Automatic Music Transcription using Structure and Sparsity

Ken O'Hanlon

Submitted for the degree of Doctor of Philosophy

Queen Mary, University of London

2014

Automatic Music Transcription using Structure and Sparsity

Ken O'Hanlon

Abstract

Automatic Music Transcription seeks a machine understanding of a musical signal in terms of pitch-time activations. One popular approach to this problem is the use of spectrogram decompositions, whereby a signal matrix is decomposed over a dictionary of spectral templates, each representing a note. Typically the decomposition is performed using gradient descent based methods, performed using multiplicative updates based on Non-negative Matrix Factorisation (NMF). The final representation may be expected to be sparse, as the musical signal itself is considered to consist of few active notes. In this thesis some concepts that are familiar in the sparse representations literature are introduced to the AMT problem. Structured sparsity assumes that certain atoms tend to be active together. In the context of AMT this affords the use of subspace modelling of notes, and non-negative group sparse algorithms are proposed in order to exploit the greater modelling capability introduced. Stepwise methods are often used for decomposing sparse signals and their use for AMT has previously been limited. Some new approaches to AMT are proposed by incorporation of stepwise optimal approaches with promising results seen. Dictionary coherence is used to provide recovery conditions for sparse algorithms. While such guarantees are not possible in the context of AMT, it is found that coherence is a useful parameter to consider, affording improved performance in spectrogram decompositions.

Submitted for the degree of Doctor of Philosophy

Queen Mary, University of London

2014

I, Ken O'Hanlon, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged and my contribution indicated. Previously published material is also acknowledged.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third partys copyright or other Intellectual Property Right, or contain any confidential material.

Ken O'Hanlon
28/09/2013

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

First of all I would like to thank my supervisor, Mark Plumbley, for his guidance at crucial moments and ultimate patience at what otherwise could have been very frustrating times. Without his assistance and insistence on a writing deadline, this project might never have reached completion. Thanks to Hidehisa Nagano and Nicolas Kerevin, two fun characters who visited C4DM, who I had the pleasure of collaborating with. Hidehisa expended a lot of time in developing my technical writing, an endeavour which, hopefully, has filtered right down to this thesis. Nicolas visited C4DM while this thesis was being written, and daily communication with him on related topics helped to structure my thoughts and offset the solitary nature of writing such a document.

On a less work related note, my deepest gratitude to my wife, Elaine, for her persistence, proofreading and for marrying me during the duration of this project. Also to my parents, brother and sister, for their everlasting support and to my extended family and friends for their kind words of encouragement.

Part of the PhD project involves constant interaction with other researchers, and I consider myself lucky to have been located in the C4DM research group during this time, meeting with many excellent researchers, both located there and visiting. I would like to thank them all, particularly those involved in the sparse representations group and researchers with an interest in music transcription, from whom I learnt a lot through discussions.

Last but not least, I would like to thank my examiners, Pierluigi Dragotti and Patrick Wolfe for helpful comments, constructive criticism, and particularly for making the viva a pleasant, good-humoured, experience

List of Common Symbols

$\mathbf{D}$	Matrix.
$\mathbf{x}$	Vector.
a	Scalar.
$\mathbf{d}_n$	nth column of the the matrix $\mathbf{D}$.
$\mathbf{d}^m$	mth row of the the matrix $\mathbf{D}$.
$d_{m,n}, [D]_{m,n}$	Element in mth row and n th column of matrix $\mathbf{D}$.
$\mathbf{D}_\Gamma$	Submatrix containing all columns of $\mathbf{D}$ indexed by Γ.
x_n	nth element of vector $\mathbf{x}$.
$\mathbf{x}^{[a]}$	Elementwise exponentation of $\mathbf{x}$ to the power of a.
$\otimes, \oslash$	Elementwise matrix or vector multiplication and division, respectively.
$\mathbf{D}[l], \mathbf{x}[l]$	Grouped elements of dictionary and activation vector , respectively.
$\mathbf{D}^T$	Transpose of matrix, or vector.
$\mathbf{D}^\dagger$	Moore-Penrose pseudoinverse of matrix.
$\mathcal{L}$	Set containing indices of individual groups.
$\mathcal{C}$	Cost function.
$\mathcal{P}, \mathcal{R}, \mathcal{F}$	Precision, Recall and $\mathcal{F}$-measure metrics.
$\mathcal{M}$	Molecule formed from collection of atoms.
$\|\|\mathbf{x}\|\|_p$	ℓ_p vector norm.
$\|\|\mathbf{x}\|\|_{p,q}$	$\ell_{p,q}$ mixed vector norm.
$\|\|\mathbf{X}\|\|_F$	Frobenius norm of matrix

Chapter 1

Introduction

Automatic Music Transcription seeks to enable a machine to understand a musical piece. Tonal musical instruments produce pitched signal elements, known as musical notes, consisting of harmonic frequency partials. Machine understanding of a tonal piece of music is often framed in terms of activity in a pitch-time representation. While conversion of an audio signal to a frequency-time representation, or spectrogram, is simply performed using Fourier analysis, derivation of an appropriate pitch-time representation is a difficult task and Automatic Music Transcription is an active research area.

Spectrogram decompositions are a popular model-based approach to AMT. A model, referred to as an *atom*,is constructed for each note, typically using a representative frequency spectrum. A collection of these atoms form a *dictionary*, and spectrogram decompositions seek to approximate the spectrogram as an additive combination of dictionary atoms. When the individual atoms are pitch-labelled, the spectrogram approximation coefficients, or activation matrix, relating the activity of individual atoms forms a pitch-time representation. Often in spectrogram decompositions phase information of the signal is discarded and a non-negative magnitude spectrogram is employed.

While spectrogram decompositions consider fixed dictionaries that are trained offline, spectrogram factorisations learn a dictionary and its corresponding spectrogram decomposition simultaneously. Typically, as a magnitude spectrogram is used, the factorisation is performed using methods based upon the Non-negative Matrix Factorisation (NMF) methodology [74]. Spectrogram decompositions are often performed using NMF-based approaches.

Sparse representations are an active research field across the areas of signal processing, statistics and machine learning, The classic interpretation of sparse representations relates to the use of *overcomplete dictionaries*, which consist of a number of atoms greater than the dimension of each atom, and the goal is to decompose a signal using such a dictionary. Early developments in the field of sparse representations include the proposal of now well-known methods for sparse approximations. These include greedy algorithms such as Matching Pursuit (MP) [80] and Orthogonal Matching Pursuit (OMP) [98] while Basis Pursuit (BP) [23], also known as ℓ_1 minimisation, applies a ℓ_1-norm sparse penalty to the least squares problem.

When an overcomplete dictionary is used the decomposition problem is considered underdetermined. Historically the perspective was taken that a unique solution was not available for underdetermined problems. However, it was recently discovered that the uniqueness of a solution to an underdetermined problem was guaranteed if the solution was sufficiently sparse [34], and conditions on the correlation of dictionary atoms were met. Dictionary correlation is typically related through the parameter of *coherence*, a fundamental property in much sparse representations research. Later developments showed that sparse recovery was guaranteed using BP and OMP when conditions, expressed simply through coherence, were met [128].

Other interesting developments have built upon these algorithmic and theoretical foundations of sparse representations. Structured sparse representations introduce the capability to model structure that is inherent in signals, and many variants proposing the incorporation of various types of structure exist. Group, or block, sparsity, [137] [37] focusses on the interrelations of atoms in a single representation vector, while multichannel [130], molecular [30] and neighbourhood [70] sparse approximations consider the co-activation of similar signal elements in different coefficient vectors. Theoretical results show that such grouping of dictionary elements can lead to improved conditions for signal recovery in certain cases [130] [37]. Another interesting element of the sparse representations field is sparse dictionary learning, which, similar to NMF, seeks to learn a dictionary and its representation in a signal simultaneously,

1.1 Motivations and Aims

The motivation behind the research presented in this thesis is the fact that musical signals are naturally sparse in a pitch-time representation, as few notes are active at any given time. While music that is not pitch-sparse is perhaps conceivable in some sense, it is known from the cognitive precepts of auditory science that there is a limit to human capabilities to separate streams from a auditory signal [13]. AMT, when performed on a frame-level basis, attempts to transform a frequency spectrum frame into a pitch activation vector, and as such can be formulated as a sparse subset selection problem. While the desirability of sparsity in musical spectrograms decompositions has been stated [1] [125] [32], there is relatively little prior research that explicitly considers sparsity as a factor in such approximations.

The aim of this thesis is to explore the incorporation of constructs and methodologies from the sparse representations repertoire to the field of AMT. In drawing knowledge from the sparse representations community towards the AMT problem, it is important to consider the different context to which sparse representations theory, and practice, generally apply. For instance pursuit algorithms are designed for use with overcomplete dictionaries that are relatively incoherent, while theoretical conditions in sparse representations assume a similar context. Conversely, undercomplete dictionaries are often employed in the AMT problem, where correlated signal elements are expected due to the structure of tonal music. More pertinently, dictionary atoms are semantically meaningful when performing AMT, whereby one must consider that the sparsest signal representation may in fact not provide the best transcription. In this respect sparsity here may be considered a useful tool but not the ultimate goal in decompositions.

The perspective taken here is that the AMT problem may be informed by algorithmic and theoretical developments in sparse representations. For instance, note modelling in many AMT methods uses a single spectral template to represent a note, while the spectral shape of notes played by many instruments is known to evolve over time. The framework of group sparsity provides a simple explicit model for dealing with grouped atoms, affording the use of a subspace modelling approach for notes that may assist in the AMT problem. Molecular sparsity, another variant of structured sparsity, provides a simple approach to dealing with structure in activation matrices. A large problem when performing AMT is the presence of correlated sources. Consideration of the dictionary coherence parameter, which considers correlated atoms, may provide a new perspective to AMT.

Stated concisely, the stated main aim of this thesis is to explore if exploitation of the sparse representations knowledgebase can afford better decompositions for musical signal processing tasks, than are currently available. The approach taken can be considered deterministic in a sense as no explicit recourse to probabilistic methods is taken. Several algorithmic developments are proposed in this thesis, most of which are numerically based with sparsity or group sparsity of the coefficient vectors assumed. Such approaches can be considered general algorithms that may find application in other arenas. Alternatively, application-specific prior information is incorporated occasionally, in particular time-continuity of signal elements and harmonicity of dictionary atoms. In both cases this is performed in a simple manner, without the requirements of priors or penalties in subsequent calculations.

1.2 Thesis Overview

Chapter 2 is devoted to background material relevant to the research presented in this thesis. The Automatic Music Transcription (AMT) problem is introduced and described with reference to the harmonic structure of western tonal music, metrology of AMT, and a brief literature review describing some well-regarded methodologies in the field of AMT. Further sections describe some sparse decomposition and matrix factorisation methods, with a particular focus given to these problems in a non-negative framework, before reference to some prior work applying these methods to AMT.

Chapter 3 describes the experimental workflow used throughout much of this thesis. The goal of this thesis is to explore the use of different sparse, structured and non-negative decompositions for the purpose of AMT. Comparison of different approaches to be taken suggests the use of a repetitive experimental setup. However, several variations are used within the experimental framework employed. For instance, various transforms may be used to produce a spectrogram and different classes of dictionaries may be used. Each of these variations is used repeatedly in the following chapters, and they are described in detail here.

Chapter 4 explores the use of greedy OMP-based methods for AMT. In particular a comparison between datapoint modelling and subspace modelling is made in the context of these methods. Non-negative variants of algorithms from the group sparse methodology are proposed for the purpose of AMT. The non-negative constraint introduces an extra computational load and efforts to alleviate this expense are undertaken. Some problems concerning the use of greedy methods in the context of AMT are noted.

Chapter 5 introduces stepwise methods incorporating backwards and forwards steps to AMT, with a focus on methods that take locally optimal steps. Considering the non-negative framework, a backwards elimination strategy is proposed, and compared to other prior stepwise algorithms proposed in the sparse literature. Close observation of the backwards elimination criteria leads to a proposed modification to the sparse cost function normally employed. Finally, a group sparse backwards elimination method is proposed, with a similarly modified sparse cost function.

Chapter 6 continues the study of greedy and stepwise methods. In this chapter their scope is extended by the introduction of temporal structure. A greedy molecular approach is proposed to deal with coherence-induced problems observed when adapting prior methods to the AMT problem. Analysis of decomposition-based methods is proposed through a simple oracle decomposition, shedding some light on the decomposition-based AMT problem. A molecular norm is defined that suggests easy adaptation of stepwise and thresholding based approaches to molecular decompositions and experimental results compare these different approaches.

Chapter 7 considers gradient-based methods for non-negative matrix decompositions. Alternative cost functions to the Euclidean distance, typically used in sparse representations, are explored. A novel generalised cost function is proposed, and proof of monotonic descent of this cost function using a multiplicative update is given. Finally sparse and group sparse penalisation strategies are explored.

Chapter 8 provides a new analysis of the AMT problem using the sparsity-based construct of dictionary coherence. An analysis of dictionaries from different signal transforms is provided with reference to transcription results and the coherence of the dictionaries. A row-weighted decomposition approach is proposed, and a novel effective coherence measure is introduced in order to derive the row-weighting used. Experimental results show improved AMT results for the proposed method.

Chapter 9 considers the use of sparsity in NMF. First, an example is given showing that incorporation of sparsity may be important in the context of NMF for musical signals. A variant on Sparse-NMF (S-NMF) is proposed using the backwards elimination non-negative sparse approximation algorithm proposed in Chapter 5. Finally, group sparse spectrogram decompositions using a synthetic dictionary are compared with a state-of-the art NMF approach.

Chapter 10 concludes the thesis with a reflection on the presented research, with suggestions for further possible avenues of research.

1.3 Associated publications and collaborations

All of the research presented in this thesis was undertaken during a course of study lasting from 2009 to 2013 at the Centre for Digital Music, C4DM, in Queen Mary, University of London. Some of this work was undertaken in collaboration with visiting researchers at C4DM, namely Hidehisa Nagano, a visiting researcher from the NTT Communications Lab in Japan, and Nicolas Keriven, a visting student from CMAP at Ecole Polytechnique in France. During this course of research, several contributions have been presented at international peer-reviewed conferences, some of which was collaborative. The publications are listed below with respect to their presentation in the this thesis, and the contribution of visiting researchers is noted.

- Research described in Chapter 4.1-4.2 was presented at the *3*rd *IMA International Conference on Linear Algebra and Optimisation* [94].

- Research described in Chapter 5.2-5.3 was presented at the *IEEE Workshop on Machine Learning for Signal Processing (MLSP), 2013* [88]. This was a collaborative publication with Nicolas Keriven. The research presented in Chapter 5 is the work of this author, while Nicolas undertook research beyond what is presented in this thesis.

- Research described in Chapter 6.1 was presented at the *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), 2012* [91]. Chapter 6.2 describes research presented at the *International Conference on Music Modelling and Retrieval (CMMR), 2012* [90], and which is to be published in a Lecture Notes in Computer Science series. This body of work was undertaken in collaboration with Hidehisa Nagano.

- A presentation made at the *5th International Workshop in Music and Machine Learning Workshop (MML), 2012* [89] forms a precursor to work described in Chapter 7. The presentation relates some initial work on group sparse NMF undertaken with Hidehisa Nagano. However the work presented in this thesis is significant departure from that initial work.

- Research described in Chapter 8 was presented at *IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), 2013* [95].

- Research described in Chapter 9.1 was presented at the *European Signal Processing Conference (EUSIPCO), 2011* [92], and also at *4th Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS), 2011* [93]. Research described in Chapter

9.2 was presented at the *IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2013* [96].

Chapter 2

Background

In order to establish the foundation upon which the research presented in this thesis is based, a summary of relevant background material is presented in this chapter. First, the Automatic Music Transcription (AMT) problem is described and the reader is referred to some well-known approaches used in the AMT field. Following this the sparse representations methodology is introduced, with some focus given to structured sparse representations and recovery conditions. The subsequent section attends to non-negative approximations such as Non-Negative Least Squares (NNLS) and the Non-negative Matrix Factorisation (NMF) methodology. Finally, some applications of sparse representations and NMF methods to musical machine listening are outlined before concluding.

2.1 Automatic Music Transcription

Fourier analysis, developed in the early 19th century [79], affords the capability to characterise a waveform in terms of its frequency content. While the simplest waves are described by a simple sine function, more complex waveforms are often seen to be periodic and Fourier analysis allows such waveforms to be represented as a superposition of several sine waves each parametrised by their amplitude and phase. It is very useful that many naturally occurring waveforms are seen to be sparse, with energy concentrated in few significant elements, when viewed in this frequency domain. With the development of modern electronic computing devices and algorithms such as the Fast Fourier Transform (FFT), many different types of waveforms from natural signals can be analysed and their information exploited.

Many interesting applications have been developed in line with these new computational possibilities. Of particular interest here is the field of machine listening, which attempts to endow machines with the capability to make sense of an audio environment. The chief goals in machine listening are the recognition of sounds and the separation of sources, which can be viewed as symbiotic, with better recognition leading to better separation and vice versa. Machine listening is a multidisciplinary endeavour, drawing from seemingly disparate fields, often to solve the same problem. For instance, two well known approaches to source separation are Blind Source Separation [28], which uses physics and machine learning approaches, and Computational Audio Scene Analysis (CASA) [15], which uses cognitive principles defined by the the Audio Scene Analysis of Bregman (ASA) [13]. Many of the principles employed by ASA focus on the grouping of sound elements into streams, and are similar to those used in the Gestalt theory of visual perception [13]. Another knowledge base from which machine listening may draw is auditory science. For instance, it is known that the frequency response of the human inner ear is not linear. Frequency scales, such as the Equivalent Rectangular Bandwidth (ERB) scale, that aim to model the frequency response of the auditory system have been adapted for application to machine listening tasks through use of filterbank systems [69] [132].

The foremost machine listening application is Automatic Speech Recognition (ASR), which is now becoming a mature technology, starting to pervade through its integration in hand-held devices. Indeed this desire to afford interaction with machines through speech, the most concise of human communications, and the commercial potential of developing such machine capabilities has probably weighted much of the initial machine listening research in this direction. In comparison, until now, the processing of signals containing music or sounds from the natural environment has received relatively little attention. However, such technologies are becoming more desirable with the advent of big data science and mobile computing.

An interesting feature of many natural waveforms is the presence of harmonic structures in the frequency domain, consisting of regularly placed peaks in the frequency spectrum. Harmonic structure is due to the physical phenomenon known as standing waves, caused by reflection of a wave at the end of the propagating medium. Many waveforms in the audio frequency range, such as sounds caused by passing air through a tube or by striking a fixed string, conform to this harmonic structure. These harmonic sounds can be primarily parametrised using the concept of *fundamental frequency*, represented by f_0. Using this parameter the location of the energy peaks in a frequency-based representation of such a signal element can be predicted such that:

$$f_n \approx f_0 \times n \tag{2.1}$$

where f_n is the nth harmonic, or harmonic partial, also known as the $(n-1)$th overtone. The concepts of harmonicity and fundamental frequency are applicable in terms of tonal musical instruments, vowel sounds of human speech, and many other sounds experienced in the real world. The use of fundamental frequency as a parameter of these harmonic sounds may help in recognising and separating these sounds. In this sense, the similarity between CASA and multiple fundamental frequency estimation in the grouping of coincident elements has been noted [68]. While fundamental frequency is a objective measurable construct, *pitch* is defined as a perceptual attribute of sounds, defined as the frequency of a sine wave that is matched to the target sound in a psychoacoustic experiment [67]. These two concepts are closely related, and can be considered synonymous in the context of this thesis.

In terms of music, the use of the fundamental frequency parameter affords a symbolic representation that transcends the physical method of sound creation and perceptual features such as the timbre of an instrument. This symbolic representation is often referred to as a *score* and music transcription is the attempt to derive such a representation from a piece of music experienced aurally. Similarly, Automatic Music Transcription (AMT) is the machine attempt to derive a pitch-time representation of a musical piece, typically represented by a digital audio file. While a complete music transcription provides a score in musical notation, with separate instruments assigned to separate score sheets, this level of detail in output is not yet considered by AMT systems [6]. A graphical representation of an AMT output known as a *piano roll* uses a binary matrix representation to denote note activity. An example ground truth piano roll, which is augmented with note onset information, is visible in Figure 2.1. The information contained in a piano roll is often communicated in terms of the the Musical Instrument Digital Interface (MIDI) protocol. The MIDI parameter of note number is used to represent the pitch, performing a discretisation of the f_0 parameter, and the temporal onset and offset parameters are also commonly used in AMT descriptors. Further parameters can also be used to describe a transcription such as the MIDI note velocity [99], representing the intensity of the energy of a note.

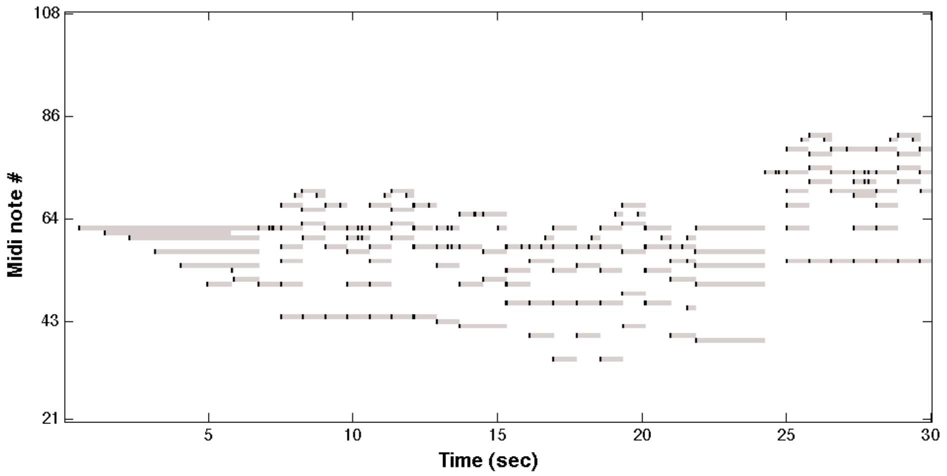

Figure 2.1: Example ground truth piano roll, with note activity marked in grey and note onsets in black.

