequency localization (i.e. energy leaking to neighboring bands is low) , the correlation between different frequency bands will be also low and the diagonal covariance matrix assumption can be extended to the CSR-BS-WPD signals.

Like we mentioned in section 4.3 simple assumption of Gaussian distribution prior does not hold for most real signals so here as well we assume GMM model. Under the assumption of diagonal covariance matrices for $\bar{S}_1, \bar{S}_2$ the estimator (2.5) takes the following form:

$$\hat{\bar{S}}_1(m) \quad = \quad \sum_{j,k=1}^{K} \gamma_{j,k}(m) \frac{\left(\bar{\sigma}_1^{(j)}\right)^2}{\left(\bar{\sigma}_1^{(j)}\right)^2 + \left(\bar{\sigma}_2^{(j)}\right)^2} \bar{X}(m) \tag{4.2}$$

and (2.6) takes the form:

$$\hat{\bar{S}}_1(m) \quad = \quad \gamma^*_{j,k}(m) \frac{\left(\bar{\sigma}_1^{(j)}\right)^2}{\left(\bar{\sigma}_1^{(j)}\right)^2 + \left(\bar{\sigma}_2^{(j)}\right)^2} \bar{X}(m) \tag{4.3}$$

where $(\bar{\sigma}_c)^2$ are on diagonal elements of Σ_c and γ^* defined like in (2.6).

4.4 Training and separation

Let L be a number of training signal time samples in CSR-BS-WPD domain. During the training stage we use signal samples of both classes $\left\{\bar{S}_1(m)\right\}_{m=1}^{L}, \left\{\bar{S}_2(m)\right\}_{m=1}^{L}$ to train two different GMM models. Both Softy-space mapping and the WPD are linear transformations, we conclude that expectation values of s, s^+ and $\bar{S}$ are zero, hence we can define a simplified zero mean GMM model that

$$\Lambda_c = \left(w_c, \left\{\Sigma_c^{(k)}\right\}_{k=1}^{K}\right), w_c \in \mathbb{R}^K, \Sigma_c \in \mathbb{R}^{M\times M} \tag{4.4}$$

where K is the GMM model order. Following the reasoning in the previous section, we assume $\Sigma_c^{(k)}$ to be a diagonal covariance matrix.

We note that often, a signal in each frequency bin is assumed to have a complex Gaussian distribution [23]. This assumption is not very accurate for voiced parts of speech which is better modeled by a complex exponent with linear phase and random initial phase. Usually, GMM model is trained on spectral magnitude signal and expected value and variance of the signal are being estimated during the training process. The estimated expected values represent spectral shapes of training signal. In this work we train GMM model using complex spectrum values. Expected value of complex signal is zero. We estimate coefficients on the covariance matrix diagonal. The output of the training process is a group of spectral shapes (one for each GMM component) described by the magnitude of values on the covariance matrix diagonal.

The training of the GMM models is performed using Expectation Maximization (EM) algorithm [35] and bootstrapped using K-Means algorithm. Expectation value of training data is assumed to be zero, it is is not updated during the expectation step of the EM algorithm and constantly set to zero.

We note that the estimation of $\hat{\bar{S}}_c(m)$ is performed for every time index m. In the rest of this section we omit time index m for the clearness of notation. In order to estimate signal sources $\hat{\bar{S}}_c$ using (2.5) for every time instance, we first estimate posterior probability $\gamma_{j,k}$:

$$\begin{aligned} \gamma_{j,k} &\propto p\left(\bar{X}|q_1=j,q_2=k\right)p\left(q_1=j\right)p\left(q_2=k\right) \\ &= g\left(\bar{X};\Sigma_1^{(j)}+\Sigma_2^{(k)}\right)w_1^{(j)}w_2^{(k)} \end{aligned} \tag{4.5}$$

where $g\left(\bar{X};\Sigma\right)$ is a zero mean multivariate Gaussian probability density function. Substituting (4.5) into (4.2) and using Λ_c estimated in the training process we acquire estimators for $\hat{\bar{S}}_1$ and $\hat{\bar{S}}_2$. The entire separation algorithm algorithm is described in Algorithm 4.1.

The summary of the training and separation stages can be seen in Algorithm 4.1. The training stage is performed offline, given a database of signals from two different classes. For each signal class, a time-frequency representation of signals is obtained using the CSR-BS-WPD transform 4.2. The GMM model of spectral magnitude vectors for each time frame is trained using EM algorithm [35]. Unlike classic EM implementation for GMM we do not learn mean value of the distribution since it is assumed to be zero.

In the separation stage, the CSR-BS-WPD transform is applied to the mixture. Then, the value of γ is calculated for all active GMM component combinations and estimates of both the sources are acquired in the time-frequency domain using PM estimator (4.2) or MAP estimator (4.3). The separated component signal estimators are recovered using inverse CSR-BS-WPD.

4.5 Experimental results

We compare the performance of our algorithm to the STFT based algorithm [2]. We demonstrate the effectiveness of the proposed algorithm on synthetic signals and on speech and piano music mixtures.

Algorithm 4.1 CSR-BS-WPD and GMM based source separation algorithm

Training

1. Compute CSR-BS-WPD time frequency representation $\bar{S}_1(m), \bar{S}_2(m)$ of training signals $s_1(n), s_2(n)$ as explained in Section 4.2.
2. Train Λ_1, Λ_2 GMM models using data vectors $\bar{S}_1(m), \bar{S}_2(m)$ and EM algorithm.

Separation

1. Compute CSR-BS-WPD time frequency representation $\bar{X}(m)$ of mixed signal $x(n)$ as explained in Section 4.2.
2. For all time indexes m
   (a) For all pairs $(j,k) \in \{(j,k) \,|\, j \in \{1,\ldots,K\}, k \in \{1,\ldots,K\}\}$
      i. Compute $\gamma_{j,k}(m)$ using (4.5)
   (b) Estimate $\hat{\bar{S}}_1, \hat{\bar{S}}_2$ using (4.2) or (4.3)
3. Compute estimates of mixture components in time domain using as explained in Section 4.2.

4.5.1 Synthetic signals

First we evaluated the performance of the separation algorithm using synthetic signals. Our goal is to verify the feasibility of the separation in the CSR-BS-WPD domain. We constructed synthetic signals by dividing the entire signal time span into 400ms segments and each segment is generated by

$$x_c(n) = \begin{cases} \sum_{i=1}^{2} \cos\left(\frac{2\pi}{f_s} f_{1,i}^{(c)} n\right) & \text{with probability } \frac{1}{2} \\ \sum_{i=1}^{2} \cos\left(\frac{2\pi}{f_s} f_{2,i}^{(c)} n\right) & \text{with probability } \frac{1}{2} \end{cases} \tag{4.6}$$

where $f_{1,1}^{(1)} = 220$Hz, $f_{1,2}^{(1)} = 440$Hz, $f_{1,1}^{(2)} = 300$Hz, $f_{1,2}^{(2)} = 600$Hz. Two different signals were generated for training and evaluation. We used GMM with two mixture components.

Table 4.1 shows the separation performance. High values of SIR indicate high rejection of interfering signal and relatively high values of SDR indicate low amount of distortion introduced by our separation algorithm. Both algorithms

	SDR_1	SIR_1	SAR_1	SDR_2	SIR_2	SAR_2
STFT	16	40	16	16	31	16
CSR-BS-WPD	20	35	20	22	40	22

Table 4.1: Separation performance measures for separation of synthetically generated signals for different algorithms. The measures are shown in *dB*.

perform very well on these synthetic signals.