2.1.1 Piano transcription

A recently published overview paper [6] aims to set out the achievements and challenges facing the AMT field. It is noted in [6] that monophonic pitch, or melody, tracking is considered a solved problem. The authors then consider the problems of AMT from a holistic viewpoint, considering possibilities such as transcription in the presence of percussive sounds or transcription of electro-acoustic music, a larger perspective than is normally taken in most papers describing AMT research. While AMT would ultimately seek to solve such problems, the simpler problem of polyphonic transcription is still seen to be problematic, with a glass ceiling reported using current methods [6].

Often in AMT research a musical signal described as polyphonic is actually played by a single instrument, typically a piano, that is capable of playing several notes simultaneously. The choice of the piano for much AMT research is interesting in that it allows several problems general to AMT to be addressed while ignoring others which need later attention. In a piano each note is activated by a hammer hitting a string and the body of the piano responds with a relatively mild percussive response. The timbre of stringed instruments is known to vary [44] according to where the string is hit, with certain harmonics attenuated according to the shape of the triangle formed when the string is struck. Piano strings are held in a fixed position and hit in a similar spot on each occasion. Therefore some consistency in the timbre of the piano can be expected, a feature which may not be present in recordings of other stringed instruments that are struck or bowed manually. Even so, the timbre of a piano is known to vary over the duration of note [17] [44] [1], with energy at higher frequencies attenuating quicker, while further moderations to the timbre can be effected through the use of pedals and the velocity of the hammer blow [17]. However, the range of timbre variation is still relatively low for a piano.

Furthermore, expressive elements such as vibrato, which may be effected through rapid physical movement of a vibrating string resulting in centred pitch variance, and pitch bending are not achievable with the machinations of a piano. Otherwise put, in the case of piano transcription sampling the pitch may be performed over a coarse discrete f_0 spectrum where each discrete value is a MIDI note, while transcription of other instruments may require a more fine-grained f_0 spectrum. However, even in this limited case, accurate polyphonic transcription is not achievable [6]. In terms of harmonicity it is noteworthy that the length of the piano string causes a slight inharmonicity in the partials of a piano note due to the high tension in piano strings, and their relatively long length [17]. An advantage of the use of a piano for AMT research is the availability of electro-mechanical MIDI pianos, such as the Yamaha Disklavier [39] [104], which allow ground truth performances to be performed mechanically affording standard comparisons of AMT methods. Previously this was only achievable with MIDI files, which are seen to be less challenging in terms of AMT performance [32] [12], while live recordings required hand-labelling of note events and temporal alignment of the score and spectrogram.

2.1.2 Musical structure

One of the chief difficulties in the AMT problem is innately related to the structure of the musical scale. Each musical note represents a single pitch, or fundamental frequency. The term *octave* refers to the relationship between two notes that have fundamental frequencies that can be expressed by the ratio 2 : 1. The music scale consists of 12 differently labelled notes, each of which can be expressed through several octaves to form a larger scale. For instance, the piano keyboard contains 88 keys, each of which represents a separate note, thereby spanning more than seven octaves. The interval between adjacent notes on the scale is referred to as a *semitone*. The ratio of the fundamental frequencies of two given musical notes is given by the expression :

$$\frac{f_0^i}{f_0^j} = 2^{\frac{i-j}{12}} \tag{2.2}$$

Name	$f_0(Hz)$	Equal	Just	Major	Minor
A	440		1	1st	1st
A♯/**B**♭	466.2	1.059	16 : 15	-	-
B	493.9	1.122	9:8	2nd	2nd
C	523.3	1.189	6:5	-	3rd
C♯/**D**♭	554.4	1.259	5:4	3rd	-
D	587.3	1.335	4:3	4th	4th
D♯/**E**♭	622.3	1.414	7:5	-	-
E	659.2	1.498	3:2	5th	5th
F	698.5	1.587	8:5	-	6st
F♯/**G**♭	740.0	1.682	5:3	6st	-
G	784.0	1.782	16:9	-	7th
G♯/**A**♭	830.6	1.888	15:8	7th	-
A	880	2	2	*octave*	*octave*

Table 2.1: One octave of musical scale with note names, fundamental frequency, decimal ratio of equal temperament and rational ratio of just intonation to first **A** note on scale, position of note on the **A** major and **A** minor scales.

where $|i-j|$ is the separation of the notes, in semitones. The frequency scale expressed by (2.2) is referred to as the *equal temperament* scale. Other temperaments, such as Pythagorean tuning and just intonation, use rational numbers for the ratios between fundamental frequencies, as these sounds are considered the most consonant. Consonance is a perceptive measure of the pleasantness of a combination of harmonic sounds played together, and is an important concept in the structure of western tonal music. The equal temperament scale, expressed through (2.2) is a compromise between consonance and other musical considerations. However, the relationship between different notes in the equal temperament approximate the rational ratios previously used. This relationship is outlined in Table 2.1.

The structure of the musical scale, with octave and approximate rational number ratios between fundamental frequencies is problematic for AMT, as many notes contain overlapping harmonic partials in the Fourier transform. Harmonic jumping refers to when a note is detected incorrectly, yet temporally coincident to a ground truth note that is pitched either an octave higher or lower. These harmonic jumps, also known as octave errors, arise from of the harmonic nature of pitched sounds [33] and are common in AMT and f_0 estimation. The problem can be compounded by the co-activation of several notes sharing regular ratios of fundamental frequency. Notes that are consonant are often played in close temporal proximity to each other, or simultaneously using the musical structures of keys and chords, respectively. A musical key refers to a fixed scale, such as the major and minor scales outlined in Table 2.1, that consist of a subset of the 12 notes, with consonance being a feature, particularly for the major scale. Often musical pieces are constrained to notes in a given key, typically resulting in consonant notes being played in sequence. A combination of notes played simultaneously is referred to as a *chord*. Commonly used chords contain the most consonant combinations of sounds. For example the major chord contains the first, third and fifth notes of the major scale in Table 2.1, which can be seen to have some of the simplest relationships in terms of the just intonation.

2.1.3 Metrology of AMT

It is an old adage that something cannot be improved if it cannot be measured. In the field of AMT, comparison of system performance is complicated by different approaches to transcription metrology being taken with different metrics employed and a variety of datasets being used [6]. The performance of an AMT system is most commonly measured in terms of a frame-based analysis in which the ground truth and AMT output are compared at each pitch-time point. In this comparison, correct detections, incorrect detections and undetected ground truth elements are labelled. The correct detections, referred to as true positives, comprise the set $\mathcal{TP}$, while $\mathcal{FP}$ denotes the set of false positives, or incorrect detections. Similarly, undetected ground truth elements, or false negatives, form the set $\mathcal{FN}$. Other forms of AMT analysis are possible. An event based analysis, that compares the onsets of the ground truth and the AMT output is proposed in [104]. In this case a true positive is denoted when a correctly pitched detection occurs within a stated time-tolerance, typically $50-200ms$, of a ground truth detection. Common machine learning and pattern recognition metrics such as Precision, $\mathcal{P}$, Recall, $\mathcal{R}$ and $\mathcal{F}$-measure given by

$$\begin{aligned}\mathcal{P} &= \frac{|\mathcal{TP}|}{|\mathcal{TP}|+|\mathcal{FP}|}\\ \mathcal{R} &= \frac{|\mathcal{TP}|}{|\mathcal{TP}|+|\mathcal{FN}|}\\ \mathcal{F} &= 2\times\frac{\mathcal{P}\times\mathcal{R}}{\mathcal{P}+\mathcal{R}}\end{aligned} \tag{2.3}$$

can be applied to both frame and onset detection based analyses. Other metrics have been proposed. One commonly used metric in AMT research papers is Accuracy :

$$\mathcal{A}=\frac{|\mathcal{TP}|}{|\mathcal{TP}|+|\mathcal{FP}|+|\mathcal{FN}|} \tag{2.4}$$

which was initially proposed by Dixon [33] and is often used in tandem with classification of individual errors.

While event-based analysis often considers onset detection only, metrics such as the Mean Overlap Ratio, $\mathcal{MOR}$, [111] which consider offset detection as well, have been developed. The $\mathcal{MOR}$ metric is most commonly used for the purpose of evaluation of note tracking AMT systems, which consider the temporal evolution of musical notes. $\mathcal{MOR}$ is defined as the mean of the overlap of all correctly detected notes, where the overlap of an individual note is given by

$$o_{\text{note}} = \frac{\max(t_{\text{on}}) - \min(t_{\text{off}})}{\max(t_{\text{off}}) - \min(t_{\text{on}})} \tag{2.5}$$

where t_{on} is a two-value set consisting of the transcribed and ground truth note onset times, while t_{off} similarly relates note offsets. Hence, o_{note} can be understood as the ratio between the temporal length of the intersection and the union of a transcribed note and its representation in the ground truth. More recently, the Sustain, Decay and mixed Sustain / Decay metrics have been proposed [45] which measure transcription by assigning scores to each note. The decay metric measures the pitch correctness and onset time accuracy, while Sustain performs a similar function to the $\mathcal{MOR}$ metric. These metrics are proposed to provide a holistic measure of AMT performance in terms of human perception, with psychoacoustic knowledge used in assignment of the scores [45].

2.1.4 Prior Research

A wide variety of methods have been proposed for AMT and multiple fundamental frequency estimation. Initial attempts, starting in the 1970s, focussed on the transcription of duets [84] [21]. Other early research in the field included methods using multi-agent blackboard systems [49] [82]. In the last decade, AMT has become a more popular research area and systems that can handle polyphonic signals have been developed, using a wide variety of signal processing and machine learning methods. A few of these approaches are outlined below. While a large body of recent literature concerning the AMT problem exists, it is intended here only to give a flavour of some of the recent methodology.

Possibly the best known method for AMT and polyphonic f_0 estimation is the iterative subtractive methodology of Klapuri. In this method, a pitch estimation step is performed at each time frame, and the dominant fundamental frequency is selected. The energy associated with that note parametrised by the dominant fundamental frequency is then subtracted from the frequency domain of the signal, before pitch estimation is again performed. This approach was first proposed in [68], where a spectral smoothness parameter was used to determine the energy in each harmonic partial associated with a selected fundamental frequency. This energy estimation step was performed before the subtractive process and aimed to counter the problem of overlapping harmonics. Further developments on this methodology employed note modelling [111] with Hidden Markov Models (HMMs) for tracking of note elements.

A recent contribution in AMT research that of Yeh et al [136], which is similar to other fundamental frequency methods, such as [68], in that an explicit pitch estimation step is involved. The method first uses a noise level estimation step, splitting the signal into a sinusoids plus noise representation. The sinusoidal peaks are expected to be harmonic partials of active sources, and pitch estimation is then performed. Unlike the iterative approach of Klapuri, this pitch estimation step leads to a set of potential f_0 candidates being extracted. A global search of combinations of these f_0 candidates is performed, with a score assigned to each set of candidates based on an aggregate of scores related to harmonicity, spectral smoothness and other spectral features. The tendency of fundamental frequency estimation systems to produce harmonic jumps is strongly considered in the development of features used to score the candidate sets.

Davy et al. propose a Bayesian harmonic model [31] for multi-pitch estimation in signal segments, which are assumed to have relatively homogeneous content. In a given segment a regular time lattice and continuous frequency spectrum are used. The Bayesian model includes parameters describing the number of notes, the number of harmonic partials in each note, the amplitudes of individual partials, noise variance and an inharmonicity parameter to model deviations from perfect harmonicity. Prior distributions are placed on many of these parameters and a Monte Carlo Markov Chain (MCMC) methodology is used to sample from the posterior distribution. The authors describe the importance of the order of the model, relating the number of notes, in achieving good performance, while the computational load associated with the MCMC methodology is noted. More recent work from the same research group [99] has focussed on matrix factorisations method using Expectation Maximisation (EM) based algorithms, incorporating prior information such as time continuity.

A discriminative method is proposed by Poliner and Ellis in [104] for polyphonic piano transcription. For each note on the piano scale, a one-versus-all Support Vector Machine classifier was trained using a large dataset of labelled piano pieces. A probability of activity is then assigned to each individual pitch-time point based on the distance from the frequency spectrum of the signal to the classifier boundary hyperplane. A pitch-time *posteriorgram* is constructed from these individual pitch-time probabilities. Hidden Markov Models (HMM) are used to track notes in the resulting posteriorgram. Experimental results show this method outperforming other well-known AMT methods, such as [111].

Bock and Schedl propose a pitched onset detection method using a recurring neural network [12]. Spectral features are formed using two-different scales STFTs in order to achieve good frequency and temporal resolution simultaneously. The STFTs are post-processed using semitone filterbanks to reduce the dimensionality of the problem. The first-order difference of the semitone filterbank output is taken and concatenated to the filterbank output itself. The data is labelled and presented to the neural network for training. Test data presented to the neural network system results are transformed to a regression matrix output, which is post-processed to produce a piano roll output. Experimental results given are considered state-of-the-art for pitched onset detection.

The use of genetic methods for AMT is proposed in [109]. The authors propose segmenting the spectrogram using an onset detector, and performing a parameter search using genetic methods to ascertain the signal elements present between each onset. A post-processing is performed to stitch together note elements that extend beyond the onset boundaries. Note template spectra are used, however the genetic method includes the ability to adapt the signal templates. Experimental results given show that the method outperforms many well known methods in terms of the combined Sustain/Decay score. However, the considerable computational expense of this approach is noted by the authors [109].

Spectrogram factorisation and decomposition methods are commonly used for AMT and musical signal processing in general. In this context, a spectrogram decomposition uses a fixed, pitch-labelled, dictionary and performs a regression on the spectrogram in order to ascertain the activations of notes, as represented by corresponding dictionary elements. Spectrogram factorisation methods seek to learn a dictionary and the activations of its atoms simultaneously. In this way spectrogram factorisations may require post-processing steps such as pitch estimation of the individual atoms. A description of some methods for matrix decomposition and factorisation is given in the following sections, after which some applications of these methods to AMT are described.

2.2 Sparse Representations

A sparse representation of a signal is characterised by a coefficient vector containing only a few non-zero elements. Sparse representations have been applied to problems in image and video processing such as de-noising, coding and compression [79], and have been seen to be particularly useful for audio processing [103]. The problem of retrieving a sparse representation for a signal is referred to as *sparse recovery* in the noiseless case, and *sparse approximation* in the presence of noise. Sparse recovery, formally, seeks to solve

$$\min_{x} ||\mathbf{x}||_{\mathbf{0}} \quad \text{s.t.} \quad \mathbf{s} = \mathbf{D}\mathbf{x} \tag{2.6}$$

while sparse approximation, the counterpart in noisy signals, seeks the minimisation

$$\min_{x} ||\mathbf{x}||_{0} \quad \text{s.t.} \quad ||\mathbf{s} - \mathbf{D}\mathbf{x}||_2^2 \leq \varepsilon \tag{2.7}$$

where $\mathbf{s} \in \mathbb{R}^M$ is the signal, $\mathbf{D} \in \mathbb{R}^{M \times N}$ is a dictionary with an atom of unit ℓ_2 norm in each column, $\mathbf{x} \in \mathbb{R}^N$ is the coefficient vector, ε is a noise component, and $||\mathbf{x}||_0$ is the ℓ_0 norm of $\mathbf{x}$ relating the number of non-zero components in $\mathbf{x}$.

The sparse representations methodology affords the use of *overcomplete* dictionaries, i.e. dictionaries where $N > M$, allowing the solution of underdetermined problems. This affords the formation of dictionaries consisting of, for example, a union of orthogonal bases [23], leading to more succinct representations of a signal containing elements that are best represented separately in disparate bases. Recovery of the sparsest solution using the ℓ_0 norm is known to be a NP-hard combinatorial problem in the general case [86] and a range of algorithms have been proposed which seek to approximate (2.6) and (2.7). One well-known approach to sparse representations is Basis Pursuit [23], also known as ℓ_1 minimization, which relaxes the ℓ_0 norm in (2.6) for a ℓ_1 norm. Similarly, for the noisy case the sparse approximation problem can be stated as

$$\min_{x} ||\mathbf{x}||_{1} \quad \text{s.t.} \quad ||\mathbf{s} - \mathbf{D}\mathbf{x}||_2^2 \leq \varepsilon \tag{2.8}$$

or, written in its Lagrangian form [79]:

$$\min_{x} ||\mathbf{s} - \mathbf{D}\mathbf{x}||_2^2 + \lambda ||\mathbf{x}||_1 \tag{2.9}$$

Algorithm 2.1 Orthogonal Matching Pursuit [98]

```
Input
  $\mathbf{D} \in \mathbb{R}^{M \times N}, \quad \mathbf{s} \in \mathbb{R}^{M}$
Initialise
  $i = 0; \quad \mathbf{r}^0 = \mathbf{s}; \quad \mathbf{x}^0 = 0; \quad \Gamma^0 = \{\}$
repeat
  $i = i + 1$
  Select atom and add to support
  $\hat{n} = \arg\max_n |\langle \mathbf{d}_n, \mathbf{r}^{i-1} \rangle|$
  $\Gamma^i = \Gamma^{i-1} \cup \hat{n}$
  Backproject supported atoms onto signal
  $\mathbf{x}^i = \mathbf{D}^{\dagger}_{\Gamma^i} \mathbf{s}$
  Calculate residual
  $\mathbf{r}^i = \mathbf{s} - \mathbf{D}_{\Gamma^i} \mathbf{x}^i$
until stopping condition met
Output $\mathbf{x}; \Gamma$
```

which is known as Basis Pursuit DeNoising (BPDN), or LASSO [124]. Both BP and BPDN can be solved using convex optimisation methods [23]. Other less computationally expensive algorithms such as Gradient Projections for Sparse Representations (GPSR) [42], Iterative Soft Thresholding (IST) [138] and stepwise methods such as (LARS) [127] and Polytope Faces Pursuit (PFP) [102] have also been proposed to solve these problems.

A faster alternative to Basis Pursuit is the use of greedy algorithms, which attempt to build up a representation by selecting the atoms most correlated with a residual in an iterative manner. Greedy methods are considered to approximate a ℓ_0 penalised least squares problem. The most well-known of these algorithms are Matching Pursuit (MP) [80] and Orthogonal Matching Pursuit (OMP) [98], which is outlined in Algorithm 2.1.

OMP is initialised by setting the residual signal, $\mathbf{r}$, equal to the initial signal, $\mathbf{s}$, and an iteration counter, i, is started. The algorithm then enters an iterative loop, calculating the inner products of the dictionary atoms with the residual signal $\mathbf{r}$ and selecting the atom, indexed by $\hat{n}$ with the largest magnitude inner product. This index, $\hat{n}$ is added to the sparse support, or set of active indices, Γ. The supported atoms are then backprojected onto the signal to get the interim coefficient vector, $\mathbf{x}^i$, from which the new residual can be calculated. This loop is repeated until a predetermined stopping condition is reached, typically a sparsity measure referred to as k-sparsity, where k is the amount of atoms to be selected. However, an energy-based threshold can also be employed, such as the residual norm, or relative error.

MP differs from OMP by not employing a backprojection step. Instead, the energy in the inner product of the selected atom is subtracted directly from the current residual:

$$\mathbf{r}^i = \mathbf{r}^{i-1} - \langle \mathbf{d}_{\hat{n}}^T \mathbf{r}^{i-1} \rangle \mathbf{d}_{\hat{n}}. \tag{2.10}$$

While the backprojection step used in OMP orthogonalises the residual to atoms that are already selected, reselection of atoms can occur using MP. It has been shown [23] that this may result in MP entering a critical loop. However, the backprojection step used in OMP may be costly compared to MP. Several variations such as Gradient Pursuits, which approximate the backprojection using a gradient descent [11], have been proposed to counter computation problems with large signals. Similarly LoCOMP [78], performs the orthogonalisation only on a subdictionary of atoms overlapping with the currently selected atom in a time-frequency dictionary. Various methods for performing the backprojection have been explored, such as rank-one Cholesky and QR updates [121]. Stagewise OMP (StOMP) [35] selects several atoms at each iteration, based on a coefficient threshold, in order to decrease the required number of backprojections.

Polytope Faces Pursuit (PFP) [102] is a greedy algorithm that seeks to optimise the ℓ_1 minimisation problem at each step, thereby performing Basis Pursuit. This is achieved using an iterative approach, similar to OMP, with the selection criteria:

$$\hat{n} = \arg\max_n \left| \frac{\mathbf{d}_n^T \mathbf{r}^{i-1}}{1 - \mathbf{d}_n^T \mathbf{c}^{k-1}} \right| \tag{2.11}$$

where $\mathbf{c}^{k-1} = [\mathbf{D}_{\Gamma^i}^{\dagger}]^T \mathbf{1}_{|\Gamma||}$.

The k-sparse problem relates to when it is known a priori that the sparse support consists of k atoms. While greedy methods can be used to solve this problem in k iterations, another family of algorithms such as CoSAMP [87], Iterative Hard Thresholding (IHT) [123], and Subspace Pursuit [29] attempt to solve this problem by initially selecting a subset of the dictionary, with size related to k. Swapping of atoms in and out of the active subdictionary is performed iteratively, while the residual is decreasing, until the algorithm converges to a fixed subset.

2.2.1 Recovery Conditions for Sparse Representations

There are theoretical conditions on the accuracy of sparse recovery for several methods, including OMP [128]. These conditions are related to the correlation between dictionary atoms, measured by dictionary coherence. Assuming unit norm atoms, the dictionary coherence μ is given by the absolute maximum inner product of dictionary atoms:

$$\mu = \max_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|. \tag{2.12}$$

Other coherence measures are also used, such as the cumulative coherence, also known as the babel function, $\mu(k)$ [128], where k is the number of atoms to be selected, given by:

$$\mu(k) = \max_{i} \max_{|\mathcal{J}|=k, i \notin \mathcal{J}} \sum_{j \in \mathcal{J}} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle| \tag{2.13}$$

from which it can be seen that $\mu = \mu(1)$, and $\mu(k) \leq k \times \mu$. An important condition for accurate recovery in sparse representations is the Exact Recovery Condition (ERC) [128] which states that recovery can be guaranteed when

$$||\mathbf{D}_0^{\dagger} \bar{\mathbf{D}}||_1 < 1 \tag{2.14}$$

where $\mathbf{D}_0$ is the subdictionary containing the correct sparse support, $\bar{\mathbf{D}}$ contains all other atoms in the dictionary and $||\mathbf{X}||_1$ is the matrix 1-norm, relating the maximum column sum of $\mathbf{X}$. It is shown [128] that the ERC is guaranteed to be met for OMP and BP when

$$k < \frac{1}{2}(\mu^{-1} + 1) \tag{2.15}$$

or similarly,

$$\mu(k) + \mu(k-1) < 1. \tag{2.16}$$

Recently, similar results have been shown for PFP [52] and ORMP [117].

Another set of conditions which can ensure guaranteed recovery for BP is the Restricted Isometry Property (RIP) [18], which relates the distance from orthogonality of a subdictionary:

$$(1 - \delta_k)||\mathbf{x}||_2^2 \leq ||\mathbf{D}_0 \mathbf{x}||_2^2 \leq (1 + \delta_k)||\mathbf{x}||_2^2 \tag{2.17}$$

where δ_k is the Restricted Isometry Constant (RIC), which measures the maximum deviation from unity of the eigenvalues of the subdictionary, $\mathbf{D}_0$ containing k atoms. It has been shown that Basis Pursuit [23] will recover the correct support with very high probability when $\delta_{2k} < \sqrt{2} - 1$ [18]. A simple relationship exists between the coherence and the RIC where $(k-1)\mu_D \geq \delta_k$, where μ_D is the coherence of the subdictionary, $\mathbf{D}_0$. The RIP is more commonly referred to in

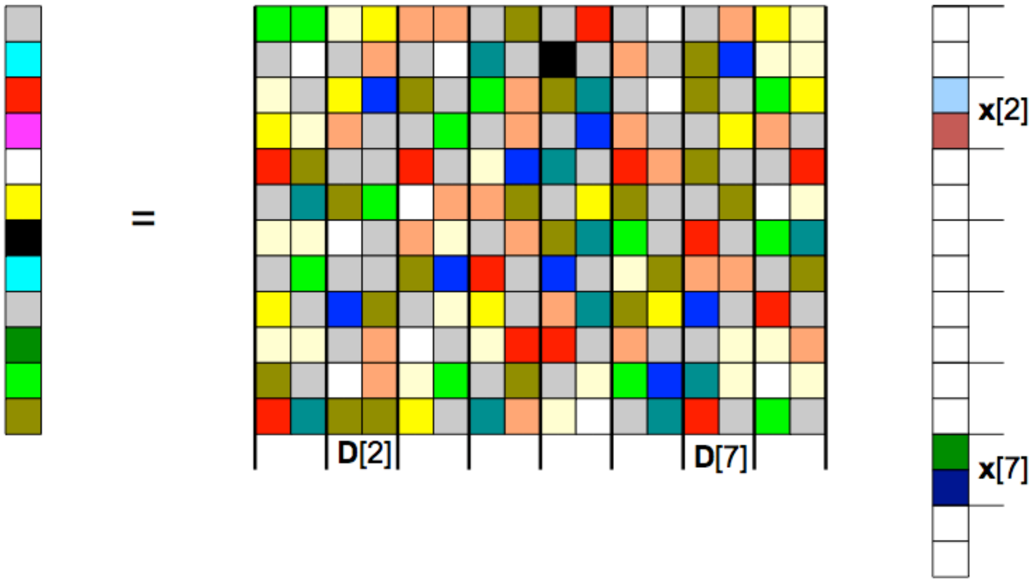

Figure 2.2: Graphical description of group sparsity. The dictionary, in centre, is partitioned into blocks that are two columns wide, and indexed correspondingly. $\mathbf{D}[2]$ refers to the second group, or block, comprised of the 3rd and 4th column atoms of the dictionary matrix. The activation vector, on right, is partitioned similarly, with two active blocks visible.

relation to Compressed Sensing, as it is known to hold with very high probability for various classes of random matrices.

2.2.2 Structured sparse representations

Structured sparse representations afford the introduction of prior knowledge to the sparse representation problem, through the implication that the activities of atoms tend to be interrelated. These interrelationships may be quite general, such as group sparsity [37] [137] where certain atoms tend to be active together in the same sparse vector, or simultaneous sparsity [130] where a similar atom tends to be active in different channels. Alternatively, more application specific forms of structured sparsity are possible where the interrelationship is based on expected outcomes, such as time continuity [30] or harmonic patterns [53].

Group sparsity, also referred to as block sparsity, extends the sparse representation framework by incorporating the assumption that certain atoms tend to be active together, as graphically described in Figure 2.2. Given the set of tuples

$$\mathcal{L} = \{\mathcal{L}^l\} \tag{2.18}$$

where $\mathcal{L}^l$ contains the column indices of the lth group leads to the notation for the lth group of the dictionary, $\mathbf{D}[l]$, and of the coefficient vector, $\mathbf{x}[l]$:

$$\begin{aligned} \mathbf{D}[l] &= [\mathbf{d}_{\mathcal{L}^l(1)}, ..., \mathbf{d}_{\mathcal{L}^l(|\mathcal{L}^l|)}] \\ \mathbf{x}[l] &= [x_{\mathcal{L}^l(1)}, ..., x_{\mathcal{L}^l(|\mathcal{L}^l|)}]^T \end{aligned} \tag{2.19}$$

where $\mathcal{L}^l(i)$ is the ith member of the lth tuple in the set $\mathcal{L}$, and $\sum_l |\mathcal{L}^l| = N$. The notation $\mathbf{x}[l,i]$ is used to refer to the ith member of the lth group of $\mathbf{x}$.

The group sparse problem is similar to the sparse problem and can also be solved using optimisation based methods, such as the Group Lasso [137] or L2-OPT [37], which seek to minimise a penalised least squares in a fashion similar to Basis Pursuit. Mixed vector or $\ell_{p,q}$ norms penalty terms given by

$$||\mathbf{x}||_{p,q} = \left(\sum_{l=1}^{L} \left(\sum_{i \in \{1,..,|\mathcal{L}^{(l)}|\}} \mathbf{x}[l,i]^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \tag{2.20}$$

are used. Similar to BP [23] the $\ell_{2,0}$ norm penalty is relaxed for an $\ell_{2,1}$ norm penalty which is seen to equal $||\mathbf{g}||_1$ where $g_l = ||\mathbf{x}[l]||_2$ for the $\ell_{2,1}$ norm (2.20). Using the $\ell_{2,1}$ norm assumes that few groups are active, but does not constrain the variety within groups. Other norms can be used, such as the $\ell_{1,2}$ norm, which conversely allows many groups to be active, but constrains the number of atoms that are active in each group. Similar to Basis Pursuit, different algorithms have been proposed in order to solve the group sparse optimisation problem, including a method based on Iterative Soft Thresholding [71] that was applied to audio denoising.

Greedy methods for group sparse recovery have been derived from Orthogonal Matching Pursuit (OMP) [98], differing only through using a group selection criteria, and adding all atoms in the selected group, indexed by $\hat{l}$, to the support: $\Gamma = \Gamma \cup \mathcal{L}^l$. The most well-known group sparse greedy method is the Block-OMP (B-OMP) [37] which uses the selection criteria

$$\hat{l} = \arg\max_l ||\phi[l]||_2 \tag{2.21}$$

where

$$\phi[l] = \mathbf{D}[l]^T \mathbf{r}^i. \tag{2.22}$$

An earlier proposed algorithm is the Subspace Matching Pursuit (SMP) [47], which uses the selection criteria.

$$\hat{l} = \arg\min_{l} \|\mathbf{r}^i - \pi_l(\mathbf{r}^i)\|_2 \tag{2.23}$$

where $\pi_l(\mathbf{y})$ is the projection of $\mathbf{y}$ onto the subspace $\mathbf{D}[l]$.

In a multi-sensor environment, several observations of the same source may be obtained, in different channels but in similar dimensions. It can be assumed that the signal is sparse in each channel at once, and the problem is referred to as simultaneous, or multichannel, sparsity. Simultaneous sparse approximation seeks to decompose all channels using a collection of a few atoms that are active across channels. This problem may be solved using optimisation based methods or greedy methods. The Simultaneous Orthogonal Matching Pursuit (S-OMP) [130] is presented as an algorithm for providing sparse solutions in the multichannel case, generalising OMP by changing the atom selection step to consider the ℓ_2 norm of the inner products of an atom with several channels :

$$\hat{n} = \max_{n} \|\mathbf{d}_n^T \mathbf{S}\|_2 \tag{2.24}$$

where $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_P]$ is the matrix containing the signal $\mathbf{s}_p$ from the pth sensor in each column. In [54], a similar algorithm is used and an average-case recovery analysis is proposed, which is shown to become more relevant as the number of channels increases. In [129] it is shown how global optimisation methods can also be used to solve the simultaneous sparse approximation problem using a $\ell_{2,1}$ mixed norm penalty term.

Several structured sparse methods have been proposed which exploit common structures found in audio signals. Harmonic Matching Pursuit [53], extends MP using a dictionary of harmonic atoms, composited from groups of Gabor atoms related by harmonicity. The coefficient of a harmonic atom is derived from the inner products of the individual Gabor atoms comprising the harmonic atom with the residual signal. The constituent Gabor atoms are selected from a small local peak search, which can allow for inharmonic tunings, and the peak and sidelobes are incorporated into the harmonic atom. Note detection and tracking are proposed as applications for this algorithm.

The Molecular Matching Pursuit (MMP) [30] is a greedy algorithm that extracts, at each iteration, a molecule consisting of a collection of structurally related atoms. MMP was proposed for the purpose of audio coding, with the aim of representing tonal elements in a Modified Discrete Cosine Transform (MDCT) and transient elements in a Discrete Wavelet Transform (DWT).

Structure is favoured in the atom selection stage by assigning locality based coefficients to each atom. From an initially selected atom, a molecule is grown based on connectivity criteria. When a tonal atom is initially selected, a search is performed, backwards and forwards through time frames along a narrow frequency window, until an energy threshold is reached. If a transient atom is selected, a wavelet tree is selected and pruned based on energy and connectivity criteria. All atoms found in the molecular search are added simultaneously to the sparse support which is reported to afford a speedup in a manner similar to StOMP [35].

2.3 Non-negative Representations

Many physical quantities and real-world data are inherently non-negative. In order to process such quantities, non-negativity often has to be explicitly considered. Non-Negative Least Squares (NNLS) is a well-studied constrained least squares problem:

$$\mathbf{x} = \arg\min_{x} ||\mathbf{s} - \mathbf{D}\mathbf{x}||_2^2 \quad s.t \quad \mathbf{x} \geq 0 \tag{2.25}$$

for which many algorithms have been proposed. The original NNLS algorithm [73], described in Algorithm 2.2 is a greedy active set algorithm. Indeed, it is similar to OMP (§2.2) with some important modifications. The atom selection is constrained to select atoms with non-negative inner products, while an inner loop is employed to eject atoms that have non-negative coefficients after backprojection of the active set onto the signal. NNLS stops iterating when no more atoms have a positive inner product with the residual, while OMP typically has a defined stopping condition.

The classic NNLS algorithm [73] is considered slow and many other algorithms have been proposed to solve the same problem. Fast-NNLS (F-NNLS) [14], proceeds similarly to the NNLS algorithm outlined above with some small modifications. The pseudoinverse of the active columns of the dictionary, $\mathbf{D}^{\dagger}$, is calculated at every iteration of NNLS, in order to perform the backprojection. In F-NNLS, the Gram matrix $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ and the projection vector $\alpha = \mathbf{D}^T\mathbf{s}$ are input. Then, instead of calculating the pseudoinverse, the inverse of a submatrix of the Gram matrix

$$\bar{\mathbf{G}} = \mathbf{G}_{\Gamma,\Gamma} = \mathbf{D}_{\Gamma}^T\mathbf{D}_{\Gamma} \tag{2.26}$$

containing the rows and columns indexed by the active set is inverted. The backprojection coef-

Algorithm 2.2 Non-Negative Least Squares (NNLS)

```
Input
  $\mathbf{D} \in \mathbb{R}^{M \times N}, \quad \mathbf{s} \in \mathbb{R}^M$
Initialise
  $\phi = \mathbf{D}^T \mathbf{s}; \quad \mathbf{x} = 0_N; \quad \Gamma = \{\};$
repeat
  $\hat{n} = \arg\max_n \phi_n$
  $\Gamma \leftarrow \Gamma \cup \hat{n}$
  $\mathbf{x}_\Gamma = \mathbf{D}_\Gamma^\dagger \mathbf{s}$
  Remove negative candidates
  While $\min(\mathbf{x}) < 0$
    $\bar{n} = \arg\min_n x_n$
    $\Gamma \leftarrow \Gamma \backslash \bar{n}; \quad x_n = 0$
    $\mathbf{x}_\Gamma = \mathbf{D}_\Gamma^\dagger \mathbf{s}$
  Endwhile
  $\mathbf{r} = \mathbf{s} - \mathbf{D}_{\Gamma^i} \mathbf{x}$
  $\phi = \mathbf{D}^T \mathbf{r}$
until $\max_n \phi_n \leq 0$
Output x
```

ficients are then calculated using

$$\mathbf{x}_\Gamma = \bar{\mathbf{G}}^{-1} \alpha_\Gamma. \tag{2.27}$$

Similarly the residual is not explicitly calculated and the vector of dictionary-residual inner products is approximated by

$$\phi = \alpha - \mathbf{G}\mathbf{x} \tag{2.28}$$

F-NNLS is seen to perform similarly to NNLS while being considerably faster. A variant of F-NNLS considers matrix decompositions, where many of the calculations may be similar and performed simultaneously. Another well known active set method is the block principal pivoting algorithm [105], which allows several atoms to be added or removed from the active set at each application. NNLS methods using descent-based approaches have been proposed. A coordinate descent based method is proposed in [46], while a Projected Quasi-Newton method is proposed in [61]. An overview of NNLS methods is given in [22].

Non-negative sparse approximations are performed using a Thresholded Non-Negative Least Squares (T-NNLS) algorithm in [114]. The authors show that the non-negativity constraint performs an innate regularisation, which may be more relevant than ℓ_1 minimisation for deriving non-negative sparse approximations. Experimental results are given showing, indeed, that T-NNLS outperforms ℓ_1 minimisation with a non-negativity constraint.

A non-negative variant of OMP (NN-OMP) is proposed in [16], in which the coherence problems inherent to non-negative sparse decompositions are noted. NN-OMP is seen to differ from OMP only by constraining the atom selection step to the maximum positive coefficient. It is suggested in [16] to use NNLS as the backprojection strategy, however it is more appropriate to consider NN-OMP as a truncated NNLS algorithm, using a defined stopping condition [100]. It is noteworthy that the normalisation of the dictionary generally used in sparse approximation is not necessary in NNLS, however it should be considered in the case of NN-OMP.

2.3.1 Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) [74] is a factor analysis algorithm that seeks to form an approximation of a non-negative matrix, $\mathbf{S} \in \mathbb{R}^{M \times T}$, such that

$$\mathbf{S} \approx \mathbf{D}\mathbf{X} \tag{2.29}$$

where both $\mathbf{D} \in \mathbb{R}^{M \times N}$, the dictionary matrix and $\mathbf{X} \in \mathbb{R}^{N \times T}$, the activation matrix are non-negative and unknown. Typically, the NMF problem is approached using an alternating projections methodology; i.e. alternating between updating the dictionary and updating the activation matrix, while the other is fixed.

The original NMF algorithm was referred to as Positive Matrix Factorisation (PMF) [97]. PMF used Alternating Non-negative Least Squares (ANLS) projections on $\mathbf{D}$ and $\mathbf{X}$ to seek the approximation (2.29) using a Euclidean distance cost function

$$\mathcal{C}_E(\mathbf{s}|\mathbf{z}) = ||\mathbf{s} - \mathbf{z}||_2^2 \quad s.t. \quad \mathbf{x} \geq 0. \tag{2.30}$$

where $\mathbf{z} = \mathbf{D}\mathbf{x}$. The ANLS projections can be slow, and optimised NNLS approaches have been proposed for use in ANLS-NMF [61] [64] [65] [57].

However, non-negative factor analysis was popularised as NMF [74] using fast multiplicative update gradient descent algorithms to solve the NMF problem using both the Euclidean distance (2.30) and Kullback-Leibler divergence

$$\mathcal{C}_{KL}(\mathbf{s}|\mathbf{z}) = \sum_i s_i \log \frac{s_i}{z_i} - s_i + z_i \tag{2.31}$$

cost functions. Again, alternating projections are used to update $\mathbf{D}$ and $\mathbf{X}$. The multiplicative updates for the Euclidean distance cost updates are given by

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes [\mathbf{D}^T\mathbf{S} \oslash \mathbf{D}^T\mathbf{D}\mathbf{X}] \tag{2.32}$$

$$\mathbf{D} \longleftarrow \mathbf{D} \otimes [\mathbf{D}\mathbf{S}^T \oslash \mathbf{X}\mathbf{S}\mathbf{S}^T] \tag{2.33}$$

where $\otimes$ denotes Hadamard or elementwise multiplication and $\oslash$ denotes elementwise division. It was demonstrated that the multiplicative updates were equivalent to the more common additive updates in gradient descent algorithms, with a fixed stepsize. It was also proved that the cost function was non-increasing in the multiplicative updates for both the Euclidean and KL multiplicative updates.

NMF algorithms using multiplicative updates have been proposed for other cost functions, the most common of which are generalised by the β divergence [27] given by

$$C_\beta(\mathbf{s}|\mathbf{z}) = \frac{1}{\beta(\beta-1)} \sum_i s_i^\beta + (\beta-1)z_i^\beta - \beta(s_i z_i^{\beta-1}). \tag{2.34}$$

The cost function given in (2.34) reduces to Euclidean distance when $\beta = 2$ and the KL-divergence and the Itakuro-Saito (IS)-divergence [40] are limiting cases given by $\beta = 1$ and $\beta = 0$ respectively. The multiplicative updates for the generalised β-divergence are given by

$$\mathbf{X} \leftarrow \mathbf{X} \otimes [\mathbf{D}^T(\mathbf{S} \otimes [\mathbf{D}\mathbf{X}]^{[\beta-2]})] \oslash [\mathbf{D}^T(\mathbf{D}\mathbf{X})^{[\beta-1]}] \tag{2.35}$$

$$\mathbf{D} \leftarrow \mathbf{D} \otimes [(\mathbf{S} \otimes (\mathbf{D}\mathbf{X})^{[\beta-2]})\mathbf{X}^T] \oslash [(\mathbf{D}\mathbf{X})^{[\beta-1]}\mathbf{X}^T] \tag{2.36}$$

where $\mathbf{X}^{[a]}$ denotes elementwise exponentiation of $\mathbf{X}$ to the power of a.

Bayesian formulations of NMF have also been proposed. Probabilistic Latent Component Analysis (P-LCA) [116] is a Bayesian NMF algorithm, allowing the use of appropriate Bayesian priors and which uses the KL-divergence NMF to update the probability distributions. The Itakuro-Saito (IS) divergence is shown in [40] to be similar to a maximum likelihood algorithm, and the authors also propose the integration of Bayesian priors, particularly in an Expectation Maximisation (EM) algorithm.

2.4 AMT using Spectrogram Decompositions

2.4.1 NMF-based approaches for AMT

NMF was originally proposed for AMT by Smaragdis and Brown [115], who show how a spectrogram, **S**, can be factorised into a dictionary, **D**, which is seen to contain atoms which represent note spectra, and a matrix, **X**, in which each row contains the activations of a corresponding atom. It is noted that the use of the KL-divergence cost function (2.31) provides a better factorisation than the Euclidean distance cost function (2.30) for the purpose of AMT. The authors suggest that each note might need to be played separately at least once in order to achieve a good separation. While the examples given in [115] are relatively simple, further work showed NMF to be a promising area of research for AMT [132] [8].

When NMF is used for AMT some common postprocessing steps are required [132]. The pitch of each atom must be determined. While this can be done by hand [1] [66], pitch estimation of single atoms is seen to be relatively straightforward and accurate [132] [8]. Once each atom is pitch labelled, a pitch-time representation can be formed by simply summing the energy contributions of all atoms of a given pitch [132]. Finally, thresholding of this pitch-time representation needs to be performed in order to to derive a piano roll [132] [8]. In [8], the effects of learning order on NMF are explored in the context of AMT, and it is found that initialisation of the dictionary with spectral templates that are similar to piano notes is optimal. Meanwhile, a simple example is given of perfect separation of notes, none of which are played in isolation. However, the authors note the ideal scenario used.