4.5.2 Real signals

We also evaluated the performance of the proposed algorithm on real signals. We separated two audio source classes: speech and piano excerpts. We used training sequence of 50 seconds for model training and 11 seconds for the performance evaluation. Different speech and piano excerpts were used for training and performance evaluation. Speech signals were taken from TIMIT database and included male speakers only. Piano excerpts were taken from Chopin's preludes. All signals were sampled at 16 KHz. We equalized the energy of all signals before training and before the signal were mixed.

We compared the separation performance for different GMM model order and different wavelet families. Fig. 4.2 depicts signal to distortion ratio of speech signal (SDR_1). A higher value of SDR_1 indicates a smaller degree of speech distortion after the separation. For low orders of GMM, the CSR-BS-WPD analysis based on the discrete Meyer (*dmey*) wavelet family outperforms other tested wavelet families and STFT based algorithm. For high orders of GMM, *dmey* based algorithm shows performance comparable to the STFT based algorithm and slightly better performance than other wavelet families. Although Fig. 4.2 depicts only the SDR_1 measure, other performance measures ($SDR_i, SIR_i, SAR_i, LSD_i$) showed similar behavior. Fig. \ref{fig:all-families-compare} shows the performance of different wavelet families and the STFT based algorithm for GMM model order of 10. The *dmey* wavelet family based algorithm shows performance superior to STFT for all objective measures com-

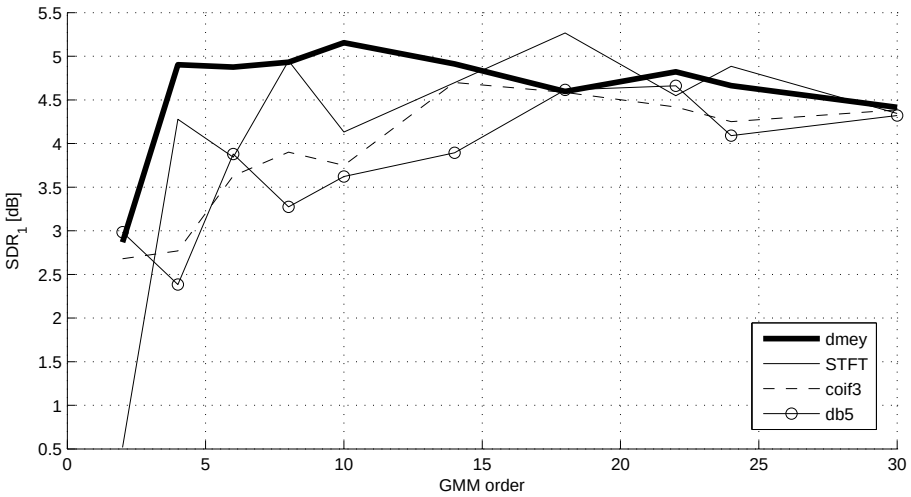


Figure 4.2: Influence of the GMM model order on the signal to distortion ratio (SDR_1) of the speech signal. STFT based algorithm [2] is compared to CSR-BS-WPD based algorithm.

pared. Other wavelet families, however show inferior separation performance to the STFT based algorithm.

In additional experiments we noticed that when *dmey* based analysis is used, the sparseness of music and speech signals in the CSR-BS-WPD domain is highest compared to other wavelet families used in separation experiments. In [31], *dmey* wavelet family is also used for speech enhancement and motivated by the regularity of the wavelet filter and its good frequency localization properties.

Informal listening tests indicate that the CSR-BS-WPD based separation produces less "jumpy" and more pleasant signal reconstruction than the STFT based version of the algorithm.

4.6 Summary

We have described how a Bark-scaled WPD can be adapted to the source separation task. Introduction of a different subsampling scheme and approximate shift invariance in the BS-WPD decomposition tree, enabled us to adapt a source

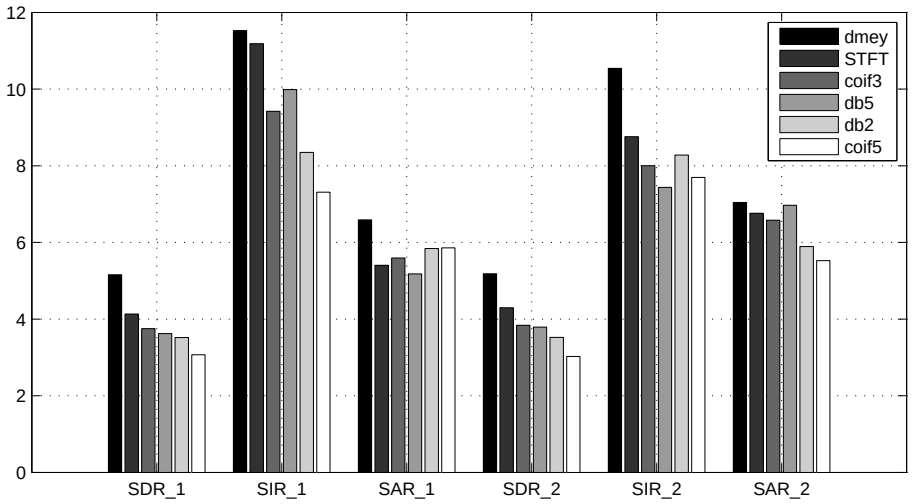


Figure 4.3: Performance comparison of all wavelet families and the STFT based algorithm for GMM order of 10.

separation algorithm that was originally designed to work in the uniformly sampled STFT domain. We found several advantages of using the proposed analysis together with a GMM based source separation algorithm. Coarse frequency resolution in high frequency range allowed us to reduce the computational burden. The perceptual quality of the extracted signal components was also improved. We found out that the Discrete Meyer wavelet based CSR-BS-WPD analysis yields improved separation performance compared to STFT analysis and other wavelet families tested. As a byproduct of our work we found out that the sparseness of signal representation is highest when the Discrete Mayer wavelet filters (compared to several other tested wavelet families) are used and we believe it is the reason for the superior performance of the separation algorithm based on this wavelet family.

Chapter 5

Short Time Spectral Kurtosis

5.1 Introduction

High order statistics is frequently used for source separation in multichannel case. In particular, kurtosis used as a measure of non-Gaussianity of the recovered mixture components. Spectral Kurtosis (SK) is a tool capable locating non-Gaussian components including their location in the frequency domain. SK was first introduced by Dwyer [59]. He defined it as a kurtosis value of real part of the STFT filterbank output. Antoni [60] introduced a different formalization of SK by means of Wold-Cramér decomposition which gave a theoretical ground for the estimation of SK of non-stationary processes. He also showed practical applications of his approach in the field of machine surveillance and diagnostics [61, 62]. Another applications of spectral kurtosis include SNR estimation in speech signals [63], denoising [64] and subterranean termite detection [65].

In Chapter 3 we used properties of subband FM signal in order to label each time-frequency bin assignment to different audio sources. We relied on the W-DO property to justify binary masking in the STFT domain as a valid method for source separation (see Section B). In this chapter we use the time localized value of the kurtosis in different subbands (short time spectral kurtosis) to label time-frequency bins and create a binary mask.