Several other related approaches exist. Abdallah and Plumbley propose learning a dictionary from power spectrograms using a non-negative sparse coding approach, in which a logarithmic prior is introduced to the atom activations. The dictionary is then pitch-labelled and used to decompose other musical spectrogram using the Itakuro-Saito divergence, again with a logarithmic sparse penalty applied. A sparse penalty and a time smoothness penalty, based on the total squared variance, augment the KL-divergence NMF for the purpose of monaural sound separation in [134].

Later research considers problems in unconstrained NMF when applied to AMT. In particular the tendency for meaningless atoms to be learnt is noted in [106]. Although NMF is considered to give a parts based representation, there are no guarantees that the atoms learnt are meaningful. In the context of AMT, a meaningful atom can be thought of as being related to only one note, with the atoms energy consisting of harmonic peaks from that note. In practice, atoms may be found consisting of energy from two or more notes, or alternatively having energy localised to one narrowband part of the spectrum, while atoms with unpitched features might also be learned. Hence, a harmonic version of NMF is proposed in [106]. In this approach, a dictionary atom is created for each note. The fundamental frequency and the expected harmonic peaks (2.1) are initialised with a fixed spectral envelope, while all other elements of the atom are set to zero. A feature of multiplicative update NMF is that zero elements are unaffected [4]. The crosstalk between the harmonic atoms is noted and some penalty terms are applied to the coefficient matrix. A sparse penalty is introduced, time continuity is encouraged by convolution with a smoothing matrix and a further penalty term on co-occurring correlated pitches are all augmented to the basic NMF updates, resulting in improved AMT [106].

A different approach to harmonic NMF is taken in [133], were it is proposed that each harmonic atom is parametrised by a collection of fixed narrowband harmonic components. A harmonic atom is learnt by finding the optimal coefficients for each narrowband element using β-NMF, with updates performed until convergence at each alternating projection. State of the art NMF results for AMT are given in [133] using this approach. The same authors also propose a Bayesian NMF approach in [9], using the same adaptive harmonic dictionaries [133].

An alternative to using the data-driven NMF approach to AMT is to use a fixed dictionary. This approach was referred to Non-negative Matrix Decomposition (NMD) in [32] and exploits the methodology of NMF, using only the coefficient matrix updates. Indeed, superior AMT results have been shown using NMD [133] [32], when the dictionary is learnt offline on a relevant dataset. A comparison of the use of β-NMD with Euclidean NMD and a sparse decomposition method [55] for the purpose of AMT is made in [32]. Superior AMT results are given for the β-divergence, with $\beta = 0.5$, for which a fast vector-based approach is proposed for the purpose of real-time approximation. In [133], NMD experiments are run using β-divergence with varying values of $\beta \in [0,2]$. Superior AMT results are again given for $\beta = 0.5$. In these experiments, degraded performance was noted when multiple atoms were used to represent one note. Furthermore, the authors show results for similar experiments with a dictionary that is not suited to the signals, with a large deprecation in performance.

An alternative perspective to spectrogram decompostion is proposed in [5] where decomposition is performed in a Bayesian setting, using Shift-Invariant P-LCA. The authors use shiftinvariance as a feature of the decompositions where differently pitched notes are represented by pitch-shifted versions of the same spectral templates. Each note is represented by three atoms, each representing a different state of the evolution of a piano note, namely attack, sustain, or decay with HMMs used to model the transitions from one state to another.

2.4.2 Spectrogram decompositions using greedy sparse methods

While NMF-based methods dominate the field of spectrogram decompositions, some researchers have explored the use of greedy methods for the same purpose. Harmonic Matching Pursuit (HMP) [53] was initially proposed for pitch tracking, a similar application to AMT, and can be seen to be a somewhat similar approach to the harmonic subtraction method of Klapuri [68] which it predates. However, the HMP algorithm suffers through not considering the problem of overlapping harmonic peaks which is addressed in the harmonic subtraction algorithm.

Matching pursuit methods have been considered on several occasions. In [20], the authors propose a variant of HMP that is augmented with spectral smoothness constraints to tackle the harmonic overlap problem. However, HMP is not used to decompose the spectrogram in this case [20]. Rather it is used to learn a dictionary of harmonic atoms. The extracted dictionary is then used with Matching Pursuit (MP) to perform spectrogram decompositions. This affords a data-driven approach based on sparse representations and results are given showing improved AMT results in comparison to NMF. However, the authors note the problem of spurious omissions when using MP, and propose a post-processing to fill gaps found in otherwise continuous signal elements [20]. An earlier work [76] uses a molecular variant of MP with a pre-learned harmonic dictionary, in which the atoms are labelled with pitch and instrument information. The molecular variant of MP used is similar to the tonal tracking used in MMP [30], but also allows for pitch variance, allowing notes from instruments with vibrato to be extracted in the one molecule.

In [125] the use of large dictionaries with OMP for AMT is considered, in which each note is represented with many atoms, each a datapoint from a spectrogram of an isolated note. The authors consider that the use of such dictionaries might allow more accurate modelling of signals. The computational complexity introduced by the large dictionary using OMP is noted and an Approximate OMP (AMP) algorithm using an approximate nearest neighbours search is proposed in [125]. AMP is seen to speed up the decomposition while a slight deterioration in AMT performance relative to OMP is noted. The problem of picking an apt stopping condition for OMP in the context of AMT is noted.

2.5 Discusssion

In this chapter, a summary of background research of relevance to this thesis was given. First the AMT problem was outlined, set in the context of the larger field of musical machine listening, and defined as a particular case of fundamental frequency estimation. A mention was given to metrics for AMT performance and some assumptions in use when AMT is applied to piano pieces were outlined before a brief summary of some prior AMT research was presented. Following this, sparse representations were introduced. Some popular algorithms for sparse approximation were described, before dictionary coherence and corresponding recovery conditions were introduced. Attention was then given to the concept of structured sparse representations. The next section focussed on non-negative representations, such as NNLS and NMF, before a return to the application of AMT, using non-negative and sparse decompositions. In later chapters more specific details of some of the prior research presented in this chapter are given.

The focus of this thesis is the development of methods encouraging sparsity and exploiting structure in the decomposition for the application of AMT. One possible initial observation might be that this could be a poor match of method to application. The sparse methodology was developed for the use of overcomplete dictionaries, and much of the research in this area assumes a union of orthogonal bases. For the AMT problem, an overcomplete dictionary is not necessary, although it may possibly be advantageous. The possibility of a mismatch becomes apparent if the high coherence in a dictionary of harmonic elements is considered in light of the recovery conditions given by the ERC (2.14). However, a direct application of sparse representations to AMT is not the goal. Rather, it is intended to draw on the knowledgebase of sparse representations in order to inform the AMT problem. Particular foci of this endeavour include the use of group sparsity, which affords a simple approach for dealing with subspaces. Dictionary coherence, while not affording the capability of guaranteed recovery in the setting of AMT, may nonetheless provide a parameter that is useful in this context. From the perspective of decomposition algorithms, greedy methods are popular for sparse approximation, and have occasionally been used, without great success, for the purpose of AMT. However, a new perspective is presented by considering algorithms that employ backtracking.

Before further description of the exploration of these ideas, the following chapter outlines the experimental setup employed throughout much of this thesis.

Chapter 3

Experimental Setup

While many different problems exist in Automatic Music Transcription (AMT) [6] the goal of this thesis is quite specific, with a focus on decomposition methods, incorporating sparsity and structure. In this regard most chapters consider different methods for performing a similar task of spectrum, or spectrogram, decompositions, and a similar homogeneity in experiments to be undertaken is suggested. A graphical description of a standard experimental workflow is shown in Figure 3.1, where the modular nature of the experiments undertaken can be observed. While different chapters propose different decomposition algorithms, the other steps in this experimental workflow vary little, while some common choices are available at each step.

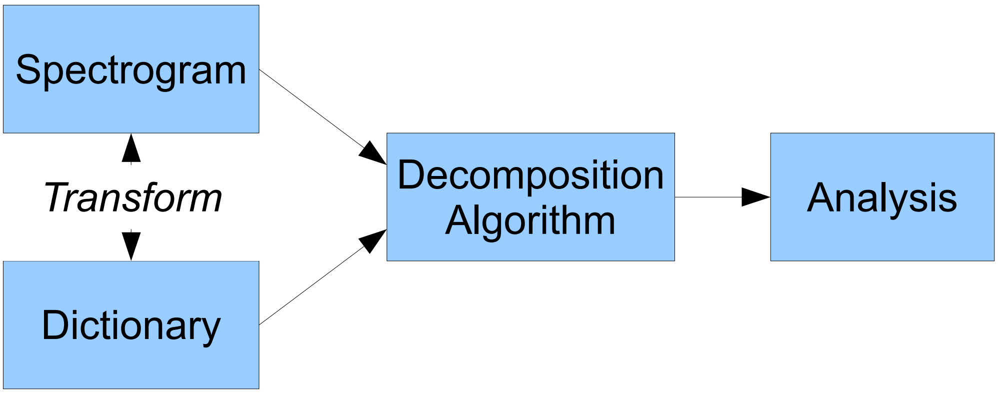

Figure 3.1: General experimental workflow

For instance, different transforms may be employed to form a spectrogram, while different types of dictionary may be used to perform the decomposition. It is also proposed to learn the dictionaries once, avoiding repetitive computation and affording direct comparison of various approaches. The analysis of the pitch-time representation may vary depending on the type of method used. The choices available in this modular experimental setup are laid out here in order to avoid repetition of description and simultaneous reference to various parts of the thesis in the description of an individual experiment. Much of the experimental setup is based upon that used in [133].

In the rest of this chapter, the dataset employed is introduced, before the various signal transforms employed are described. The dictionaries used are then derived, before post-processing steps employed in the analysis of performance are described. An example experiment, based upon reported state-of-the-art framewise spectrogram decompositions for AMT, is then performed to consolidate the workflow, and to set a benchmark for comparison of the approaches proposed in subsequent chapters.

3.1 Dataset

Many different datasets are used in the AMT community, a fact which in itself can lead to confusion as to the comparative efficacy of different methods [6]. Part of the reason for the wide range of datasets available is the large range of different AMT problems. For instance, multi-instrument AMT requires a different testbed to polyphonic piano transcription. In order to provide a direct comparison of the different approaches proposed in this thesis, one dataset is used throughout all AMT experiments in this thesis.

Polyphonic piano transcription is considered an appropriate application and a dataset taken from the MIDI-Aligned Piano Sounds (MAPS) database [39], was chosen for all AMT experiments. The MAPS database is a collection of digital audio files of piano sounds, including classical piano pieces, individual notes and chords. There are ten datasets in MAPS, eight of which are high quality synthetic Musical Instrument Digital Interface (MIDI) recordings. The remaining two datasets are recorded live on an electro-mechanical Disklavier piano, capable of playing automatically from a MIDI file input. The dataset used in this thesis is the *EnStDkcl* dataset, which is recorded with a microphone placed close to the piano body. The *EnStDkcl* contains 30 classical pieces of varying duration and complexity and the first 30*s* of each piece comprise the dataset employed in this thesis, referred to as the *standard dataset*. This provides a similar dataset to other research performed in AMT [133]. Indeed, much of the experimental setup outlined here is derived from that work.

Several reasons led to the choice of this dataset. As well as providing a challenging, varied set of piano pieces, the dataset is recorded live, thereby presenting a realistic challenge to an AMT system, for instance introducing room-specific echoic effects. An advantage of this dataset is the fact that due to the electromechanical nature of the recordings, a ground truth is provided, thereby avoiding the task, and related vagaries, of hand-labelling the notes in a piece and score to pianoroll alignment. Using the onset and offset times given for each note, a ground truth piano roll is derived in order to analyse performance. The ground truth piano roll is a grid with 23*ms* temporal delineations and 88 different pitch frames, representing MIDI notes #21-#108. All pitch-time points containing an onset or offset of a note, as defined from the supplied ground truth, are designated as active. Similarly pitched points intermediate to the onset and offset of a note are also set as active.

3.2 Spectrogram Transforms

Different signal transforms may be used to produce a spectrogram. The most commonly used spectrogram is the Short-Time Fourier Transform (STFT), which can be calculated very quickly by windowing a signal into frames of a desired sample size which are individually processed using the Fast Fourier Transform. However the STFT is known to suffer from several problems in relation to musical signal processing. In particular the time and frequency resolutions are related which can be problematic in dynamic signals. In musical signals the fundamental frequency scale is logarithmic (§2.1.2). A reasonable time resolution, in terms of human perception, can lead to the fundamental frequencies of neighbouring lower pitched notes sharing the same frequency bin in the spectrum. Overlapping signal windows are often used to afford greater frequency resolution in the STFT. Alternatively, other transforms with logarithmic frequency scales have been proposed for use in musical signal processing. Two commonly used logarithimic scale transforms are the Equivalent Rectangular Bandwidth Transform (ERBT) [132] and the Constant-Q Transform (CQT) [113], both of which are implemented using filterbanks. While the STFT partitions the time frequency domain in a regular manner, with all time-frequency points of identical shape, the ERBT and CQT use irregularly-sized frequency partitions. The CQT places an equal

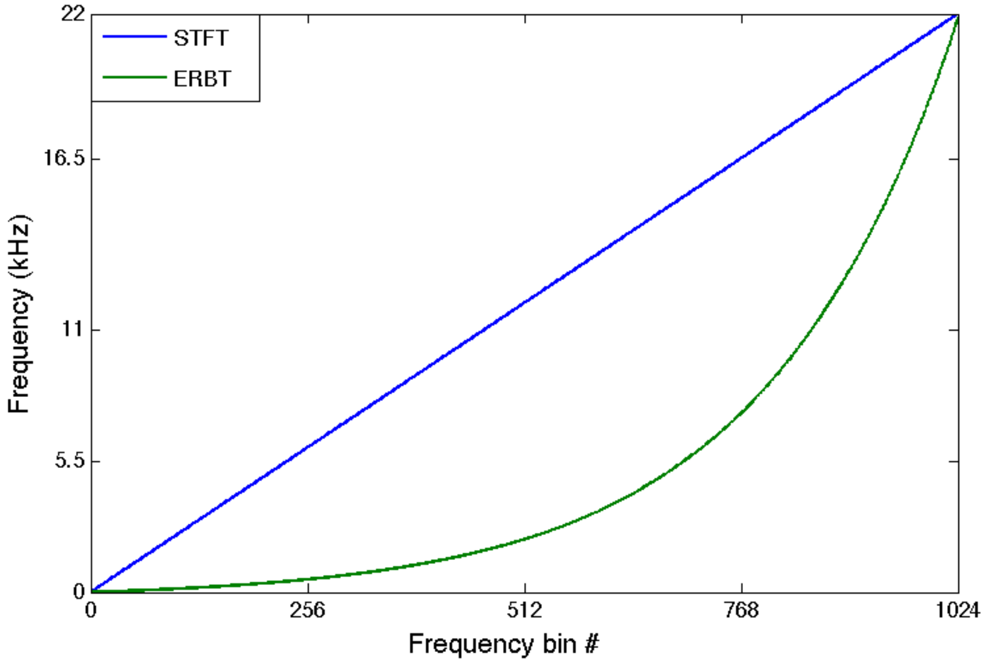

Figure 3.2: Central frequency of each ERB dimension relative to STFT. Both transforms are sampled at $44.1kHz$ and have similar atom dimension 1024

amount of frequency partitions in each octave, while the ERBT places the frequency bins linearly on the ERB scale given by [132]:

$$\nu_f^{ERB} = 9.26 \times \log(0.00437\nu_f^{Hz}) \tag{3.1}$$

where ν_f^{ERB} is the frequency in terms of ERBs and ν_f^{Hz} is the frequency as measured conventionally in Hertz (Hz). Both the ERB and CQT are known to introduce some distortion into the spectrogram at low-frequencies. The higher resolution at low frequencies requires the use of longer signal windows than is needed at higher frequencies, and both the CQT and ERB are often interpolated onto a rectangular matrix [113] [132] for convenient post-processing. The relationship between the linear frequency scale used in the STFT and the log scale used in the ERB can be observed in Figure 3.2.

The ERBT and STFT were compared for the purpose of AMT in [132]. Signals with sampling frequency $22.05kHz$ were transformed using an ERBT with atom dimension 250 interpolated onto a $23ms$ grid. This ERBT was designed so that no two fundamental frequencies shared the same frequency frame [132]. A comparison was made with an STFT of atom dimension

Name	Transform	$f_S(kHz)$	$Window(ms)$	$Overlap$	Dimension
$\mathbb{S}1$	STFT	22.05	92	75 %	1024
$\mathbb{S}2$	STFT	44.1	92	75 %	2048
$\mathbb{E}1$	ERBT	22.05	23	-	250
$\mathbb{E}2$	ERBT	22.05	23	-	512
$\mathbb{E}3$	ERBT	44.1	23	-	512
$\mathbb{E}4$	ERBT	44.1	23	-	1024

Table 3.1: Transforms used in experiments - ERB transforms are interpolated onto a grid

1024, requiring 92*ms* windows and using a 75% window overlap, giving a similar temporal resolution as the ERBT. Similar results were recorded for both transforms, while it was noted that the ERBT afforded faster processing due to its smaller dimension.

To further this exploration, a comparison of several different dimension ERBTs as well as two different scale STFTs is proposed, with some experiments in this thesis comparing all six transforms. In particular, larger dimension transforms are used, and it hoped that doing so might result in less coherent dictionaries due to the spread of energy. A 23*ms* time partitioning of the spectrogram is used in all cases, and two different sampling frequencies are considered. The higher sampling frequency, $44.1kHz$, is the same rate that is used on compact discs and covers the range of frequencies accessed by normal human hearing. The lower sampling rate is $22.05kHz$, the same as that used in [132] and simply effected on the MAPS dataset by downsampling using the *resample* function in Matab, which applies a low pass anti-aliasing filter. The details of the transforms employed are shown in Table 3.1. Each transform is coded with an alphabetic and a numeric character. The alphabetic character relates the type of transform used while the numeric character relates a transform specific ordering in terms of sampling frequency and atom dimension.

3.3 Dictionaries

To perform AMT in a supervised manner, a spectrogram is decomposed with a dictionary of pitch-labelled atoms that represent notes in the corresponding transform. Different types of dictionary can be used, the choice of which may be related to the method. Greedy methods are flexible and can be used with both large and small dictionaries. NMD-based methods have previously been observed to perform optimally when each note is represented by one atom [133]. Audio files containing isolated notes signals are available in the MAPS database and are used to derive the dictionaries proposed here. The isolated note signals that are used are from the

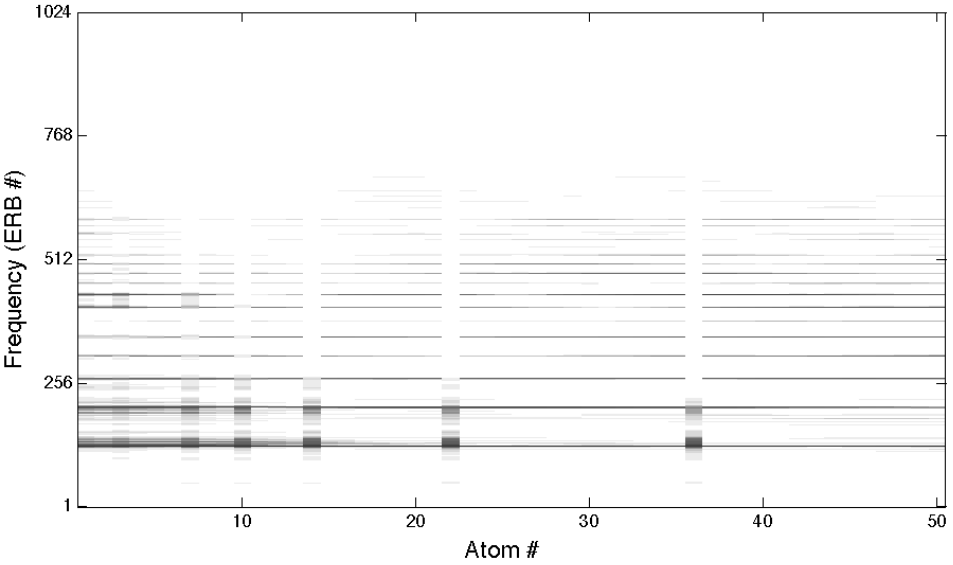

Figure 3.3: 50 atoms from *datapoint dictionary* in ERB (transform $\mathbb{E}4$) used to represent one note. Echoic artefects visible

EnStDkcl dataset, similar to the dataset. Several different types of fixed dictionary are used in the experiments in this thesis. To form each of these types of dictionary, a subdictionary is first obtained for each note of the piano scale, 88 notes in all, from MIDI notes #21 to #108. The atoms of each subdictionary are pitch-labelled according to the isolated note from which they are derived and the dictionary is constructed through concatenation of these subdictionaries. This dictionary construction is performed for all transforms listed in Table 3.1.

3.3.1 Datapoint Dictionaries

A stated advantage of greedy methods is the ability to easily handle overcomplete dictionaries (§2.2). For example, Tjoa et al [125] propose using OMP with an overcomplete dictionary, comprised of datapoints from spectrograms of isolated notes, to perform AMT. A similar dictionary is constructed here using a process akin to that employed in [125] . To construct each pitched subdictionary, a spectrogram of an isolated note file was produced, and onset detection was performed to capture the start of the note. The 50 time frames of the spectrogram following the onset were individually normalised to unit ℓ_2 norm and collated to form the subdictionary representing the given note. Concatenation of the individual subdictionaries was performed to construct the dictionary, as previously mentioned. These dictionaries are referred to as the *datapoint dictionaries*, and an example subdictionary representing one note is shown in Figure 3.3.

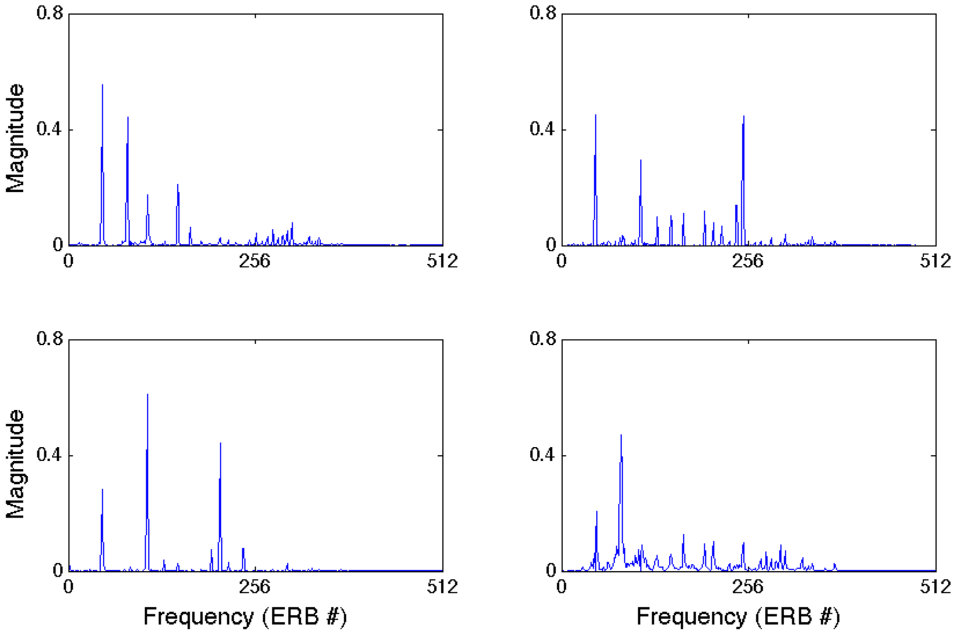

Figure 3.4: Group of atoms from *subspace dictionary* (P=4) used to represent a note

3.3.2 Subspace and Atomic Pitch Dictionaries

Usually a dictionary much smaller than the datapoint dictionary described above is used for spectrogram decompositions [32] [133]. While a dictionary can be learnt from signals containing musical pieces with hand-labelling of pitches [1], it is also possible to learn each subdictionary from isolated notes, an approach taken by [32] [133] and adapted here.

A spectrogram of an isolated note signal is formed using the desired transform. Non-negative Matrix Factorisation (NMF) [74] is used to perform a factorisation of predetermined rank, P, of the spectrogram. The resultant subspace, of size P, then forms the subdictionary corresponding to the note played in the isolated signal. Again, the dictionaries are formed by concatenation of the individual subspaces and each atom is normalised to have to unit ℓ_2 norm.

Dictionaries were learnt for all transforms in Table 3.1 for all values of $P \in \{1,...,7\}$. These dictionaries are referred as *subspace dictionaries* collectively for $P \in \{1,...,7\}$. The term *atomic pitch dictionary* is used to refer to a dictionary with one atom used to represent each note, also considered a subspace dictionary with $P = 1$. An example subspace of size $P = 4$ is shown in Figure 3.4.

3.4 Analysis of AMT

The aim of this thesis is to enhance the performance of spectrogram decomposition-based AMT. Spectrograms are decomposed on a frame-by-frame basis in many of the proposed approaches. Therefore a frame-based analysis is considered most appropriate to measure the performance of the proposed methods, although some onset-based analysis is also described in Chapter 6. A plethora of measures have been proposed for AMT, as described in (§2.1.3). However as it is intended mostly to measure the effectiveness of decomposition methods, the $\mathcal{F}$-measure is employed as the primary performance metric, with occasional auxiliary use of the Precision, $\mathcal{P}$, and Recall, $\mathcal{R}$, (2.3) metrics commonly used in pattern recognition. These metrics, when used here, are expressed as percentages.

The Accuracy, or $\mathcal{A}$-measure (2.4) is commonly used in AMT, which, similar to $\mathcal{F}$-measure, incorporates both false detections and true omissions. It is worth noting that the $\mathcal{F}$-measure (2.3) can, through substitution of the definitions of the Recall and Precision (2.3), be written in similar form to the Accuracy measure:

$$\mathcal{F} = \frac{2|\mathcal{TP}|}{2|\mathcal{TP}| + |\mathcal{FP}| + |\mathcal{FN}|} \tag{3.2}$$

differing only from the Accuracy metric through replacement of $|\mathcal{TP}|$ with $2|\mathcal{TP}|$. Whether the $\mathcal{F}$- or $\mathcal{A}$-measure is more appropriate for the purpose of AMT is a philosophical point. The $\mathcal{F}$-measure is preferred here, as in the case when the polyphony is known the $\mathcal{A}$-metric will penalise a false detection twice. For example, in the known polyphony case, if 80% of detections are correct, leading to 20% false detections / true omissions, an Accuracy of 67% is recorded, while an $\mathcal{F}$-measure of 80% is given. Two different post-processing steps are commonly used prior to calculation of the metrics. These post-processing steps are referred to here as δ-thresholding and k-sparse thresholding.

3.4.1 δ-thresholding

Often in the context of AMT thresholding is performed as a post-processing step to a decomposition method such as NMF (§2.3.1) or NNLS (§2.3) in order to determine the final output piano roll [132] [133] [32]. The thresholding setup used in [133] is adapted here, whereby a parameter,

δ, is used to adapt the threshold λ to the data:

$$\lambda = \delta \times \max_{l,t}[H]_{l,t}. \tag{3.3}$$

where δ is typically related in decibels (dB) and $\mathbf{H}$ is a (group) coefficient matrix. For each individual piece, the size of the sets, $\mathcal{TP}, \mathcal{FP}, \mathcal{FN}$, are recorded for a range of values of δ. Summation of the cardinalities of these sets, across all pieces, is performed for each value of δ, allowing the $\mathcal{F}$-measure to be calculated similarly. Results given relate the optimal $\mathcal{F}$-measure achieved, recorded at δ_{opt}. When values for Recall and Precision are given, these are also the values recorded at δ_{opt}. A graphical example of this procedure is shown in Figure 3.5, later in this chapter.

3.4.2 *k*-sparse thresholding

Occasionally in AMT it is assumed that the polyphony at a given time frame of the spectrogram is known, in which case the problem can be conceived of as a k-sparse problem (§2.2). Some algorithms, such as OMP can be initialised with a k-sparse stopping condition, selecting only k atoms. However other algorithms, such as NNLS or NMF/NMD, perform a global optimisation. While such approaches are typically analysed using a global threshold, such as the δ-thesholding strategy employed above, it may be useful occasionally to compare these different classes of algorithms. In order to do this a k-thresholding is performed at each time frame of the coefficient matrix from a global optimisation based method to identify the k-sparse support Γ such that

$$\Gamma_i = \begin{cases} 1, & \text{if } i \in \mathcal{J}^{(1:k)} \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

where $\mathcal{J} = \{j | h_j > h_{j+1}\}$ is the set of the indices of ordered values of a vector $\mathbf{h}$ and $\mathcal{J}^{(1:k)}$ is a subset of $\mathcal{J}$ consisting of the first k elements. When k-sparse analysis is used, the results are described in terms of $\mathcal{F}$-measure. Due to the known sparsity level the number of false positives and false negatives are equal and $\mathcal{P} = \mathcal{R} = \mathcal{F} = |\mathcal{TP}|/(|\mathcal{TP}| + |\mathcal{FP}|)$.

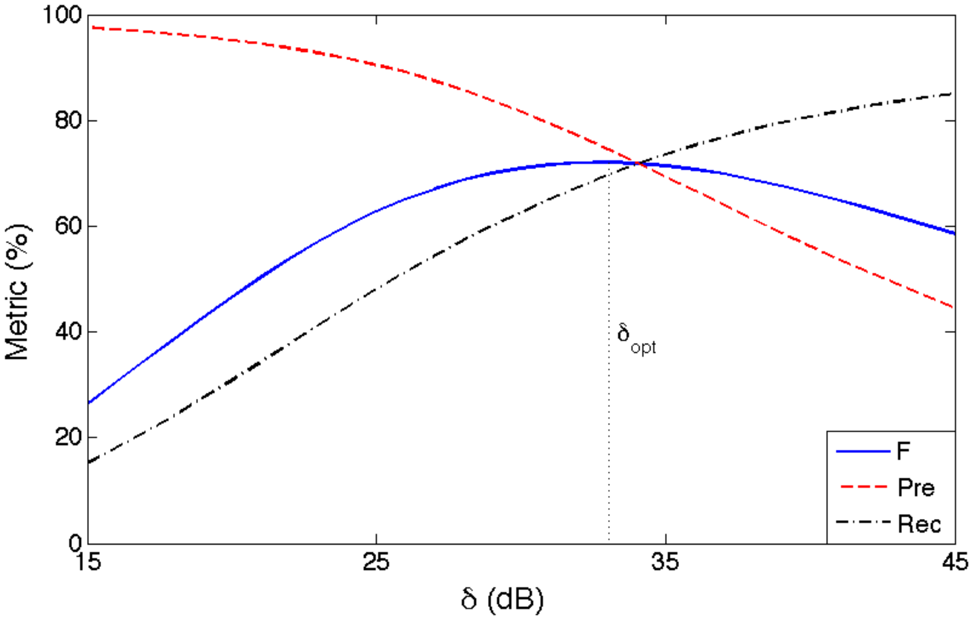

Figure 3.5: $\mathcal{P},\mathcal{R},\mathcal{F}$ metrics relative to δ threshold parameter for benchmark experiment. δ_{opt} indicated.

3.5 Example experiment

An example experiment is performed to provide a benchmark. Spectrograms using Transform $\mathbb{E}1$ (§3.2), are decomposed with a atomic pitched dictionary (§3.3.2) for all pieces of the standard dataset (§3.1). δ-thresholding (§3.4.1) was performed for a range of values of $\delta \in \{15,...,40\}$dB, in steps of 1dB. In this way a similar dataset, dictionary and thresholding strategy are employed in this set of experiments to those used in the NMD experiments described in [133]. The results for these experiments are tabulated in Table 3.2 and a graphical description of the evolution of the $\mathcal{P},\mathcal{R},\mathcal{F}$ metrics relative to δ is given in Figure 3.5 with δ_{opt} outlined.

It should be noted that the dictionary used in [133] was learnt from isolated note signals in the RWC [51] database. The best results presented in [133] give an $\mathcal{F}$-measure of 67%. The results presented here use a similar transform and a $\mathcal{F}$-measure of 72% is achieved, an increase of 5%. This can be explained as a result of using a dictionary learnt from isolated note signals recorded in similar conditions to the pieces. This result, with $\mathcal{F}$-measure of 72% is

Transform	P	$\mathcal{F}$	$\delta_{opt}(dB)$	$\mathcal{P}$	$\mathcal{R}$
$\mathbb{E}1$	1	72.0	33	73.0	71.0

Table 3.2: Results from benchmark experiment in terms of $\mathcal{P},\mathcal{R},\mathcal{F}$. δ_{opt} noted.

considered as a benchmark for the researched presented here, as results using this experimental setup, performing NMD with β-divergence with $\beta = 0.5$ with ERB of dimension 250 has recently been considered state-of-the-art for supervised decomposition based AMT.

3.6 Other details

All code was written in Matlab version 10.1. and experiments were run on an Apple Mac using the OSX 10.6 operating system with $8Gb$ of RAM running at $1067MHz$, with an Intel Core 2 Duo processor clocking $3.06GHz$.

3.6.1 Reproducible Research

The level of reproducibility of much research in the signal processing community is underwhelming. As the methodologies used in the field become more complex, ambiguity in written descriptions may arise, leading to difficulty in building upon established published literature. In light of this, efforts should be undertaken to provide code which makes results reproducible, and allows competing methodologies to be easily and fairly compared. The research described in this thesis has benefited greatly from the availability of code and datasets made possible through the endeavours of other researchers. Hoping both to acknowledge the benefit of this open availability to this research, and to promote the reproducible research philosophy, the code used to perform all experiments will become available at *www.soundsoftware.ac.uk/ken/thesis*.

Chapter 4

Matching Pursuits

Spectrogram decompositions are often used for the application of Automatic Music Transcription (AMT) and other musical signal processing tasks, and are usually performed using a gradient descent methodology. In particular, a magnitude spectrogram is usually decomposed using multiplicative update algorithms based on Non-negative Matrix Factorisation [115] [133] [32]. Greedy sparse Matching Pursuit [80] algorithms provide an alternative strategy to perform matrix decompositions and may be attractive as they are fast and allow explicit control of the sparsity level. Another advantage of greedy methods is that they allow the use of an overcomplete dictionary, which may afford greater modelling capabilities than the single atom per note model commonly used in supervised Non-negative Matrix Decompositions (NMD) [133] [32]. The most well-known greedy pursuits are Matching Pursuit (MP) [80] and Orthogonal Matching Pursuit (OMP) [98], outlined in (§2.2).

The use of greedy methods has previously been proposed on a few occasions for the application of AMT. Cariabas-Orti et al [20] use a variation of Harmonic Matching Pursuit (HMP) [53] with an extra constraint placed on the smoothness of the spectral envelope to learn a dictionary of harmonic atoms. After the dictionary is derived, a spectrogram decomposition is then performed using MP. The authors note the problem of correctly detected note events being represented in a fractured manner, due to spurious pitch-time omissions.

Tjoa et al [125] propose to use large datapoint dictionaries, similar to those outlined in (§3.3.1), for AMT, using OMP to perform spectrogram decompositions. The authors note the resultant computational expense associated with the large dictionary-residual multiplications at each iteration of OMP. An Approximate OMP (AMP) algorithm is proposed to counter this expense, using an approximate nearest neighbour search based on Locality Sensitive Hashing. The AMP algorithm is seen to be faster than OMP, while incurring a small relative deficit in performance. The authors note the difficulty in selecting apt stopping conditions for OMP-based methods in the context of AMT.

One recent avenue of research in the field of sparse representations is the concept of group, or block, sparsity, which assumes that certain groups of dictionary atoms tend to be correlated in their activity in a given coefficient vector (§2.2.2). Greedy methods for the group sparse problem have been proposed, such as Block-OMP (BOMP) [37] and Subspace Matching Pursuit (SMP) [47]. BOMP and SMP are based upon the OMP algorithm, outlined in (§2.2), with two important alterations to incorporate the group structure. The selection criteria consider all atoms in the group. For BOMP the ℓ_2-norm of the coefficients belonging to a group is used :

$$\hat{l} = \arg\max_{l} ||\phi[l]||_2 \tag{4.1}$$

where $\Phi = \mathbf{D}^T\mathbf{r}$ is the inner product of the dictionary and the residual, and $[l]$ denotes the lth group using the group notation defined in (§2.2.2). SMP considers an orthogonal projection of the current residual signal onto a given subspace

$$\hat{l} = \arg\min_{l} ||\mathbf{r} - \pi_l(\mathbf{r})||_2^2 \tag{4.2}$$

where $\pi_l(\mathbf{r})$ is the projection of the residual onto the subspace represented by the lth block of the dictionary. In both BOMP and SMP, all atoms in a newly selected group are added to the sparse support simultaneously.

The use of group sparsity has not been previously explored in the context of AMT. It has previously been shown that the use of several atoms to represent one note may result in better AMT performance [1], due to improved modelling ability. Conversely, it was observed in [133] that the use of more than one atom to represent a note had a negative effect on AMT performance. It is one of the goals of this thesis to explore the use of group sparsity with subspace-based models for each note and this chapter begins that exploration, using greedy methods.

In the remainder of this chapter, non-negative variants of group sparse greedy algorithms are first proposed. Experiments are described that compare subspace modelling, which employs

Algorithm 4.1 Non-negative Group OMP-based algorithms

```
Input
  $\mathbf{D} \in \mathbb{R}^{M \times N}$,   $\mathbf{s} \in \mathbb{R}^{M}$,   $\mathcal{L}$
Initialise
  $\Gamma = \{\}$;   $\mathbf{r}^0 = \mathbf{s}$;   $i = 0$
repeat
  $i = i + 1$
  Select $\hat{l}$ using selection criteria (4.3), (4.4), (4.5) or (4.6)
  $\Gamma = \Gamma \cup \mathcal{L}^{\hat{l}}$
  Back project support onto signal
    $\mathbf{x}_{\Gamma_i} = \arg\min_x \|\mathbf{s} - \mathbf{D}_{\Gamma_i}\mathbf{x}\|_2^2 \quad s.t. \quad \mathbf{x} \geq 0$
  Update residual
    $\mathbf{r}^i = \mathbf{s} - \mathbf{D}_{\Gamma_i}\mathbf{x}_{\Gamma_i}$
until stopping condition met
```

these group sparse methods, with datapoint modelling and single atom per pitch modelling using standard unstructured greedy methods. Some computational problems using greedy group methods are observed in the non-negative framework, and strategies to counter these issues are proposed. In particular, a non-negative variant of the SMP [47] is seen to be computationally demanding and a fast variant of this algorithm is proposed. Furthermore the use of a gradient step for the backprojection is explored in this non-negative structured framework. Finally the chapter concludes with a discussion, summarising the findings of the chapter.

4.1 Greedy Non-Negative Group Sparsity

Subspace dictionaries, described in (§3.3) are formed from a union of pitch-labelled subspaces, each learnt from the spectrogram of an isolated note using Non-negative Matrix Factorisation (NMF) [74]. The factorisation used to learn individual subspaces also produces an activation matrix in which each row displays the activations for a corresponding atom. Co-activation of atoms is generally observed in individual columns of the activation matrix of this factorisation, although not all atoms are necessarily active at each time frame, particularly when the size of the subpace $P \geq 3$. An implication of this observed co-activation is that individual atoms cannot be expected to form a good model of a given note spectrum, unlike atoms from the datapoint or single atom dictionaries. Indeed, their modelling capability lies in their interdependency, and it is essential to consider some grouping strategy when they are used to decompose a spectrogram, providing the rationale for the use of group sparsity.

As magnitude spectrograms are considered and the dictionaries consist of non-negative atoms, group sparsity is considered in a non-negative framework, where $\mathbf{s}, \mathbf{D}, \mathbf{x} \geq 0$, and variants of group sparse pursuits that incorporate the non-negative constraint are proposed. An outline of a general non-negative group sparse OMP-based algorithm is given in Algorithm 4.1. Similar to the NN-OMP [16], the use of NNLS as the backprojection step is used for the general non-negative group OMP. Different group sparse algorithms can be effected simply by using different group selection criteria.

A non-negative variant of the BOMP selection criteria (2.21) [37] referred to as NN-BOMP is proposed by considering only the positive inner product coefficients:

$$\hat{l} = \arg\max_{l} ||\phi^{+}[l]||_2 \tag{4.3}$$

where ϕ =D$^T\mathbf{r}$ and $\phi^{+} = \mathcal{I}\phi$ where $\mathcal{I}$ is a binary elementwise "is positive" indicator function. The use of the ℓ_∞ and ℓ_1 norms as group selection criteria for greedy methods may be appropriate for some applications and non-negative versions of these selection criteria are proposed, being easily derived in a similar manner to (4.3) :

$$\hat{l} = \arg\max_{l} ||\phi^{+}[l]||_1 \tag{4.4}$$

for the ℓ_1-norm, or alternatively for the ℓ_∞ norm:

$$\hat{l} = \arg\max_{l} ||\phi^{+}[l]||_\infty \tag{4.5}$$

which are referred to as NN-ℓ_1-BOMP and NN-ℓ_∞-BOMP respectively. A non-negative variant of the SMP algorithm is proposed using the selection criteria :

$$\hat{l} = \arg\min_{l} ||\mathbf{r} - \mathbf{D}[l]\hat{\mathbf{x}}[l]||_2 \tag{4.6}$$

where $\hat{\mathbf{x}}[l]$ is the NNLS solution (2.25) to $\mathbf{r} \approx \mathbf{D}[l]\mathbf{x}[l]$. This selection criteria (4.6) is referred to as Non-Negative Nearest Subspace OMP (NN-NS-OMP) [91], although it could also be referred to as Non-negative SMP. The difference in how the selection criteria for NN-NS-OMP and SMP are written is notable, with SMP using the projection operator, $\pi_l(\mathbf{r})$ onto the subspace $\mathbf{D}[l]$. As some atoms in a group may not be active the projection operator is not explicitly available in the non-negative case. Hence, NNLS is required in order to ascertain the active atoms in each group so that the selection criteria (4.6) may be calculated.

Experiments

Experiments were performed to compare the use of the different modelling approaches for AMT using greedy algorithms. As the methods employed are greedy in both cases, a comparison can be performed by using a k-sparse stopping condition. The problem in picking a good stopping condition for greedy methods when used for AMT has previously been noted [125] and the k-sparse stopping condition avoids such complications when a comparison is required. Experiments were run on the complete standard dataset (§3.1). Separate experiments were run for each signal transform described in Table 3.1 in order to compare their performance in an AMT task.

For the subspace modelling, all non-negative group sparse algorithms are used by employing the selection criteria (4.3)-(4.6) in the general non-negative group sparse OMP described in Algorithm 4.1. Experiments were run for all groupsizes, $P \in \{1, ..., 7\}$, noting that all group algorithms default to NN-OMP when $P = 1$. In the case of the group sparse algorithms, the k-sparse stopping condition is effected when k different groups have been selected.