The remainder of this chapter is structured as follows. In Section 5.2 we present the concept of spectral kurtosis. Then, Section 5.3 extends the idea of spectral kurtosis to to non-stationary signals. In Section 5.4 we describe simple source separation algorithm based on the spectral kurtosis analysis and experimental study is described in Section 5.5. Section 5.6 concludes this chapter.

5.2 Spectral kurtosis

Let $x(n)$ be a real, discrete time, random vector. Let X_k be its N points Discrete Fourier Transform (DFT) defined as

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}$$

Following [66], the spectral kurtosis is defined as

$$\mathcal{K}_x(k) = \frac{\kappa\{X_k, X_k^*, X_k, X_k^*\}}{(\kappa\{X_k, X_k^*\})^2} \tag{5.1}$$

Using the circularity the definition can be rewritten as

$$\mathcal{K}_x(k) = \frac{\mathbb{E}\left\{|X_k|^4\right\}}{\left(\mathbb{E}\left\{|X_k|^2\right\}\right)^2} - 2 \tag{5.2}$$

Gaussian processes

Let $x_{\mathrm{WG}}(n)$ be white Gaussian vector. Its DFT is a complex Gaussian. Cumulants with order greater than 3 is zero for Gaussian and complex Gaussian random variable. By (5.1) the SK of $x_{\mathrm{WG}}(n)$ is zero for all m.

Complex sine with random phase

Let $x_{\mathrm{sine}}(n) = ae^{j\left(2\pi\frac{m_0}{N}n+\varphi\right)}$. When $\varphi \sim U(0, 2\pi)$, $x_{\mathrm{sine}}(n)$ is a stationary process. By (5.2) and noticing that $\mathbb{E}\left\{|X_k|^4\right\} = \left(\mathbb{E}\left\{|X_k|^2\right\}\right)^2 = (Na)^4$ we

conclude that $\mathcal{K}_{x_{\text{sine}}}(k) = -1$.

5.2.1 Kurtosis of signal mixture

Assume mixing model (2.1) and that $s_1(n)$ and $s_2(n)$ are statistically independent stationary processes. Let $\phi_A(k) \triangleq \mathbb{E}\left(|A_k|^2\right)$ and $\gamma \triangleq \phi_{s_1}(k)/\phi_{s_2}(k)$. It is shown in [67] that

$$\begin{aligned}\mathcal{K}_x(k) &= \left|\frac{\phi_{s_1}(k)}{\phi_{s_1}(k)+\phi_{s_2}(k)}\right|^2 \mathcal{K}_{s_1}(k) + \left|\frac{\phi_{s_2}(k)}{\phi_{s_1}(k)+\phi_{s_2}(k)}\right|^2 \mathcal{K}_{s_2}(k) \\ &= \left|\frac{1}{1+1/\gamma}\right|^2 \mathcal{K}_{s_1}(k) + \left|\frac{1}{1+\gamma}\right|^2 \mathcal{K}_{s_2}(k) \qquad (5.3)\end{aligned}$$

When $\gamma \gg 1$, $\mathcal{K}_x(k) \approx \mathcal{K}_{s_1}(k)$. Similarly, when $\gamma \ll 1$, $\mathcal{K}_x(k) \approx \mathcal{K}_{s_2}(k)$.

If two signals that we try to separate have W-DO property, then for almost every time-frequency bin either $\gamma \gg 1$ or $\gamma \ll 1$. According to (5.3) the source of the signal in each time-frequency bin can be determined.

5.2.2 Kurtosis estimation

Let $\{X(i)\}_{i=1}^{L_K} \in \mathbb{R}^N$ a set of samples. If $\{X(i)\}$ are i.i.d, Vrabie et al. propose the following unbiased estimator of the SK [66]

$$\hat{\mathcal{K}}_X = \frac{L_K}{L_K - 1}\left(\frac{(L_K+1)\sum_{i=1}^{L_K}|X(i)|^4}{\left(\sum_{i=1}^{L_K}|X(i)|^2\right)^2} - 2\right)$$

Antoni [60] proposed another estimator for the SK assuming Wold-Cramér decomposition of non-stationary process. It is based on the STFT transform and requires analyzed signal to be quasi-stationary at the scale of STFT analysis windows. This requirement is common requirement in audio signal processing. The estimator is defined using $\langle . \rangle_t$ time averaging operator (time averaging is

done with the respect to t)

$$\hat{S}_{2nX}(k) \triangleq \left\langle |X_k(m)|^{2n} \right\rangle_m \tag{5.4}$$

where $X_k(m)$ is defined in (3.11). The STFT based estimator of SK is defined as:

$$\hat{\mathcal{K}}_X(k) \triangleq \frac{\hat{S}_{4X}(k)}{\hat{S}_{2X}^2(k)} - 2 \tag{5.5}$$

The analysis of the statistical properties of this estimator can be found in [60].

The kurtosis estimator (5.5) estimates frequency localized kurtosis values. For the purpose of source separation we also need to localize SK estimation in time. In order to do so, we define a time localized $2n$-th order empirical spectral moment of $|X_k(m)|$ as

$$\hat{S}_{2nX,k}(m) \triangleq \sum_{i=-\lfloor L_K/2 \rfloor}^{\lfloor L_K/2 \rfloor} w_{\mathcal{K}}(m+i) |X_k(i)|^{2n} \tag{5.6}$$

where $w_{\mathcal{K}}(m)$ is an averaging window with $\sum_m w_{\mathcal{K}}(m) = 1$. Equations (5.4) and (5.6) are similar except (5.6) is also localized in time. Finally we define

$$\hat{\mathcal{K}}_{X,k}(m) \triangleq \frac{\hat{S}_{4X,k}(m)}{\hat{S}_{2X,k}^2(m)} - 2 \tag{5.7}$$

In the rest of this paper we refer to (5.7) as Short Time Spectrum Kurtosis (STSK).

5.2.3 Physical interpretation

Equation (5.7) can be written as follows:

$$\hat{\mathcal{K}}_X(k) \triangleq \frac{\left\langle |X_k(m)|^4 \right\rangle_m - \left\langle |X_k(m)|^2 \right\rangle_m^2}{\left\langle |X_k(m)|^2 \right\rangle_m^2} - 1$$

The expression can be interpreted as normalized empirical variance of signal energy in different bandpass channels (up to a subtracted constant -1) which is as a sort of measure for time dispersion of $|X_k|^2$ [60]. Similar interpretation can be applied to STSK.

5.3 Short time spectral kurtosis of real audio signals

In following examples we use audio signals sampled at 16 KHz. The STFT analysis is performed using $N = 1024$, $M = 128$. The spectral moments estimator average over time window of $L_K = 31$ and $w_{\mathcal{K}}$ is a square window:

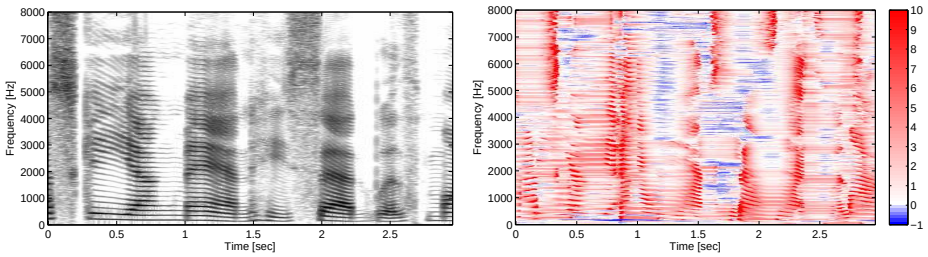
$$w_{\mathcal{K}}(m) = \begin{cases} 1/L_K & -\lfloor L_K/2 \rfloor \leq m \leq \lfloor L_K/2 \rfloor \\ 0 & \text{otherwise} \end{cases}$$

.

Figures 5.1,5.2,5.3 and 5.4 show spectrograms and STSK of speech, piano play, fast piano play and speech piano mixture signals respectively. High values of STSK can be observed in time-frequency regions where the power spectrum has non-stationary behavior in time.

Piano play signal mostly composed of harmonic partials that are well localized in frequency and maintain same statistical properties on a relatively long segments in time. As such, they can be well described by a complex sinusoidal model with a random phase. As we saw in Section 5.2, this model induces kurtosis value of -1.

Speech signal can be roughly divided into three categories: voiced phonemes,



(a) Spectrogram of speech signal (b) STSK of speech signal

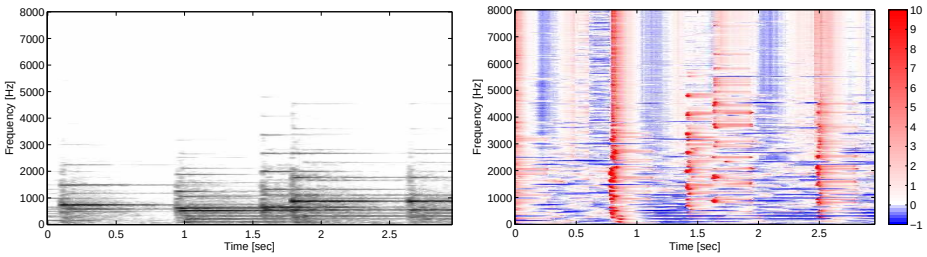
Figure 5.1: Power spectrum and STSK analysis of speech

plosive and fricative phones. Fricative phonemes are well described by a colored Gaussian noise model. A complex time signal in each frequency band of the STFT has complex Gaussian distribution. As we saw in Section 5.2, this model induces kurtosis value of 0. Plosive phonemes produce an energy peak in time that resembles a sample from a heavy tail distribution, as such, results in high values of kurtosis.

Voiced phonemes, like piano play, are harmonic signals. Unlike, piano play, their fundamental frequency changes continuously and rapidly. Harmonic partials rapidly enter to and exit from different frequency bands. This produces a sample set with a mix of high and low energy samples. It can be seen as samples from a heavy tail distribution. Examination of STSK (Fig. 5.1) indeed shows that measured values of STSK for speech are high.

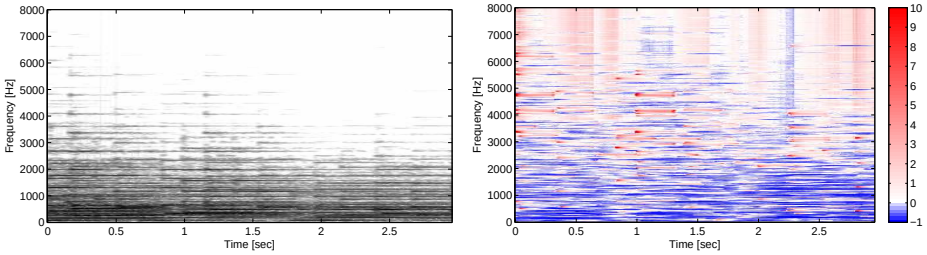
The STSK of piano play signals is much lower then STSK of speech (Fig. 5.2,5.3). The piano play signal is composed mostly of harmonic sounds having a constant fundamental frequency over a relatively long time periods. The onsets of notes show themselves as a spikes of STSK values.

The STSK of speech and piano mixture (Fig. 5.4) has low values of STSK in time-frequency regions that originate from the piano play and high values of STSK in the regions originating from speech. This observation implies that it should be possible to separate speech from piano play by binary masking the interfering signal using STSK for binary time-frequency bin classification.



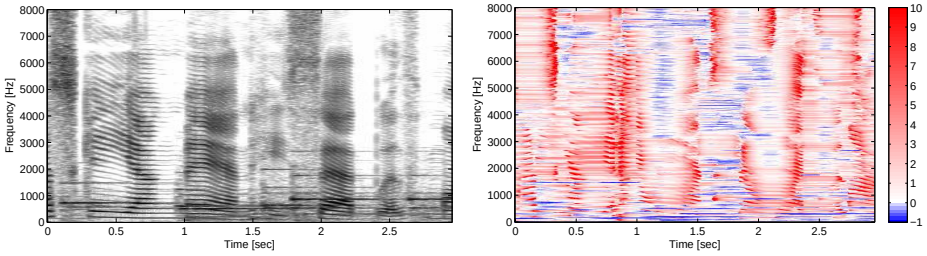
(a) Spectrogram of piano play (slow) (b) STSK of piano play (slow)

Figure 5.2: Power spectrum and STSK analysis of slow piano play



(a) Spectrogram of piano play (fast) (b) STSK of piano play (fast)

Figure 5.3: Power spectrum and STSK analysis of fast piano play



(a) Spectrogram of speech and piano mixture (b) STSK of speech and piano mixture

Figure 5.4: Power spectrum and STSK analysis of speech and piano play mixture

5.4 Source separation using STSK

In the previous section we observed that the STSK values of speech signal are generally higher than the STSK values of piano play. We also saw that pitch tracks of the piano play has low values of SK. Using this intuition we define the following binary masks in the STFT domain.

$$M_k^{(1)}(m) = \begin{cases} 1 & \hat{\mathcal{K}}_x(m,k) > \delta_{\mathrm{SK}} \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

$$M_k^{(2)}(m) = 1 - M_{1,k}(m) \tag{5.9}$$

where δ_{SK} is a threshold chosen based on experiments.

We reconstruct mixture components by masking the interfering signal time-frequency bins and performing inverse short time Fourier transform A.4

$$\hat{s}_c(n) = \mathrm{ISTFT}\left(M^{(c)} \circ X\right) \tag{5.10}$$

where $\circ$ denotes elementwise multiplication.

Algorithm 5.1 presents all steps of the STSK based source separation algorithm. First, the STSK of the mixed signal is calculated. Then STSK is estimated using (5.7). A simple thresholding (5.8), (5.9) is used to create binary masks. Threshold value δ_{SK} established from experiments. Finally, binary masks are applied to the time-frequency mixture signal obtained in step 1 of the algorithm and the separated components signal is recovered using the ISTFT transform.

5.5 Experimental results

We used same test samples as in Section 3.5. The STFT analysis parameters used for the SK based algorithm are shown in Table 5.1. The value of δ_{SK} was

Algorithm 5.1 STSK based source separation algorithm

1. Find the time-frequency representation $X_k(m)$ of the mixture using STFT A.3.
2. Estimate STSK $\hat{\mathcal{K}}_X(m,k)$ of the mixture using equation (5.7)
3. Create binary time-frequency masks $M_k(m)$ using equations (5.8), (5.9)
4. Filter out the interfering signal and get a time domain estimate using equation (5.10)

Sampling frequency	16KHz
STFT analysis window length	1024
STFT overlap (samples)	128
Short time SK estimation window length (samples/secs)	71 samples, (~0.5 sec)
δ_{SK}	1

Table 5.1: STFT analysis parameters

chosen experimentally.