For the datapoint modelling, NN-OMP and a non-negative variant of MP (NN-MP) are used to perform the decompositions. NN-MP differs from standard MP only by constraining the selection step to consider only atoms displaying positive inner products with the residual. Using the datapoint dictionaries, the greedy algorithms are faced with some difficulty in relation to the k-sparse stopping condition, as many atoms labelled with the same pitch could be selected in the same time frame of the spectrogram. Two separate formulations of k-sparsity are compared; first, a k-note-sparse stopping condition is used, where any number of atoms may be selected and the algorithm stops iterating when k notes have been selected; in the second case k-sparsity is maintained by constraining the selection step to only select one atom for any given pitch. In previous AMT research using OMP [125], non-negativity was enforced in the atom selection step while the backprojection step was performed by simply using Least Squares (LS), in typical OMP fashion. Conversely, the NN-OMP algorithm proposed in [16] uses a NNLS backprojection step. The OMP experiments with the datapoint dictionaries were run twice to compare these backprojection strategies in the context of AMT, particularly to test if the use of the LS backprojection has a negative effect in terms of AMT performance.

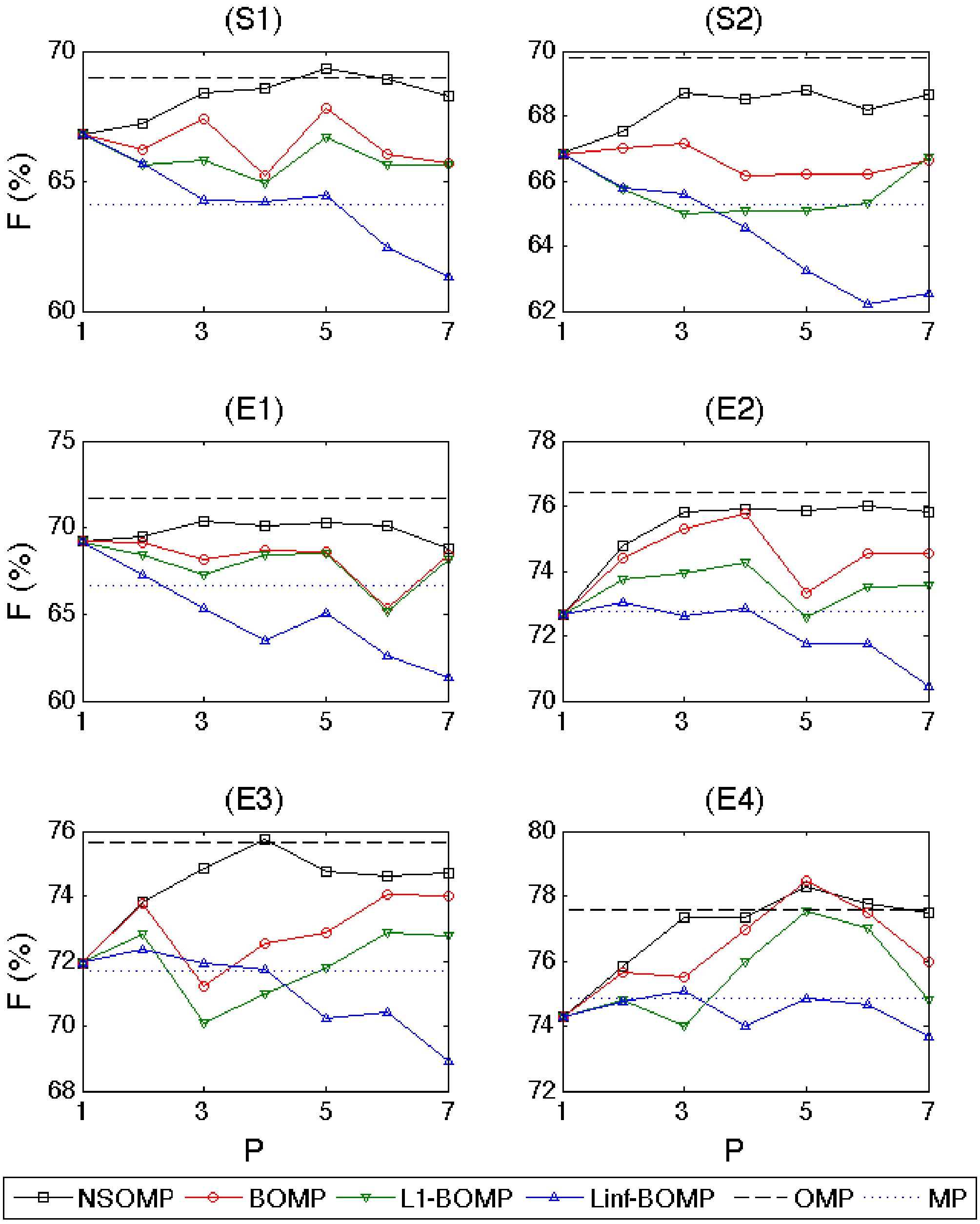

Figure 4.1: $\mathcal{F}$-measure for k-sparse transcription experiments comparing datapoint modelling approaches with subspace modelling approaches for various transforms. Note differing scales used.

Results

The results for the experiments described are shown in Figure 4.1. It should first be noted that the scaling on each graph differs, due to the different range of results found with the individual transforms. The difference in $\mathcal{F}$-measure, between Transforms $\mathbb{S}1$ and $\mathbb{E}4$, which display the worst and best performance, respectively is of the order of 10%. This pattern is observable across all algorithms and amounts to a substantial difference in performance, being greater than the differences between individual algorithms on a given transform.

Similar patterns in algorithm performance are observable across all transforms. When data-point modelling is used, OMP outperforms MP in all cases. In the worst performing transforms this performance gap was $\sim 5\%$, while in the transforms displaying improved results, the performance gap was shortened to $\sim 2.5\%$. Recalling that both algorithms were run in two different modes, a constrained k-sparse and a k-note sparse, the results shown in Figure 4.1 are for the k-note sparse approach, which performed better in all cases. Little difference was observed in the performance of the two approaches; for OMP the difference was $\sim 0.4\%$ for all transforms, while in the case of MP the difference was even smaller, at $\sim 0.1\%$. The comparative use of LS and NNLS backprojection steps was also compared by running OMP with both strategies, separately. However the difference in performance relative to the backprojection employed was negligible, and the results therefore are not recorded here.

Amongst the group sparse algorithms, NN-NS-OMP was seen to perform best consistently, followed by the NN-ℓ_2, NN-ℓ_1- and NN-ℓ_∞- BOMP algorithms, in that order. Better performance with another group sparse algorithm was seen only once across the 36 different combinations of groupsize and transform. This occured using the NN-BOMP algorithm for Transform $\mathbb{E}4$ with $P = 5$. The difference in performance between the NN-NS-OMP and NN-BOMP ranged up to 5%, averaging at around 2%. A tendency for NN-NS-OMP to have a smooth transition of $\mathcal{F}$-measure across values of P can be observed in Figure 4.1. Here it is observed that the other group sparse algorithms were less consistent with respect to the groupsize used, with large performance differences between adjacent groupsizes. Indeed sympathetic undulations are seen for the NN-BOMP algorithms, possibly indicating underlying properties of certain dictionaries. In terms of group size, performance tended to peak at $P = 4$ or 5, and to decrease at larger values of P, which could be construed as a overfitting phenomenon. Alternatively this could be an effect of the dictionary having meaningless atoms introduced when larger subspaces are learnt due to the simple unconstrained method used to learn the subdictionaries.

The use of datapoint modelling is seen, in Figure 4.1, to increase performance using NN-OMP by $2-4\%$ relative to the single atom per note model in each transform. Similar improvements relative to the single atom modelling are observable when the subspace modelling is used with an optimal value of P, and this improvement is enhanced when Transforms $\mathbb{E}2-4$, the larger dimension ERBTs are used. When the optimal groupsize dictionaries are used little difference in performance is observed between the subspace modelling, using NN-NS-OMP, and the datapoint modelling. Results using NN-NS-OMP are seen to outperform those of OMP with datapoint dictionaries in the case of Transforms $\mathbb{S}1$ and $\mathbb{E}3\&4$.

Experiments using mixed group sizes

Further experiments were run to test how the different algorithms perform when presented with subspace based dictionaries where the subspaces are of varying size. While it is straightforward to learn a dictionary with a fixed subspace size when isolated note signals are available, such signals may not be available and a dictionary may need to be learnt from a signal, or signals [20] [1], in which case it may be necessary to accommodate mixed group sizes.

As the NN-ℓ_1- and NN-ℓ_2- BOMP selection criteria employ a summation in their calculation, it may be expected that performance for these algorithms will suffer when presented with mixed group sizes, with a preference given to notes represented with by larger subspaces. Conversely, NN-NS-OMP, which considers an orthogonal projection, and NN-ℓ_∞-BOMP which uses a groupwise maximum can be expected to avoid mostly these issues related to the inner coherence of individual groups.

Experiments similar to those described in the previous section were run, differing only through the use of mixed group sizes. The group size relating to a given note was randomly selected from $P \in \{2,...,5\}$, and the relevant block was taken from the subspace dictionary of corresponding P. These individual note-specific blocks were then concatenated into a dictionary, and the group structure, $\mathcal{L}$, (§2.2.2) indexed accordingly. Experiments were run for all transforms outlined in (§3.2). Each piece in the dataset was assigned its own specific random pattern of groupsizes, which was used for all transforms and algorithms.

The results are presented in Table 4.1. Here it is observed that NN-NS-OMP and NN-ℓ_1-BOMP record similar performance to the standard case when the blocks are of equal size, as may be expected. However, NN-ℓ_1- and NN-ℓ_2- BOMP are seen to deteriorate significantly.

Transform	NN-ℓ_1	NN-BOMP	NN-ℓ_∞	NS-OMP
$\mathbb{S}1$	55.3	62.7	64.3	**68.0**
$\mathbb{S}2$	55.0	62.4	64.8	**67.8**
$\mathbb{E}1$	64.1	57.8	65.3	**69.4**
$\mathbb{E}2$	62.6	69.8	72.6	**75.1**
$\mathbb{E}3$	60.6	68.2	71.1	**74.1**
$\mathbb{E}4$	65.3	72.8	74.6	**76.8**

Table 4.1: $\mathcal{F}$-measure for various greedy group sparse approaches for a range of transforms in experiments with mixed group sizes.

4.2 A fast implementation of NN-NS-OMP

It is observed in the experimental results in the prior section that NN-NS-OMP outperforms the other non-negative group sparse methods, displaying a consistency not seen with other group methods and adpating well when a dictionary with mixed group sizes is employed. However, this improved performance comes at a very high computational expense. The selection criteria of the NN-BOMP algorithms can easily be calculated by vector and matrix multiplications. This simplicity is not afforded to NN-NS-OMP as the non-negative constraint on the groupwise solution vector, $\hat{\mathbf{x}}[l]$, (4.6) requires that NNLS be used to find the active set of atoms in a given group in order to calculate the value of the selection criteria (4.6). While these NNLS projections may use small subdictionaries, a significant computational expense is incurred as a NN-NS-OMP decomposition of a $60sec$ spectrogram with $23ms$ time frames requires of the order of 10^6 of these small NNLS calculations. While this could be approximated by using a Least Squares projection, as in SMP [47], some prior experiments determined that this was an oversimplification, leading to a deterioration in performance.

In order to derive a fast NN-NS-OMP algorithm, it is necessary to recall the SMP selection criteria (4.2) and in particular the coefficient for each group, which can be rewritten:

$$\begin{aligned} ||\mathbf{r} - \pi_l(\mathbf{r})||_2^2 &= ||\mathbf{r} - \mathbf{D}[l](\mathbf{D}[l])^\dagger \mathbf{r}||_2^2 \\ &= ||\mathbf{r}||_2^2 - \phi[l]^T \hat{\mathbf{x}}[l] \end{aligned} \tag{4.7}$$

where $\phi = \mathbf{D}^T\mathbf{r}$ and $\hat{\mathbf{x}}[l]$ is the LS solution vector for the lth group. This enables the SMP selection criteria (4.2) to be expressed as

$$\hat{l} = \arg\max_l \phi[l]^T \hat{\mathbf{x}}[l] \tag{4.8}$$

which is referred to here as the Fast SMP (FSMP) selection criteria. One way that FSMP can be solved is through introduction of an auxiliary dictionary

$$\Theta[l] = [\mathbf{D}[l]^{\dagger}]^{T} \tag{4.9}$$

where $\mathbf{D}[l]^{\dagger}$ is the Moore-Penrose pseudoinverse of $\mathbf{D}[l]$. Using the auxiliary dictionary the groupwise LS solution vector $\hat{\mathbf{x}}[l] = \Theta[l]^{T}\mathbf{r}$ is easily calculated, while ϕ is available using the original dictionary allowing the FSMP to be calculated using a two-dictionary approach.

A more straightforward single dictionary approach is possible, through orthogonalisation of the individual subspaces. Given that $\hat{\mathbf{x}}[l] = (\mathbf{D}[l]^{T}\mathbf{D}[l])^{-1}\phi[l]$ from the definition of the pseudoinverse, the FSMP group coefficient can be rewritten

$$\begin{aligned} \phi[l]^{T}\hat{\mathbf{x}}[l] &= \mathbf{r}^{T}\mathbf{D}[l](\mathbf{D}[l]^{T}\mathbf{D}[l])^{-1}\mathbf{D}[l]^{T}\mathbf{r} \\ &= \mathbf{r}^{T}\Psi[l]\Psi[l]^{T}\mathbf{r} = \|\Psi[l]^{T}\mathbf{r}\|_{2}^{2} \end{aligned} \tag{4.10}$$

where $\Psi[l] = (\mathbf{D}[l]^{T}\mathbf{D}[l])^{-1/2}\mathbf{D}[l]^{T}$ is the polar decomposition of $\mathbf{D}[l]$. The final expression in (4.10) suggests the BOMP algorithm can be used to solve SMP when supplied with the dictionary Ψ rather than the original dictionary $\mathbf{D}$. Other subspace orthogonalisations to derive Ψ are also possible.

However, this formulation using orthogonalised subspaces is not appropriate in the non-negative framework. While projection onto a given subspace and its orthogonal complement are equivalent, there is no simple correspondence between the individual atoms or their coefficients and the set of atoms that would form a non-negative backprojection coefficient vector is not simply discernible. However, a Fast NN-NS-OMP (F-NS-OMP) selection criteria defined by:

$$\hat{l} = \arg\max_{l} \hat{\mathbf{x}}[l]^{T}\phi[l] \tag{4.11}$$

in which $\hat{\mathbf{x}}$ is the NNLS solution vector is amenable to fast expression using the two dictionary formulation. This is still not straightforward as the NNLS solution vector, $\hat{\mathbf{x}}[l]$, is not simply derivable by dictionary calculation. An alternative approach is proposed in which the F-NS-OMP coefficient for each group is bounded. These bounds are first defined.

Fact 1. *The F-NS-OMP selection criteria (4.11) is upper-bounded by the FSMP selection criteria (4.8).*

Algorithm 4.2 Fast NN-NS-OMP (F-NS-OMP)

```
**Input** $\mathbf{D} \in \mathbb{R}^{\mathbf{M} \times \mathbf{N}}, \quad \Theta \in \mathbb{R}^{M \times N}, \quad \mathbf{r} \in \mathbb{R}^{M}$
$\phi = \mathbf{D}^T \mathbf{r}; \quad \phi^+ = \mathcal{I}\phi; \quad \mathbf{x} = \Theta^T \mathbf{r}$
$\mathbf{f}_l = \min[\mathbf{x}[l]^T \phi[l], \phi^+[l]^T \phi[l]]$
$\mathcal{J} = \{j | \mathbf{f}_j > \mathbf{f}_{j+1}\}$
$t = 0; \quad \zeta_{max} = 0$
**while** $\mathbf{f}_{\mathcal{J}(t+1)} > \delta_{max}$ **do**
  $t = t + 1$
  $l = \mathcal{J}(t)$
  $\mathbf{x}^+[l] = \arg\min_x \|\mathbf{r}^i - \mathbf{D}[l]\mathbf{x}\|_2^2 \quad s.t. \quad \mathbf{x} \geq 0$
  $\zeta = \mathbf{x}^+[l]\phi[l]$
  **if** $\zeta > \zeta_{max}$ **then**
    $\zeta_{max} = \zeta$
    $\hat{l} = l$
  **end if**
**end while**
**Output** $\hat{l}$
```

Proof. This is a simple result of least squares projections. Given that **B** is a column submatrix of **A** it follows that $\pi^{\mathbf{A}}(r) \geq \pi^{\mathbf{B}}(r)$, while a NNLS solution refers to the least squares solution of a column submatrix. □

Fact 2. *The F-NS-OMP selection criteria (4.11) is upper-bounded by the NN-BOMP selection criteria (4.3) in the non-negative case, i.e. when* $\mathbf{D} \geq 0$

Proof. In a non-negative framework, it is necessary that $\mathbf{x}_i \leq \phi_i$ where $\mathbf{x}$ and ϕ are the NNLS and inner product coefficients given a non-negative dictionary **D** and a signal **s**. □

The subprocedure outlined in Algorithm 4.2 can be used to calculate the NN-NS-OMP selection criteria (4.6) using these bounds. This subprocedure is initialised by calculating the blockwise least squares coefficient vector $\hat{\mathbf{x}}$ and the inner products, ϕ, of the dictionary and the residual The FSMP (4.8) and NNBOMP (4.3) group coefficients are then derived using $\hat{\mathbf{x}}$ and ϕ. The variable f_l expressing the groupwise minimum of the FSMP and NN-BOMP for the lth group then sets an upper bound to the F-NS-OMP selection criteria (4.11). The ordered set, $\mathcal{J}$, of group indices, l, sorted in descending order of f_l, is formed before an iterative loop is entered after a counter, t, is initialised to zero. At each iteration, NNLS is run for the group indexed by $l = \mathcal{J}(t)$ to derive the solution vector $\mathbf{x}^+[l]$. The F-NS-OMP group coefficient, ζ, is then calculated using $\mathbf{x}^+[l]$, and compared to the current best estimate, ζ_{max}. If ζ is larger than ζ_{max}, its value is assigned to ζ_{max}, and l is assigned to $\hat{l}$. This iteration continues until the next group index

$l = \mathcal{J}(t+1)$ in the set relates a group with an upper bound, f_l that is lower than the current best estimate, ζ_{max}. When this happens, the loop stops and the subprocedure then returns the selected index $\hat{l}$.

In this way it is expected that the amount of NNLS iterations required to assert the NN-NS-OMP selection criteria will be pruned, while the calculation and ordering of the bounds is relatively computationally inexpensive. It is notable that the proposed accelerated F-NS-OMP does not constitute an approximation of the NN-NS-OMP, and the same results are derived using both approaches.

Experiments

Experiments were run to test if the proposed F-NS-OMP provides an expected acceleration relative to the initial implementation in which a NNLS calculation is performed for each group. For comparison, timing experiments are also run using OMP and MP with the datapoint dictionaries, and NN-BOMP with the subspace dictionaries.

For all OMP-based algorithms, including the group sparse variants, the inner product of the dictionary atoms with the residual is approximated at each iteration using

$$\phi^{i+1} = \alpha - \mathbf{D}^T \mathbf{D}_\Gamma \mathbf{x}^i_\Gamma \tag{4.12}$$

where $\alpha = \mathbf{D}^T \mathbf{s}$ and $\mathbf{D}^T \mathbf{D}$ is precalculated, as suggested for OMP in [121] and [110], and also used as part of the F-NNLS algorithm [14]. In MP, the inner product vector is updated in a similar fashion, while only considering the coefficient of the last atom to be selected:

$$\phi^{i+1} = \phi^i - \mathbf{D}^T \mathbf{d}_{\hat{l}} \phi^i_{\hat{l}}. \tag{4.13}$$

The timing results given for OMP consider the case when the non-negative backprojection step is omitted and the solution is k-sparse, when only one atom is selected to represent each note, i.e. the fastest variant of the OMP algorithms described in the experiments in (§4.1). For the group sparse algorithms, the groupsize is set to $P = 5$, as this value was generally seen to be optimal in the earlier experiments. The group sparse algorithms use the F-NNLS algorithm [14] configured for warm restarts for the back projection step.

The timing results for all algorithms are outlined in Table 4.2. Here it is seen that MP is generally the fastest algorithm, as can be expected due to the lack of a backprojection step.

Transform	NN-MP	NN-OMP	NN-BOMP	F-NS-OMP	NS-OMP
$\mathbb{S}1$	**29.7**	38.4	34.4	55.3	1087
$\mathbb{S}2$	48.8	53.3	**34.3**	55.6	1073
$\mathbb{E}1$	**17.0**	26.7	33.7	53.9	1048
$\mathbb{E}2$	**21.5**	29.2	33.6	50.7	1072
$\mathbb{E}3$	**21.0**	29.2	35.4	52.7	1067
$\mathbb{E}4$	**28.7**	38.0	33.7	48.8	1084

Table 4.2: Time taken in seconds to transcribe 15 minutes of music using OMP based algorithms.

However, NN-BOMP is seen to be faster in the case of Transform $\mathbb{S}2$, which is of the largest dimension, 2048, of all dictionaries.

OMP and MP are both seen to vary greatly between transforms, due solely to the initial inner product calculation introducing a computational load which scales with the atom/spectrogram dimension. It is worth noting that the authors of [125] ignore the fast inner product vector update used (4.12), instead using an Approximate OMP (AMP) with approximate nearest neighbour search. Some experiments using OMP without the inner product approximation (4.12) suggest that (4.12) effects a large speedup in the execution time, as the matrix multiplication of the dictionary and the residual is not performed. Indeed it would appear that this speedup is larger than that described using AMP [125] without suffering the reported degradation of results.