Table 5.2 compares the performance of the proposed algorithm comparing to a GMM based separation algorithm (see Section (2.3.3)). We show the performance of the EFMS based separation algorithm explained in Chapter 3. All performance measures show superior performance of the STSK based algorithm over EFMS and GMM algorithm. The subjective tests indicated superior quality of the STSK over GMM algorithm and comparable or slightly better performance of the STSK over EFMS algorithm.

	SDR_1	SIR_1	SAR_1	LSD_1	SDR_2	SIR_2	SAR_2	LSD_2
Oracle mask	18.9	42.6	18.9	0.73	17.9	47.2	18.0	0.8
GMM	2.4	9.3	3.8	2.9	2.6	7.9	4.8	2.5
STSK	**7.7**	**19.5**	**8.0**	**2.8**	**8.2**	**21.3**	**8.4**	**2.3**
EFMS	6.1	11.8	7.8	2.4	6.4	19.7	6.6	2.0

Table 5.2: EFMS based separation algorithm performance. All performance measures show superior performance of the STSK based separation algorithm.

5.6 Summary

We have defined a concept of short time spectral kurtosis (STSK). Recent work on the Spectral Kurtosis focused mainly on the machine surveillance and diagnostics. In these publications estimation of the SK was done using large sets of independent samples. In our work we were interested in the value of SK localized in time.

The proposed estimation of the STSK is done on a much shorter time periods when only several or at best tens samples are available. Windowing and overlaps that are part of the STFT analysis, introduce inter-sample correlation and result in estimation bias [60]. We defined an ad-hoc estimator of the STSK but we did not study its statistical properties. Graphical plots of the STSK estimated by the proposed estimator produced meaningful images: the harmonic tracks are shown to have low values of the STSK while fricatives and noise had higher values of STSK. We note that in the proposed algorithm, the bias of the STSK estimator has little importance since only relative values of the STSK affect the algorithm.

We presented a very simple source separation algorithm that uses STSK values to classify time-frequency bins into one of two classes. We assume that piano play has more harmonic components and hence lower values of the STSK than in speech. The classification between two classes is done by simply comparing the STSK to a threshold and the separated sources are extracted by simple binary masking of the time-frequency bins followed by the ISTFT. Good experimental results suggest that the STSK information provides meaningful information about audio signals and may be a useful tool for audio signal processing applications.

Chapter 6

Conclusion

6.1 Summary

In this thesis we addressed a problem of blind and semi-blind source separation from a single sensor. We presented three novel algorithms for source separation.

The first two algorithms are the EFMS and the STSK based separation algorithms. The separation is based on individual time-frequency bin classification. Time-frequency mask is created for every mixture component and time-frequency bins that belong to the interfering class are zeroed. Extracted components are recovered using inverse time-frequency transform.

Similar single-channel source separation approaches found in the literature rely on global time-frequency information such as CASA cues or spectral shapes learned in the training stage. In our method we use only local information in order to assign each time-frequency bin to one of the sources. We use information from a single frequency band and time vicinity of few hundreds of millisecond to make an assignment decision.

We do not deal with complicated spectral statistical models and use simple a-parametric p.d.f. estimation using a normalized histogram, hence the training procedure is significantly simplified. Our method is also insensitive to spectral shape variations between different source instances.

The downside of using proposed time-frequency localized features is the ability to differentiate between a limited set of audio signals. For example, two harmonic musical instruments might have different timbre, but same amount of frequency modulation. These signals are indistinguishable in the local view of the time-frequency domain.

For a successful separation It is important that two signals do not overlap in the STFT domain. Previously studied W-DO requirement is not strong enough for the proposed method since it is necessary that the entire time vicinity used to estimate EFMS or STSK would contain mostly a single signal energy.

It is interesting to note that both the EFMS and the STSK algorithms produce similar separation results, both in the sense of objective measures, similar perceptual quality and similar artifacts. The former algorithm relies mostly on the phase information and the later on the fourth order statistics of the amplitude. For the harmonic parts of the signal, high frequency modulation implies that the carrier constantly exists and enters a subband. This directly affects the subband amplitude and is reflected through the STSK values.

The third algorithm is the CSR-BS-WPD based separation algorithm. It is closely related to a GMM based source separation presented earlier in the literature except for a psychoacousticaly motivated signal analysis frontend. It reduces dimension of the signal space and computational complexity of similar STFT filterbank based algorithm. Due to the similarity of the overall separation algorithm to the STFT filterbank based algorithm, the performance and pitfalls of both algorithms are similar. Only in some scenarios, when GMM order was chosen to be low, the separation performance was improved.

An important observation that was made during the performed experiments is that the discrete Meyer wavelet family showed superior performance compared to other wavelet families. Besides, we also found correlation between the performance of the separation algorithm to the sparsity of the representation coefficients which depends strongly on the wavelet family used.

6.2 Future research

Several source separation approaches and related techniques studied in this thesis open a number of interesting topics for future study:

1. Use CSR-BS-WPD analysis instead of STFT analysis in the EFMS and the STSK based separation algorithms presented in this work. The psychoacoustically motivated CSR-BS-WPD filterbank (opposed to the STFT filterbank used in our work) may improve the robustness of time-frequency bin assignment done by EFMS and STSK algorithm. It could also address the difficulty of selecting correct subband width for low and high frequencies discussed in subsection 3.3.3.

2. In the proposed EFMS based separation algorithm, we use a low-pass filter to reduce the variance of the EFMS estimation. This results in smoothed boundaries of signal onsets and offsets and hence deteriorates separation performance. Some sort of non-linear filtering, such as bilateral filtering, frequently used in image processing, can be applied to reduce variance of EFMS estimation without blurring onset and offset regions.

3. The EFMS based algorithm uses the energy of the FM signal which gives a crude description of the signal. Other FM signal analysis may address spectral structure of frequency modulating signal or time varying statistics and extend the range of possible applications of the subband frequency modulating signal based analysis as well as improve the performance of source separation.

4. We used a binary mask to reject interfering signal from the mixture. It may be possible to use some sort of a soft mask, (having real values in the 0 to 1 range instead of a discrete set of 0 and 1). The value of the mask may be determined using some probabilistic function. Use of soft mask has a potential to improve perceptual quality of separated signals.

5. The EFMS and the STSK based algorithm use local time-frequency infor-

mation for the source assignment of time-frequency bins. The information from other spectral regions, such as spectral shape, is discarded. The localized time-frequency approach may be combined with another source separation algorithm, such as GMM based algorithm used in this work to account for currently unused information.

6. It may be interesting to investigate applicability of subband frequency modulating signal analysis to other signal processing applications, such as speech enhancement or various audio classification tasks. For example, some preliminary experiments not reported in this work showed that EFMS can be used very efficiently as signal feature for different signal classes.

7. Little can be found in the literature on the topic of spectral kurtosis estimation. In our thesis we define and perform the STSK estimation in an ad-hoc manner. A rigorous approach to the STSK estimation such as rigorous definition estimators and study of their statistical properties may improve the performance of the proposed algorithms. It could also find its use in other audio processing applications such as signal enhancement, signal classification, etc.

8. We notice that the separation performance of the CSR-BS-WPD based separation algorithm depends on the sparsity of the time-frequency representation. It may be possible to define an optimization procedure that would try to improve sparsity of the representation for some given signal class by changing mother wavelet function. The resulting wavelet analysis would have a perfect reconstruction properties on one hand and more sparse signal representation on the other. We speculate that a source separation algorithm based on this signal representation would have better performance. The sparsity of signal representation is often a desired property for many other applications as well.

9. Future work on the CSR-BS-WPD analysis may address various audio pro-

cessing tasks traditionally performed in the STFT domain, which require instantaneous spectral shapes and have some relations to critical bands in human auditory system.

Appendix A

Joint Time-Frequency Analysis

A.1 Short time Fourier transform

Short time Fourier transform (STFT) and its inverse transform are well known tools used for time-frequency signal analysis. It is especially useful for non stationary signals analysis such as audio signals like speech or music.

We use the following definition of the continuous version of the STFT

$$X_{\omega}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} w_c(\tau - t)\, x_c(\tau)\, e^{-j\omega\tau} d\tau \tag{A.1}$$

where $w_c(t)$ is an analysis window and $x_c(t)$ a continuous time signal.

Let $x(n) = x_c(nT)$ be a discrete time domain signal where T is a sampling period. In the discrete case, the STFT transform is defined by

$$X_k(m) = \sum_{n=-\infty}^{\infty} w(mM - n)\, x(n)\, e^{-j\frac{2\pi}{N}kn} \tag{A.2}$$

where k and m are frequency and time indices respectively, $w(n)$ is an analysis

window and M is number of overlapping samples between consequent analysis frames. In practical application, the support of the analysis window $w(n)$ is finite. Let N be the support of $w(n)$. In this case (A.2) reduces to

$$X_k(m) = \sum_{n=0}^{N-1} w(mM - n)\, x(n)\, e^{-j\frac{2\pi}{N}kn} \tag{A.3}$$

The inverse STFT (ISTFT) transform is given by

$$x(n) = \sum_{m=-\infty}^{\infty} \sum_{k=0}^{N-1} X_k(m)\, \tilde{w}(mM - n)\, e^{j\frac{2\pi}{N}k(n-mM)} \tag{A.4}$$

where $\tilde{w}(n)$ is a synthesis window. Substituting (A.2) into (A.4) gives the completeness condition

$$\sum_m w(n - mM)\, \tilde{w}(n - mM) = \frac{1}{N}$$

which ensures perfect reconstruction. For a given analysis window w there might be infinitely many perfectly reconstructing synthesis windows. We use synthesis window $\tilde{w}$ which is bi-orthogonal to w.

A.2 Discrete wavelet transform

Discrete Wavelet Transform (DWT) is a time-frequency signal analysis. Unlike STFT transform that provides uniform time-frequency resolution at all frequencies, DWT provides varying time-frequency resolution at different frequencies: low frequency regions are analyzed with fine frequency resolution and course time resolution while high frequency regions analyzed with course frequency resolution and high time resolution [68].

DWT can be implemented as a series of half-band filters. Input signal is filtered using a high pass half-band filter $h(n)$ and a low pass half-band filter $g(n)$. The output of both filters are decimated by 2. The output signal $d_{j+1,k}$ of

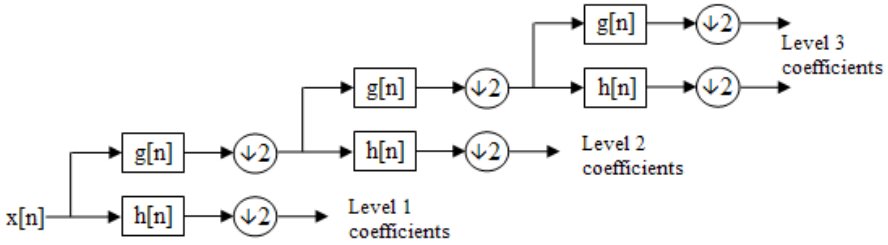


Figure A.1: Discrete wavelet decomposition

"detail filter" h are usually referred as "detail signal" and the output signal $a_{j+1,k}$ of "approximation filter" g is referred as "approximation signal". The "filtering and decimation" procedure may be repeated several times. The detail signals from all levels of the decomposition, together with the approximation signal of the last level of the decomposition are the DWT coefficients. The entire DWT process is displayed in Fig. A.1. The number of times the half-band filters are applied (hence the depth of the decomposition tree) depends on the application requirements.

More precisely, let $j \in \{1, \ldots, J\}$ and J be the number of DWT decomposition levels. The DWT transform is given by following recursive formulas:

$$\begin{aligned} d_{j+1,m} &= \sum_{n=-\infty}^{\infty} h\left(n-2m\right) a_j\left(n\right) \\ a_{j+1,m} &= \sum_{n=-\infty}^{\infty} g\left(n-2m\right) a_j\left(n\right) \end{aligned}$$

Under some conditions on h and g the inverse DWT (IDWT) exists.

Wavelet Packet Decomposition (WPD) is different from the DWT in the sense that it allows to perform the "filtering and decimation" procedure not only on the approximation signals but also on the detail signal of each tree level. The removal of this constraint imposed by the DWT makes it possible to control the resolution of the WPD analysis at different frequencies.

A.3 Mapping based complex wavelet transform

In this section we describe disadvantages of the standard Discrete Wavelet Transform (DWT) and present a Mapping Based Complex Wavelet Transform (CWT), introduced in [58] that mitigates these disadvantages to some degree.

A major disadvantage of the DWT that reduces its usefulness in audio signal processing applications is the lack of shift invariance. Let $x(n)$ be a time domain signal and $X_{l,n}(m) = \mathrm{DWT}\{x(n)\}$ its DWT transform. Let $x_\Delta(n) = x(n-\Delta)$ be a shifted version of the time signal. The DWT coefficients of $x_\Delta(n)$ change significantly compared to $X_{l,n}(m)$. The reason for this behavior lies in the downsampling performed by the DWT on the dilated signals. A short survey of techniques used to mitigate lack of shift invariance may be found in [58].

Let $L^2(\mathbb{R} \to \mathbb{C})$ denote a function space of square integrable complex-valued functions on a real line and $L^2(\mathbb{R} \to \mathbb{R})$ its subspace comprised of real-valued functions. Hardy-space $H^2(\mathbb{R} \to \mathbb{C})$ is defined by

$$H^2(\mathbb{R} \to \mathbb{C}) \triangleq \left\{ f \in L^2(\mathbb{R} \to \mathbb{C}) : F(\omega) = 0 \text{ for a.e. } \omega < 0 \right\}$$

where F is a Fourier transform of f.

In [58], a function space $L^2(\mathbb{R} \to \mathbb{R})$ is shown to be isomorphic to Hardy-space $H^2(\mathbb{R} \to \mathbb{C})$ under certain conditions. It is also shown therein, that the mapping of a function in $L^2(\mathbb{R} \to \mathbb{R})$ into Hardy-space cannot be implemented using a digital filter. Softy-space is a practical approximation of a Hardy-space. The mapping into Softy-space is done using a digital filter h^+. From now on, we denote signals in Softy-space by superscript "+".

Forward CWT transform is defined by a map of a time domain signal $x(n)$ into its Softy-space image $x^+(n)$ followed by standard DWT transform. The inverse CWT transform consists of Inverse Discrete Wavelet Transform (IDWT) followed by the inverse mapping from the complex valued Softy-space back to the real valued time signal.