In comparison, the group sparse methods are seen to vary little relative to the transform used as the dictionary size is much smaller and the computation load is dominated by the backprojection step. The F-NS-OMP is seen to be expensive relative to NN-BOMP typically taking more than 1.5 times as long to calculate. However, F-NS-OMP is seen to accelerate the NN-NS-OMP, being about 20 times faster, while again noting that this acceleration does not come with a degradation in terms of performance, as the F-NS-OMP selection criteria, as described in Algorithm 4.2 is not an approximation.

Comparing the OMP with datapoint dictionaries against the group sparse methods, it is seen that the group sparse methods are faster on all the transforms with dimension larger than 512, i.e. the two STFTs, and Transform $\mathbb{E}4$, the largest ERBT, which has been observed in the previous section to give superior AMT performance to the other transforms.

4.3 Non-negative Group Gradient Pursuits

The largest computational expense incurred by OMP is associated with the backprojection step [121]. The backprojection step consists of a LS projection of the supported atoms onto the signal that is performed after each new atom is added to the sparse support. Different approaches such as using rank-one QR or Cholesky decomposition updates to ameliorate the computational expense associated with backprojection have been considered and a comparison of some of these approaches is described in [121]. Gradient Pursuits [11] provide a fast alternative to OMP, from which they differ only through use of an approximate backprojection step which may be performed using gradient descent methods [11],

It was observed in previous experiments that the group greedy non-negative OMPs suffer in computational terms, by the need to use a NNLS backprojection step. Performance is not seen to degrade when a LS backprojection is used in OMP for the context of AMT when the datapoint dictionaries are used. However, this is not the case when the subspace dictionaries are used, possibly an effect of the interdependency of the atoms in each group of the subspace dictionaries

Several attempts to counter the computational load in terms of the backprojection in non-negative group greedy algorithms were undertaken. While the execution time is reduced by a considerable amount when a warm restart is used, the addition of a group may still require several iterations of the F-NNLS algorithm in order to add new atoms to the active set, while possibly ejecting others. Adding all atoms of positive coefficient from a newly selected group to the active set, similar to a stagewise approach [35] and also proposed for greedy group sparsity [37] [47] may seem an appropriate strategy. However, scaling errors were incurred in some initial experiments using this approach, assumed to be due to badly conditioned dictionaries.

Similarly, experiments were run using the block pivoting approach [105]. However this approach required calculation of a pseudoinverse at each iteration of NNLS, and even with a warm restart was slower than FNNLS. A further attempt, in which explicit calculation of the pseudoinverse in the block pivoting method was replaced through use of the inverse of the Gram matrix, in a manner similar to that employed by F-NNLS [14], was proposed. However, scaling errors were again seen using this approach.

It is useful to consider, in this case, the use of a gradient pursuit, and the NMF coefficient matrix update (2.32) is known to provide a fast gradient descent update with proven monotonic descent in a non-negative framework. A gradient variant of the general Non-Negative Group OMP, Algorithm 4.1, is proposed by substituting the NNLS backprojection step with a fixed number, w, of the Euclidean distance NMF coefficient matrix updates (2.32).

		NNLS	Gradient (w)				
			1	*2*	*5*	*10*	*20*
Time (*secs*)	NN-BOMP	33.7	11.5	13.0	17.4	24.7	39.6
	F-NS-OMP	51.2	39.6	41.1	44.3	49.7	61.3
$\mathcal{F}$-measure (%)	NN-BOMP	72.9	**73.7**	73.5	73.0	72.9	72.9
	F-NS-OMP	**74.8**	73.4	73.7	74.1	74.2	74.3

Table 4.3: Comparision of NNLS and gradient based backprojections for F-NS-OMP and NN-BOMP in terms of time and AMT performance using $\mathcal{F}$-measure. The number of iterations of the multiplicative update, w, is varied between 1 and 20 (indicated in italics).

Experiments

Some experiments were run to test the effectiveness of the gradient approach, in terms of execution time and potential deterioration in AMT performance. Gradient variants of both the NN-BOMP and F-NS-OMP, with a varying number of iterations of the NMF multiplicative update, $w \in \{1,2,5,10,20\}$ were used to decompose the spectrograms of Transform $\mathbb{E}2$ using subspace dictionaries with $P = 5$. A k-sparse stopping condition was used, with k_t equal to the known polyphony at each time frame. The assumption is made that performance will decrease in the gradient algorithms, as an approximation is used.

The multiplicative update (2.32) does not effect any difference on a coefficient set to zero, so initialisation of the coefficients of atoms newly added to the sparse support is necessary. Different initialisation strategies were used with the two algorithms. For Gradient NN-BOMP, only atoms belonging to the selected group which displayed a positive inner product with the residual were added to the support. The inner products of these atoms with the residual were used to initialise the corresponding coefficients in the vector, $\mathbf{x}_t$. In the case of Gradient F-NS-OMP, the selection criteria requires calculation of an NNLS coefficient vector $\hat{\mathbf{x}}_t[l]$ for the newly added group. This NNLS coefficient vector is used to initialise the coefficients of the newly selected group.

Results are given in terms of computation time and $\mathcal{F}$-measure in Table 4.3. Some differing patterns are seen in the results, relative to the algorithm used. In the case of F-NS-OMP it is observed that the gradient backprojection decreases the performance in terms of $\mathcal{F}$-measure with the worst performance seen when $w = 1$ and the best when $w = 20$. However, with NN-BOMP the opposite occurs. Improved $\mathcal{F}$-measure is observed when the gradient backprojection is employed and the best performance is seen when $w = 1$. This was originally assumed to be an effect of using the different initialisations for each algorithm. However, further experiments showed that this was not the case and both algorithms performed worse than reported in Table

4.3 when employing the alternative initialisation.

Furthermore, the timing results show some unexpected phenomena. In both cases, a speedup is seen in all cases when $w < 20$. However, the scale of the speedup is larger for NN-BOMP, which drops from $33.7sec$ to $11.5sec$, a decrease of $\sim 22sec$, being now faster than even MP. However, in the case of the F-NS-OMP, the speedup is $\sim 11sec$, around half of that of the NN-BOMP. This might suggest that the original NNLS backprojection was for some reason more efficient with the NS-OMP, possibly requiring less backtracking due to the use of the prior NNLS projection in the group selection criteria.

4.4 Discussion

In this chapter, the use of greedy methods for the application of AMT has been explored. In particular the use the group sparse greedy methods with subspace modelling of piano notes was introduced, a novel approach in AMT. Several greedy non-negative group sparse methods were proposed, the best of which, NN-NS-OMP, was seen to be computationally demanding. However, a strategy based on bounded projections was seen to afford a large speed up of this algorithm, with zero effect on performance. Further speed-ups were shown using a gradient approach to perform a backprojection step, however this was seen to effect the performance, albeit positively in some cases.

The subspace modelling, using group sparsity, was compared with the use of both single atom per pitch modelling and datapoint modelling with standard sparsity. It was found that the subspace-based decompositions performed better than the atomic pitched dictionary based decompositions and similar to the datapoint dictionary based decompositions, when the size of the group of atoms representing the subspace is optimal. This suggests that the subspace model is an apt model for AMT. Further exploration of this model is undertaken in the following chapters. One interesting observation was the large discrepancy seen between performance relative to different transforms. Indeed, the differences in performance observed for different transforms were larger than those recorded for different algorithms with a given transform employed. This is suspected to be an effect of varying dictionary coherence in the different transforms, which leads to further analysis provided in Chapter 8.

To summarise, in the context of AMT, it has been seen that the use of different transforms can affect the performance, and the subspace model can also provide an improvement similar to that found using a large datapoint dictionary, while having similar computational expense.

While it has been observed that subspace modelling and the use of different transforms may be useful for AMT in the context of greedy methods, the usage of this methodology has not necessarily been advanced. Fractured time continuity is a previously reported problem using greedy methods for AMT [20] that was observed in the experiments undertaken, even though some improvement in this aspect was seen in the better performing transforms. Previously noted advantages of greedy methods are similarly unaffected, such as their potential for use in multi-instrument signals [76], as greedy methods discourage co-activity of instruments observed using gradient based methods [133]. While speed is a noted advantage of greedy methods [23], this may not be so important in a non-negative framework where fast algorithms are available mostly due to the use of multiplicative updates.

One flaw of greedy methods observed in the experiments undertaken is that many mistakes in atom selection can happen at a very early stage; even at the first iteration upwards of 10% of selections are incorrect. In a purely greedy method these early atom selection mistakes are irreversible. A previously noted problem with OMP for AMT is selection of an apt stopping condition when the polyphony, k, is unknown [125]. These two observations lead to an exploration in the next chapter of stepwise methods that include backwards elimination steps, and locally optimise the ℓ_0 sparse cost function at each step.

Chapter 5

Stepwise Optimal Methods

In the previous chapter, the use of methods based upon Orthogonal Matching Pursuit (OMP) was explored in the context of AMT. Subspace modelling was seen to be equivalent to datapoint modelling in this context, while the use of different signal transforms was observed to result in large relative differences in AMT performance. However, problems using matching pursuits in the context of AMT that have previously been described in the literature, such as fractured temporal continuity [20] and the difficulty in selecting an apt stopping condition [125] were still observed.

Fractured temporal continuity is observed when a ground truth atom remains undetected, while temporally adjacent time-pitch points are correctly selected. Often when a ground truth atom is undetected using OMP, an atom representing a harmonically related note is detected instead, a phenomenon that can be understood as a dictionary coherence problem. High coherence is expected between musical atoms, particularly those that are consonant due to the harmonic overlap and non-negativity.

A graphical description of this problem with OMP and musical dictionaries is shown in Figure 5.1, in a simple, noiseless problem using a dictionary with three atoms. Two of these three atoms are active in the signal. However, the inactive atom has a higher correlation with the signal. In the described problem, an incorrect selection is made at the first iteration if a greedy method is used, an example of the myopic nature of greedy methods [83], while NNLS would be expected to recover the correct representation in this noiseless scenario. Alternatively, it has long been recognised that backwards steps, as used in the LARS/LASSO algorithms [127] and

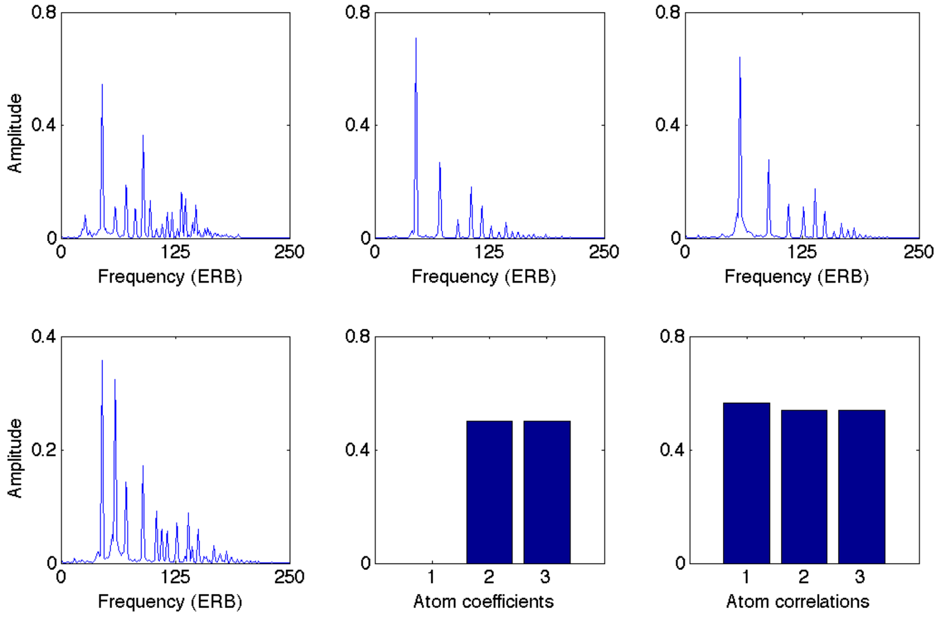

Figure 5.1: Graphical description of problems of overlapping harmonic partials. Three musical atoms are shown on top, representing the first, its octave and the twelfth note of a scale. On the bottom are shown, from left, a signal composed from a superposition of two atoms, the synthesis coefficents and the correlations of each atom with the signal

.

suggested for Polytope Faces Pursuit [102], may be needed to correct false detections when a greedy method is used. Such an approach has not previously been proposed in the context of AMT.

Optimal Steps

Orthogonal Matching Pursuit can be seen as part of a larger family of algorithms referred to as stepwise regression. In stepwise regression different selection criteria can be used while several stepwise strategies are common. Forward selection, referred to as Order Recursive Matching Pursuit (ORMP) [86] in the sparse representations literature, adds a single atom at each iteration, in a similar fashion to OMP [98]. OMP selects the atom with the largest gradient in terms of the least squares cost function, related through the inner products of the dictionary with the residual. ORMP looks deeper in its atom selection, seeking to select the atom, indexed by $\hat{n}$ which when added to the current sparse support will cause the largest reduction in the residual error norm. The difference in the residual norm effected by adding an atom indexed by n to the sparse support is denoted by :

$$\Delta_F \mathbf{r}^n = ||\mathbf{r}^i||_2^2 - ||\bar{\mathbf{r}}^n||_2^2 \tag{5.1}$$

where $\mathbf{r}^i$ is the residual at the current iteration, and $\bar{\mathbf{r}}^n$ is the residual for the hypothetical sparse support $\Gamma^n = \Gamma \cup n$, where Γ contains the indices of the current sparse support. Using this definition the forward selection criteria is given by

$$\hat{n} = \arg\max_n \Delta_F \mathbf{r}^n. \tag{5.2}$$

ORMP is considered a slower algorithm than OMP [10] and several variants have been proposed in the sparse representations literature [10] [108] [86] [83] offering optimised computational performance.

Backwards elimination is an alternative strategy in which the initial estimate contains all, or many candidates. Similar to forward selection, the elimination cost of an atom is given by

$$\Delta_B \mathbf{r}^n = ||\bar{\mathbf{r}}_B^n||_2^2 - ||\mathbf{r}^i||_2^2 \tag{5.3}$$

where $\bar{\mathbf{r}}_B^n$ is the residual given the hypothetical sparse support $\Gamma^n = \Gamma \backslash n$. At each iteration one atom indexed by

$$\hat{n} = \arg\min_n \Delta_B \mathbf{r}^n \tag{5.4}$$

is eliminated from the sparse support Γ.

OMP is considered an ℓ_0 approximation algorithm and is guaranteed to solve the k-sparse problem when the Exact Recovery Condition (ERC) [128] is met. However, in the case of highly coherent dictionaries, a condition such as ERC can be considered irrelevant. Indeed, the ℓ_0-penalised sparse cost function

$$\mathcal{C}_S = ||\mathbf{s} - \mathbf{D}\mathbf{x}||_2^2 + \lambda ||\mathbf{x}||_0 \tag{5.5}$$

is not necessarily optimised by any iteration of OMP, except for the initial selection, which selects the atom that displays the largest inner product with the current residual. The forwards selection criteria selects the atom that minimises the residual when added to the sparse support and therefore can be considered to make the locally optimal step with regard to the sparse cost function (5.5). Indeed the forward selection step is seen to reduce the value of (5.5) as long as $\Delta_F \mathbf{r}^{\hat{n}} > \lambda$ in which case the reduction in $||\mathbf{s} - \mathbf{D}\mathbf{x}||_2^2$ is offset by in the increment of λ in the sparse penalty term. The backwards elimination step performs a similar local optimisation, when $\Delta_B \mathbf{r}^{\hat{n}} < \lambda$. Hence, ORMP and backwards elimination can be referred to as *stepwise optimal*. One advantage of stepwise optimality is that the parameter λ, representing a threshold, can be used as a stopping condition in pursuit algorithms.

Fast calculation of the forwards selection (5.2) and backwards elimination (5.4) criteria are proposed as part of the Greedy Sparse Least Squares (GSLS) algorithm in [83]. Given that $||\mathbf{r}||_2^2 = ||\mathbf{s}||_2^2 - \alpha_\Gamma^T \mathbf{x}$, where $\alpha = \mathbf{D}^T \mathbf{s}$ and $\mathbf{x} = (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \alpha_\Gamma$ the forward selection coefficient (5.1) becomes

$$\begin{aligned} \Delta_F \mathbf{r}^n &= \alpha_\Gamma^T (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \alpha_\Gamma - \alpha_{\Gamma_n}^T (\mathbf{D}_{\Gamma_n}^T \mathbf{D}_{\Gamma_n})^{-1} \alpha_{\Gamma_n} \\ &= \alpha_\Gamma^T (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \alpha_\Gamma - \begin{bmatrix} \alpha_\Gamma \\ \alpha_n \end{bmatrix} \begin{bmatrix} \mathbf{A}_n & \mathbf{b}_n \\ \mathbf{b}_n^T & c_n \end{bmatrix} \begin{bmatrix} \alpha_\Gamma & \alpha_n \end{bmatrix} \end{aligned} \tag{5.6}$$

where $\mathbf{b}_n = \mathbf{D}_\Gamma^\dagger \mathbf{d}_n$; $c_n = 1/(\mathbf{d}_n^T \mathbf{d}_n - [\mathbf{D}_\Gamma^T \mathbf{d}_n]^T \mathbf{b}_n)$ and $\mathbf{A} = (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} + c_n [\mathbf{D}_\Gamma^\dagger \mathbf{d}_n][\mathbf{D}_\Gamma^\dagger \mathbf{d}_n]^T$ are given by the block matrix inverse formulae [7]. Cancellations in (5.6) and the fact that $||\mathbf{d}_n||_2 = 1 \, \forall n$ lead to the reconfigured selection criteria :

$$\Delta_F \mathbf{r}^n = \frac{\phi_n^2}{1 - \mathbf{g}_\Gamma^{n\,T} (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \mathbf{g}_\Gamma^n} \tag{5.7}$$

where $\mathbf{g}_\Gamma^n = \mathbf{d}_n^T \mathbf{D}_\Gamma$ is a submatrix of the Gram matrix $\mathbf{G} = \mathbf{D}^T \mathbf{D}$. Some matrix manipulation conveniently allows (5.7) to be calculated for all atoms simultaneously, using matrix and vector operations:

$$[\Delta_F \mathbf{r}] = \phi^{[2]} \oslash \left(\mathbf{1}_{|\Gamma|} - \left[\mathbf{G}_\Gamma \otimes [(\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \mathbf{G}_\Gamma] \right] \mathbf{1}_{|\Gamma|} \right) \tag{5.8}$$

where $\mathbf{G}_\Gamma = \mathbf{D}^T \mathbf{D}_\Gamma$; $\mathbf{1}_{|\Gamma|}$ is a column vector of dimension $|\Gamma|$ in which each element is equal to one, $\oslash$ denotes elementwise division and $\mathbf{x}^{[a]}$ denotes elementwise exponentiation of $\mathbf{x}$ to the power of a. Using a similar methodology based on block matrix inverse updates, the backwards GSLS [83] elimination criteria can also be derived:

$$\Delta_B \mathbf{r}^n = \arg\min_n \frac{\mathbf{x}_n^2}{[(\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1}]_{n,n}} \tag{5.9}$$

which, even more simply than (5.8) can be computed solely in terms of matrix and vector operations.

In the remainder of this chapter the use of stepwise optimal methods for AMT is explored. In the next section several bi-directional strategies from the literature are described and a backwards elimination strategy from an initial NNLS solution is proposed (BF-NNLS). Comparison using k-sparse experiments shows similar performance to NNLS for the stepwise optimal methods. A modified sparse cost function is then proposed, which is incorporated into the BF-NNLS. This backwards elimination approach is then extended to incorporate group sparsity, before concluding with a summary of the findings of the chapter.

5.1 Bi-directional stepwise optimal strategies

A bi-directional stepwise method incorporates both forwards and backwards steps, using a strategy to define the ordering of the steps. Recently, several papers in the sparse representations literature have proposed pursuit algorithms that incorporate different combinations of backwards and forwards steps in order to counter problems encountered with coherent dictionaries [131] [83] [60] [120].

One bi-directional strategy is the use of cyclic replacement, proposed in [120], where two different cyclic algorithms, Cyclic MP (CMP) and Cyclic OLS (COLS) are derived from MP and ORMP, respectively. The COLS algorithm is considered here as it can be seen as a bi-directional stepwise strategy that uses optimal forward selection. COLS, outlined in Algorithm 5.1, proceeds by selecting an initial support estimate through the normal iterations of ORMP. Once the support is initialised, the COLS algorithm iteratively deselects each atom. When an atom is deselected, the residual is recalculated and an atom is then selected using the forward selection criteria (5.7). If the newly selected atom and the deselected atom are the same, a counter, j, is incremented. Otherwise, the newly selected atom replaces the deselected atom in the sparse support. COLS converges when $j = k$; that is, a full cycle through all supported atoms occurs without any atoms being replaced.