An algorithm presented in Chapter (4), benefits from approximate shift invariance property. As explained therein, we train the GMM model using the wavelet transform coefficients. Lack of shift invariance adds redundancy to the signal space making it larger and the amount of training data required will grow accordingly.

Appendix B

Approximate W-DO orthogonality

W-disjoint orthogonality (W-DO) was introduced and studied by Yilmaz and Richard [34] for speech signal sources. It is not obvious that using binary mask in the case of two speakers or speaker and music, in order to reject an interfering source in the STFT domain should lead to a separation result of satisfying quality. In this section we describe a concept of approximate W-DO and the way it is used to justify binary mask usage for source separation.

Two continuous functions $s_1(t)$, $s_2(t)$ are called disjoint orthogonal if

$$s_1(t)\, s_2(t) = 0 \qquad \forall t \tag{B.1}$$

Two continuous functions $s_1(t)$, $s_2(t)$ are called W-disjoint orthogonal if the following condition hold for their STFT (A.1) mappings $S_{1,\omega}(t)$, $S_{2,\omega}(t)$

$$S_{1,\omega}(t)\, S_{2,\omega}(t) = 0 \qquad \forall \omega, t \tag{B.2}$$

The "W" in the "W-DO" stands for the analysis window of the STFT transform.

In equation (A.1) it is denoted by a small w_c instead of W. For a special case where $w_c(t) = \delta(t)$, the W-DO condition (B.3) reduces to disjoint orthogonality in time (B.1). For a special case of $w_c(t) \equiv 1$ the STFT transform reduces to a regular Fourier transform and W-DO condition (B.3) means that two signals are disjoint in the frequency domain.

In a discrete signal case, a discrete STFT transform is used. In this case very few speech or music coefficients will actually be zero. However, due to the sparsity of two signals in the STFT domain and their independence, only a small amount of energy from both signals will reside in same time-frequency bins. A relaxed version of (B.3) can be used

$$S_{1,\omega}(t)\, S_{2,\omega}(t) \approx 0 \qquad \forall \omega, t \tag{B.3}$$

and some measures of the approximate W-DO can be defined based on the analysis of energy in different time-frequency bins.

Let $M_\omega(t)$ be a time frequency mask. The preserved-signal ratio ($\mathrm{PSR}^{\mathrm{W}}$) defines the portion of a signal $c \in \{1, 2\}$ that remains after applying a mask $M_\omega(t)$

$$\mathrm{PSR}^{\mathrm{W}}_{M,c} \triangleq \frac{\left\| M_\omega(t)\, \hat{S}_{c,\omega}(t) \right\|^2}{\left\| \hat{S}_{c,\omega}(t) \right\|^2}$$

and signal-to-interference ratio ($\mathrm{SIR}^{\mathrm{W}}$) as

$$\mathrm{SIR}^{\mathrm{W}}_{M,c} \triangleq \frac{\left\| M_\omega(t)\, \hat{S}_{c,\omega}(t) \right\|^2}{\left\| M_\omega(t)\, \hat{S}_{3-c,\omega}(t) \right\|^2}$$

In the notations we use a super index W to avoid confusion with SIR measure

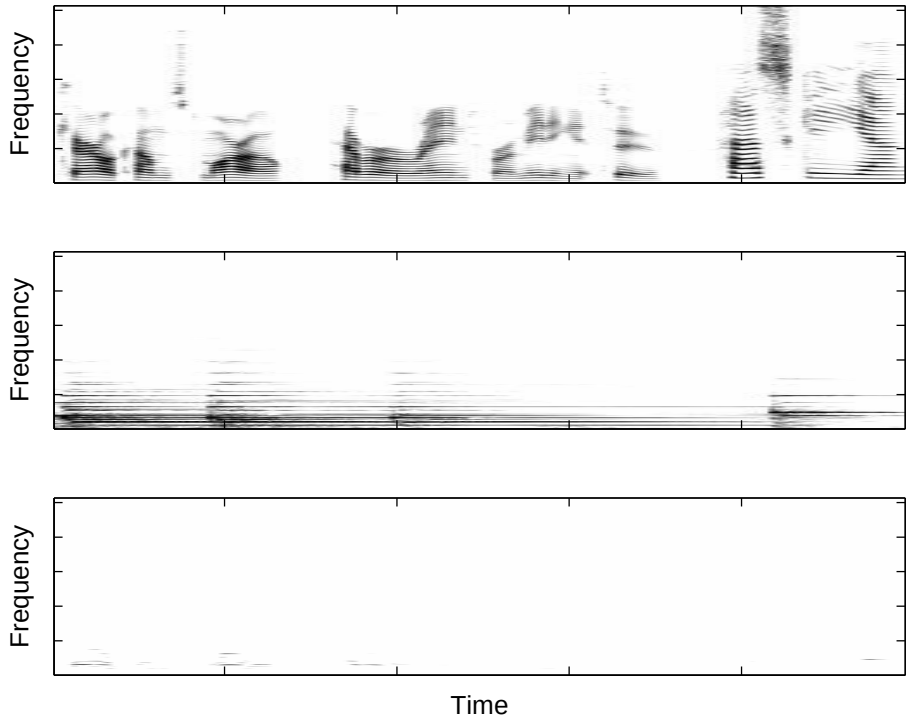


Figure B.1: Two upper panes show spectrograms of sample speech and piano music signals $S_{1,k}(m)$, $S_{2,k}(m)$ (the intensity map is logarithmic). Both signals are normalized to have unit energy in time. The lower pane shows $S_{1,k}(m)\, S_{2,k}(m)$ (also on a logarithmic scale of gray scale intensity). We see that only a small amount of energy resides at the same time-frequency bins, i.e. the property of the approximate W-DO holds for these signals.

defined in Section 2.4. The measure of W-DO is defined by

$$\begin{aligned} \mathrm{WDO}_M &\triangleq \frac{\left\| M_\omega(t)\,\hat{S}_{c,\omega}(t) \right\|^2 - \left\| M_\omega(t)\,\hat{S}_{3-c,\omega}(t) \right\|^2}{\left\| \hat{S}_{c,\omega}(t) \right\|^2} \\ &= \mathrm{PSR}^{\mathrm{W}}_{M,c} - \frac{\mathrm{PSR}^{\mathrm{W}}_{M,c}}{\mathrm{SIR}^{\mathrm{W}}_{M,c}} \end{aligned}$$

If signals are W-disjoint orthogonal and we choose a binary mask $M_\omega(t)$ that coincide with support of signal c. In this case The $\mathrm{PSR}^{\mathrm{W}}_{M,c} = 1$ and $\mathrm{SIR}^{\mathrm{W}}_{M,c} = \infty$ (because the $M_\omega(t)$ and the support $\hat{S}_{3-c,\omega}(t)$ are disjoint) hence the $\mathrm{WDO}_M = 1$. On the other hand, if the support of both signals is the same, then $\mathrm{PSR}^{\mathrm{W}}_{M,c} = 1$ and $\mathrm{SIR}^{\mathrm{W}}_{M,c} = 1$ hence $\mathrm{WDO}_M = 0$.

Subjective listening tests conducted in [34] indicate that WDO have good correlation with rating given by human listeners. For example, WDO values that are greater then 0.8 coincide with perfect perceptional quality indicated by human listeners. WDO values between 0.6 and 0.8 coincide with "minor artifacts or interference" rating.

Bibliography

[1] E. Vincent, C. Févotte, L. Benaroya, and R. Gribonval, "A tentative typology of audio source separation tasks," in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 715–720.

[2] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 957–961.

[3] C. E. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[4] S. H. Godsill and P. J. Rayner, *Digital Audio Restoration: A Statistical Model Based Approach*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1998.

[5] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5 – 25, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1C-4D98BTC-2/2/ef01d0b17c355314957bd15af09c496e

[6] "The elisa systems for the nist'99 evaluation in speaker detection and tracking," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 143 – 153, 2000. [On-

line]. Available: http://www.sciencedirect.com/science/article/B6WDJ-45F541V-B/2/53de85033c951dfa215175ba40619208

[7] M. Goto and S. Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis, pp.31-40, Aug 1999.

[8] M. Casey, "General sound classification and similarity in mpeg-7," *Organised Sound*, vol. 6, no. 02, pp. 153–164, 2001.

[9] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 4, pp. 354–358, 1978.

[10] B. Hanson and D. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-84*, vol. 9, Mar 1984, pp. 65–68.

[11] J. Herault, C. Jutten, and B. Ans, "Detection de grandeurs primitives dans un message composite par une architecture de calcul neuromimetique en apprentissage non supervise." GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 1985.

[12] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[13] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Interscience, May 2001.

[14] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Comput.*, vol. 13, no. 4, pp. 863–882, 2001.

[15] A. Bregman, *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1990.

[16] F. R. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 65–72.

[17] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-00*, vol. 2, 2000, pp. II765–II768 vol.2.

[18] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. on ASLP*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[19] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on ASLP*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19 – 41, 2000.

[21] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. on ASLP*, vol. 14, no. 1, pp. 163–176, 2006.

[22] ——, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. on ASLP*, vol. 15, no. 2, pp. 441–452, 2007.

[23] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Springer, 2009.

[24] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[25] B. Wang and M. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in *In Proc. UK Digital Music Reserarch Network (DMRN) Summer Conf.*, 2005.

[26] P. O. Hoyer, "Non-negative sparse coding," *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565, 2002.

[27] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *International Computer Music Conference, ICMC*, 2003.

[28] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-06*, vol. 5, pp. V–V, May 2006.

[29] ——, "Shifted non-negative matrix factorisation for sound source separation," *IEEE/SP 13th Workshop on Statistical Signal Processing*, pp. 1132–1137, July 2005.

[30] M. Kim and S. Choi, "Monaural music source separation: Nonnegativity, sparseness, and shift-invariance," in *ICA*, 2006, pp. 617–624.

[31] I. Cohen, "Enhancement of speech using bark-scaled wavelet packet decomposition," in *Eurospeech*, 2001, pp. 1933–1936.

[32] F. Fernandes, R. van Spaendonck, and C. Burrus, "A new framework for complex wavelet transforms," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1825–1837, July 2003.

[33] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. 4th International*

Symposium on ICA and BSS (ICA2003), Nara, Japan, Apr. 2003, pp. 763–768.

[34] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[36] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide revision 2.0," IRISA, Rennes, France, Tech. Rep. 1706, April 2005. [Online]. Available: http://www.irisa.fr/metiss/bss-eval/

[37] H. M. Teager and S. M. Teager, "A phenomenological model for vowel production in the vocal tract," in *Speech Science: Recent Advances*, R. G. Daniloff, Ed. San Diego, CA: College-Hill Press, 1985, ch. 3, pp. 73–109.

[38] ——, *Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract*, W. J. Hardcastle and A. Marchal, Eds. Boston: Kluwer Academic, 1989, vol. 55.

[39] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-90*, Apr 1990, pp. 381–384 vol.1.

[40] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct 1993.

[41] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-91*, 1991, pp. 421–424 vol.1.

[42] P. Maragos, J. Kaiser, and T. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1532–1550, Apr 1993.

[43] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," vol. 1, May 1995, pp. 784–787 vol.1.

[44] ——, "Time-frequency distributions for automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 196–200, Mar 2001.

[45] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Speaker identification using FM features," in *Proc. of 11th Australasian International Conference on Speech Science and Technology*, Auckland, New Zealand, 2006, pp. 148–152.

[46] D. V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 621–624, Sept. 2005.

[47] J. Jankowski, C.R., T. Quatieri, and D. Reynolds, "Measuring fine structure in speech: application to speaker identification," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-95*, vol. 1, May 1995, pp. 325–328 vol.1.

[48] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an am-fm modulation model," *Speech Communication*, vol. 28, no. 3, pp. 195 – 209, 1999.

[49] R. Sussman and M. Kahrs, "Analysis and resynthesis of musical instrument sounds using energy separation," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-96*, vol. 2, 1996, pp. 997–1000 vol. 2.

[50] S. Schimmel, L. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proc. Int. Conf. on*

Acoustics, Speech and Signal Processing, ICASSP-07, vol. 4, April 2007, pp. IV–605–IV–608.

[51] L. Atlas and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-05*, vol. 4, March 2005, pp. iv/461–iv/464 Vol. 4.

[52] S. M. Schimmel, "Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices," Ph.D. dissertation, University of Washington, 2007.

[53] Q. Li and L. Atlas, "Coherent modulation filtering for speech," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-08*, 31 2008-April 4 2008, pp. 4481–4484.

[54] L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-04*, vol. 2, May 2004, pp. ii–761–4 vol.2.

[55] M. Poletti, "The homomorphic analytic signal," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 1943–1953, Aug 1997.

[56] J. F. Kaiser, "On teager's energy algorithm and its generalization to continuous signals," *Proc. IEEE DSP Workshop New Paltz, NY*, Sept 1990.

[57] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.

[58] F. C. A. Fernandes, "Directional, shift-insensitive, complex wavelet transforms with controllable redundancy," 2002, chairman-Burrus, C. Sidney.

[59] R. Dwyer, "Detection of non-gaussian signals by frequency domain kurtosis estimation," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-83*, vol. 8, 1983, pp. 607–610.

[60] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 282 – 307, 2006.

[61] J. Antoni and R. Randall, "The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 308 – 331, 2006.

[62] J. Antoni, "Fast computation of the kurtogram for the detection of transient faults," *Mechanical Systems and Signal Processing*, vol. 21, pp. 108–124, Jan. 2007.

[63] E. Nemer, R. Goubran, and S. Mahmoud, "Snr estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Process. Lett.*, vol. 6, no. 7, pp. 171–174, Jul. 1999.

[64] P. Ravier and P. O. Amblard, "Denoising using wavelet packets and the kurtosis: application to transient detection," in *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 6–9 Oct. 1998, pp. 625–628.

[65] J. J. G. de la Rosa, C. G. Puntonet, and A. Moreno, "Subterranean termite detection using the spectral kurtosis," in *Proc. 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS 2007*, 2007, pp. 351–354.

[66] V. Vrabie, P. Granjon, and C. Servière, "Spectral kurtosis: from definition to application," in *IEEE-EURASIP International Workshop on Nonlinear Signal and Image Processing*, Grado, Italy, 2003.

[67] J. Benesty, private communication, Jul 2009.

[68] S. Mallat, *A Wavelet Tour of Signal Processing, Second Edition (Wavelet Analysis & Its Applications)*. Academic Press, September 1999.