An alternative formulation of the k-sparse problem is the family of subspace pursuits, including algorithms such as Subspace Pursuit [29], CoSaMP [87], and Iterative Hard Thresholding [123], which are known to afford accurate recovery under ERC and RIP conditions. However, problems when dealing with coherent dictionaries with subspace pursuits are noted in [131] where a related algorithm called Stepwise Optimal Subspace Pursuit (SOSP) is proposed. SOSP, outlined in Algorithm 5.2, is a bi-directional pursuit algorithm that is initialised, similar to COLS, by selecting k atoms using ORMP. The algorithm then enters an inner loop, in which Δ extra

Algorithm 5.1 Cyclic OLS

```
**Input** $\mathbf{D} \in \mathbb{R}^{M \times N}$, $\mathbf{s} \in \mathbb{R}^{M}$, $k$
**Initialise** $\Gamma = \{\}$; $i = 0$; $j = 0$
**repeat**
  Select atom index $\hat{n}$ using (5.2); $\Gamma = \Gamma \cup \hat{n}$
**until** $|\Gamma| = k$
**repeat**
  $i = i + 1$; $\hat{i} = i \bmod k$; $\bar{n} = \Gamma^{\hat{i}}$; $\Gamma^{\hat{i}} \leftarrow \Gamma \backslash \bar{n}$
  Get new atom index $\hat{l}$ using (5.2)
  **If** $\bar{n} = \hat{n}$
    $j = j + 1$
  **Else**
    $j = 0$; $\Gamma = \Gamma^{\hat{i}} \cup \hat{n}$
  **EndIf**
**until** $j = k$
**Output** $\Gamma$
```

atoms are added using forward selection steps, followed by removing Δ atoms using backwards elimination steps until the size of the sparse support is equal to k. The forwards and backwards steps proposed in the Greedy Sparse Least Squares (GSLS) algorithm [83] are used. After the inner loop of atom selection and elimination is completed, the new residual error, $\mathcal{C}_i$, is calculated, and Δ is decremented when a decrease in the current residual error relative to that of the previous iteration is not encountered. SOSP converges when Δ disappears.

Non-negative bi-directional stepwise strategies

In the context of a non-negative framework, such as is used for AMT throughout this thesis, some modifications have to be made to the COLS and SOSP algorithms. In particular, a non-negative variant of the fast forward selection criteria (5.7), proposed in [83] is required. The denominator of (5.7) can also be written as $\mathbf{d}_n^T(\mathbf{I} - \mathbf{D}_\Gamma \mathbf{D}_\Gamma^\dagger)\mathbf{d}_n$, where $\mathbf{D}^\dagger$ is the Moore-Penrose pseudoinverse. Indeed, the bracketed term is the projector onto the subspace $\mathbf{D}_\Gamma$ and is therefore positive semi-definite, leading to a positive denominator in all cases. Hence, a non-negative variant of the forward step of GSLS, referred to as NN-GSLS, is proposed simply by taking the square root of (5.7):

$$\hat{n} = \arg\max_n \frac{\phi_n}{(1 - \mathbf{g}_\Gamma^{n\,T}(\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1}\mathbf{g}_\Gamma^n)^{\frac{1}{2}}} \tag{5.10}$$

as the sign of the inner product of a given atom and the the residual, ϕ_n, is preserved due to the positive denominator. This can also be written in the form of (5.8) allowing calculation of the coefficients for all N atoms simultaneously. It is noted however, that this does not ensure a

Algorithm 5.2 SOSP

```
Input   $\mathbf{D} \in \mathbb{R}^{M \times N}$,   $\mathbf{s} \in \mathbb{R}^{M}$,   $k$,   $\Delta$
Initialise   $i = 0$;   $\Gamma = \{\}$
repeat
  Select atom index $\hat{n}$ using (5.2);   $\Gamma = \Gamma \cup \hat{n}$
until $|\Gamma| = k$
$\mathcal{C}^0 = \|\mathbf{s} - \mathbf{D}_\Gamma \mathbf{x}_\Gamma\|_2^2$
$i = 0$
repeat
  $i = i + 1$; $\Gamma^i = \Gamma^{(i-1)}$
  repeat
    Select atom index $\hat{n}$ using (5.2);   $\Gamma^i = \Gamma^i \cup \hat{n}$
  until $|\Gamma^i| = k + \Delta$
  repeat
    Select atom index $\hat{n}$ using (5.4);   $\Gamma^i = \Gamma \backslash \hat{n}$
  until $|\Gamma^i| = k$
  $\mathcal{C}^i = \|\mathbf{s} - \mathbf{D}_{\Gamma^i} \mathbf{x}_{\Gamma^i}\|_2^2$
  If $\mathcal{C}^i \geq \mathcal{C}^{i-1}$
    $\Gamma^i = \Gamma^{i-1}$;   $\mathrm{C}^i = \mathcal{C}^{i-1}$;   $\Delta = \Delta - 1$
until $\Delta = 0$
Output $\Gamma^i$
```

completely non-negative solution vector.

A non-negative variant of ORMP (NN-ORMP) is achieved using several iterations of the NN-GSLS selection criteria (5.10), with some subsequent replacements if necessary to maintain the non-negative constraint. After the initial support of k atoms was derived, elements displaying non-negative coefficients in a least squares (LS) backprojection were removed from the sparse support and atom additions were performed until the support was again of size k atoms. This removal and addition was repeated until a non-negative solution of cardinality k was found.

Similar steps are required for non-negative variants of both COLS and SOSP. For COLS, non-negativity of the current solution vector is checked after all replacements. When a replacement occurs, LS backprojection is performed to check if the non-negative constraint is met. If this constraint is violated, the newly selected atom is removed from the sparse support, and the atom with the next best non-negative forward selection value (5.10) is added. If necessary this is repeated until an atom was found for which the non-negative constraint is met.

SOSP adds a further Δ atoms to the support, after initialisation of the support using NN-ORMP. The authors of [131] propose that $\Delta = k$, that is the support is initially grown to $2k$, similar to the Subspace Pursuit [29]. However, in some cases it may be found that $|\mathbf{x} > 0| < 2k$, where $\mathbf{x}$ is the NNLS solution vector. A non-negative variant of SOSP needs to be aware of this, stopping

Algorithm 5.3 BF-NNLS algorithm

```
Input
    $\mathbf{D} \in \mathbb{R}^{M\times N}, \quad \mathbf{s} \in \mathbb{R}^{M}, \quad \mathbf{s},\mathbf{D} \geq 0, \quad k$ or $\lambda$
Initialise
    $\mathbf{x}^0 = \arg\min_x ||\mathbf{s}-\mathbf{Dx}||_2^2 \quad$ s.t. $\quad \mathbf{x} \geq 0$
    $\Gamma = \{j|x_j^0 > 0\}$
    $\mathbf{r}^0 = \mathbf{s}-\mathbf{Dx}^0$
repeat
    $\hat{n} = \arg\min \Delta_B \mathbf{r}^n \quad$ where $\quad n \in \Gamma$
    $\Gamma = \Gamma\backslash\hat{n}; \quad x_{\hat{n}} = 0;$
    $\mathbf{x}_\Gamma = \arg\min_x ||\mathbf{s}-\mathbf{D}_\Gamma\mathbf{x}||_2^2 \quad$ s.t. $\quad \mathbf{x} \geq 0$
until $||\mathbf{x}||_0 = k \quad$ or $\quad \Delta_B \mathbf{r}^{\hat{n}} > \lambda$
Output $\mathbf{x}$
```

when no more atoms display a positive inner product with the residual to avoid a potential critical loop. When Δ extra atoms have been added, a check needs to be made for non-negativity, and atoms displaying non-negative coefficients are removed from the sparse support. SOSP then uses a backwards elimination step. Some initial experiments led to the observation that the non-negative constraint is rarely broken in the backwards elimination stage when the forward solution was positive, and backwards elimination was performed naively, without consideration being given to the non-negative constraint.

It is seen above that the non-negative constraint introduces some instability to the bi-directional strategies, particularly in the forward selection step. It is worthwhile to reconsider that the dictionaries used for the application of AMT are often undercomplete, avoiding the scenario encountered with overcomplete dictionaries where many solutions are possible. Even if an overcomplete dictionary, such as a datapoint dictionary (§3.3.1) or a subspace dictionary (§3.3.2) with large groupsize, P, and a small transform dimension, is used the active set NNLS algorithm is known to give a full rank solution and converges to a minimum [14].

Considering that NNLS provides a stable solution and backwards elimination was seen to generally maintain the non-negative constraint an algorithm referred to as Backwards From NNLS (BF-NNLS) is now proposed, in order to avoid the problems introduced to the forward step by the non-negative constraint. BF-NNLS is outlined in Algorithm 5.3 and is seen to be conceptually simple. Active set NNLS is used to initialise the solution vector, which can be considered as having taken all possible forward steps. This is followed by backwards elimination steps and a stopping condition can be employed that considers the number of atoms selected, k, or a threshold λ.

Transform	OMP	ORMP	SOSP	COLS	BF-NNLS	T-NNLS
$\mathbb{S}1$	66.8	66.9	68.1	69.6	69.9	**70.0**
$\mathbb{S}2$	66.8	66.9	68.1	69.7	69.9	**70.0**
$\mathbb{E}1$	69.2	69.3	70.4	71.9	72.2	**72.3**
$\mathbb{E}2$	72.6	72.7	73.8	74.4	**74.5**	**74.5**
$\mathbb{E}3$	71.9	72.1	73.2	74.0	74.0	**74.1**
$\mathbb{E}4$	74.3	74.3	75.3	**75.7**	75.6	75.6

Table 5.1: OMP and bi-directional algorithms compared in terms of $\mathcal{F}$-measure with T-NNLS in k-sparse experiments on standard dataset across all transforms using single atom dictionaries.

Experiments

Experiments were run to compare the OMP, T-NNLS and BF-NNLS with non-negative variants of SOSP and COLS. NN-ORMP is also compared as SOSP and COLS initialise with ORMP, as described above. It is noteworthy that all the stepwise methods ultimately form a subset NNLS solution, and it is queried whether the stepwise approaches may be able to capture the more relevant signal elements, while countering some of the problems introduced by harmonic overlaps experienced using OMP.

Decompositions of the standard dataset (§3.1) are performed with the atomic pitch dictionaries (§3.3.2) using all transforms outlined in (§3.2). A k-sparse experimental setup was used with the stepwise methods set to select k atoms while k-sparse thresholding (§3.4) was used as the thresholding strategy for T-NNLS.

Results

The results for the experiments are outlined in Table 5.1. Here it is seen that the ORMP performs slightly better than OMP, while all bi-directional methods outperform these forwards only strategies. There is little to divide the different bi-directional approaches, with SOSP performing worst, and BF-NNLS generally performing best. However, improvement on NNLS results are only seen once. Hence, COLS and SOSP can be considered unuseful in this context, as BF-NNLS is faster, simpler and provides marginally better results.

In terms of performance relative to signal transform, a similar pattern is seen in the variation of all algorithms, with the STFTs performing worst, and improvements with the larger dimension ERBTs with the largest of these, Transform $\mathbb{E}4$, performing the best. The variation in performance for the NNLS based methods is reasonably large at over 5%. However this is considerably smaller than the variation observed with OMP-based decompositions.

5.2 A modified sparse cost function

In the experiments in the previous section, the BF-NNLS was seen to perform similarly to NNLS in terms of AMT in a k-sparse experimental setup. It is worth reconsidering the backwards elimination cost (5.4) for an atom that, using the block matrix inversion formulae, can also be written

$$\Delta_B \mathbf{r}^n = \frac{x_n^2}{\left(\mathbf{d}_n^T(\mathbf{I}-\mathbf{D}_{\Gamma_n}\mathbf{D}_{\Gamma_n}^{\dagger})\mathbf{d}_n\right)} = \frac{x_n^2}{(1-\mathbf{d}_n^T\mathbf{D}_{\Gamma_n}\mathbf{D}_{\Gamma_n}^{\dagger}\mathbf{d}_n)} \tag{5.11}$$

where $\Gamma^n = \Gamma\backslash n$ as $||\mathbf{d}_n||_2 = 1\forall n$. From (5.11) it can be seen that the backwards elimination cost of an atom is related to two elements. The first of these is the square of its NNLS coefficient, x_n, given the current support. The second is the bracketed term $(1-\mathbf{d}_n^T\mathbf{D}_{\Gamma_n}\mathbf{D}_{\Gamma_n}^{\dagger}\mathbf{d}_n)$. This bracketed term shows that the elimination cost of an atom is related to correlation with other atoms. The value of the bracketed term is equal to 1 in the case of zero correlation with other supported atoms if $||\mathbf{d}_n||_2 = 1$. As correlation increases the value of the bracketed term can be expected to increase, and can be considered to apply a coherence-based weighting to the coefficient of an atom.

As noted earlier in this chapter, the backwards elimation step provides a local optimisation to the sparse ℓ_0 penalised least squares cost function (5.5) when $\lambda > \Delta_B \mathbf{r}^{\hat{n}}$. This allows AMT to be performed in a polyphony-blind manner using BF-NNLS, with λ used as a stopping condition, as seen in Algorithm 5.3. A similar global approach is possible using NNLS with the δ-thresholding strategy (§3.4.1), performed on the NNLS coefficient matrix $\mathbf{X}$.

It is recalled from above that the BF-NNLS elimination cost is constructed (5.11) in terms of $x_{\hat{n}}^2$, the square of the NNLS coefficient at a given iteration. It may be preferable to derive a coefficient in the backwards elimination that is based upon the atom coefficient rather than its square. One motivation is that the coefficients of individual atoms may be expected to scale well relative to a maximum coefficient value, as employed in the δ-thresholding approach (§3.4.1). Hence a *modified sparse cost function* is proposed using the sparse penalised residual norm

$$\mathcal{C}_{mod} = ||\mathbf{s}-\mathbf{D}\mathbf{x}||_2 + \lambda\,||\mathbf{x}||_0 \tag{5.12}$$

which differs from the usual sparse cost function (5.5) which uses a penalised least squares. In terms of the BF-NNLS algorithm, a modified elimination cost is given by

$$\hat{\Delta}_B \mathbf{r}^n = \|\bar{\mathbf{r}}_B^n\|_2 - \|\mathbf{r}^i\|_2 \tag{5.13}$$

which uses the same elimination criteria (5.9) as the standard ℓ_0-penalised least squares (5.5), as the ordering of the elimination criteria coefficients for both cost functions, $\hat{\Delta}_B \mathbf{r}$ and $\Delta_B \mathbf{r}$, is the same. The value of the modified elimination cost of the selected atom is easily calculated by :

$$\hat{\Delta}_B \mathbf{r}^{\hat{n}} = \sqrt{\|\mathbf{r}^i\|_2^2 + \Delta_B \mathbf{r}^{\hat{n}}} - \ \|\mathbf{r}^i\|_2 \tag{5.14}$$

Hence, the BF-NNLS algorithm can also be used to form an approximation for the modified sparse cost function (5.12) by using a stopping condition, $\lambda > \hat{\Delta}_B \mathbf{r}^{\hat{n}}$.

Experiments

Experiments were run to compare the BF-NNLS approach, using both the standard sparse (5.5) and the modified sparse (5.12) cost functions, with the T-NNLS. It is expected that the use of the modified sparse cost function will result in consistency in the value of δ_{opt} as scaling is expected to be relatively consistent regardless of the transform used. Furthermore, it can be expected that the backwards elimination approaches may produce better AMT results.

A δ-thresholded experimental setup was used. For all transforms outlined in Table 3.1 the standard dataset (§3.1) was decomposed using the atomic pitch dictionaries (§3.3.2). The results are tabulated in terms of the $\mathcal{F}, \mathcal{P}, \mathcal{R}$ ensemble of metrics. The value of δ_{opt}, representing the value of δ which gives the maximum value of $\mathcal{F}$-measure when applied across all pieces, is also tabulated in order to compare the scaling of the different approaches across transforms. The values of Precision, $\mathcal{P}$, and Recall, $\mathcal{R}$ tabulated are those recorded at δ_{opt}.

Results

The experimental results are shown in Table 5.2. In terms of $\mathcal{F}$-measure BF-NNLS with the modified sparse cost function is seen to outperform T-NNLS for all transforms by $1.4 \sim 2.0\%$, while BF-NNLS with the standard cost function is seen to perform worse than T-NNLS in all cases. The variability of the δ_{opt} value is seen also to validate the use of the modified sparse cost function. For five of the transforms the value is $\delta_{opt} = 41$dB, while it is 40dB for Transform $\mathbb{E}1$. In stark contrast, the value of δ_{opt} is seen to range between 15 - 46dB, showing that the modified sparse cost function scales better in the signals of interest. T-NNLS is also seen to have a small

Transform	Metric	T-NNLS	BF-NNLS ($\mathcal{C}_S$)	BF-NNLS ($\mathcal{C}_{mod}$)
$\mathbb{S}1$	$\mathcal{F}$	64.0	63.9	**65.8**
	δ_{opt}(dB)	30	21	41
	$\mathcal{P}$	**63.4**	60.9	61.7
	$\mathcal{R}$	64.6	67.2	**70.4**
$\mathbb{S}2$	$\mathcal{F}$	64.0	63.9	**65.8**
	δ_{opt}(dB)	28	45	41
	$\mathcal{P}$	**63.4**	61.8	62.8
	$\mathcal{R}$	64.7	66.2	**69.0**
$\mathbb{E}1$	$\mathcal{F}$	66.6	66.2	**68.0**
	δ_{opt}(dB)	30	46	40
	$\mathcal{P}$	**65.3**	61.4	63.4
	$\mathcal{R}$	68.0	71.8	**73.3**
$\mathbb{E}2$	$\mathcal{F}$	68.4	67.9	**70.0**
	δ_{opt}(dB)	28	45	41
	$\mathcal{P}$	65.5	64.3	**65.8**
	$\mathcal{R}$	71.5	71.8	**74.8**
$\mathbb{E}3$	$\mathcal{F}$	68.1	67.5	**69.6**
	δ_{opt}(dB)	28	42	41
	$\mathcal{P}$	64.4	63.4	**65.2**
	$\mathcal{R}$	72.3	72.3	**74.7**
$\mathbb{E}4$	$\mathcal{F}$	68.6	68.4	**70.6**
	δ_{opt}(dB)	27	41	41
	$\mathcal{P}$	67.2	66.6	**68.1**
	$\mathcal{R}$	70.0	70.3	**73.3**

Table 5.2: Comparision of T-NNLS, and BF-NNLS using standard sparse cost function $\mathcal{C}_S$ (5.5) and modified sparse cost function $\mathcal{C}_{mod}$ (5.12).

range of optimum values of δ_{opt}, ranging from 27-30dB, confirming the observations for NMF based algorithms in [133].

In terms of Recall, $\mathcal{R}$, the BF-NNLS with modified sparse cost function is seen to effect an improvement over the other approaches, reaching 5.8% relative to T-NNLS with Transform $\mathbb{S}1$, while a minimum improvement of 2.4% was observed for Transform $\mathbb{E}3$. BF-NNLS with the standard sparse cost function is also seen to improve over T-NNLS in terms of Recall. In terms of Precision, two patterns emerge. For the transforms which perform worst, Transforms $\mathbb{S}1, \mathbb{S}2\&\mathbb{E}1$, the Precision is seen to be highest for the T-NNLS followed by BF-NNLS using the modified sparse cost function. For the other transforms, BF-NNLS with the proposed cost function is seen to achieve the highest Precision values followed by T-NNLS. Variability of Precision is seen to be much lower than for Recall.

To summarise, the use of the modified sparse cost function is seen to improve AMT results, and to scale well relative to the signal. Conversely, backwards elimination using the standard cost is seen to result in a deterioration in AMT results relative to NNLS.

5.3 Group BF-NNLS

In the previous chapter, group sparsity with subspace dictionaries was seen to improve AMT performance when OMP-based methods were used. Meanwhile, backwards elimination with a modified sparse cost function is seen to improve NNLS based decompositions for AMT when the atomic pitch dictionaries are used. Therefore, it is proposed to introduce group sparsity to the backwards elimination framework.

Using a similar approach to the GSLS algorithm [83], using block inverse matrix updates, the downdate for a group indexed by $[l]$ is given by

$$\begin{aligned}\Delta_B \mathbf{r}[l] &= \begin{bmatrix}\alpha_{\Gamma_l}\\ \alpha[l]\end{bmatrix}\left[\mathbf{D}_\Gamma^T\mathbf{D}_\Gamma\right]^{-1}\left[\alpha_{\Gamma_l} \quad \alpha[l]\right] - \alpha_{\Gamma_l}^T\left[\mathbf{D}_{\Gamma_l}^T\mathbf{D}_{\Gamma l}\right]^{-1}\alpha_{\Gamma_l}\\ &= \begin{bmatrix}\alpha_{\Gamma_l}\\ \alpha[l]\end{bmatrix}\begin{bmatrix}\mathbf{A}_l & \mathbf{B}_l\\ \mathbf{B}_l^T & \mathbf{C}_l\end{bmatrix}\left[\alpha_{\Gamma_l} \quad \alpha[l]\right] - \alpha_{\Gamma_l}^T\left[\mathbf{D}_{\Gamma_l}^T\mathbf{D}_{\Gamma l}\right]^{-1}\alpha_{\Gamma_l} \qquad (5.15)\end{aligned}$$

where Γ_l is the hypothetical support $\Gamma\backslash\mathcal{L}^{(l)}$ and $\alpha = \mathbf{D}^T\mathbf{s}$. $\mathbf{A}_l, \mathbf{B}_l$ and $\mathbf{C}_l$ are given in the inverse of the Gram matrix of the current support. Using the block matrix updates it is seen that $[\mathbf{D}_{\Gamma_l}^T\mathbf{D}_{\Gamma l}]^{-1} = \mathbf{A}_l = \mathbf{B}\mathbf{B}^T/\mathbf{C}$, from which it follows that

$$\begin{aligned}\hat{l} &= \arg\min_l \Delta_B \mathbf{r}[l]\\ &= \arg\min_l \frac{\mathbf{x}[l]^T\mathbf{x}[l]}{[(\mathbf{D}_\Gamma^T\mathbf{D}_\Gamma)^{-1}][l,l]} \qquad (5.16)\end{aligned}$$

where $\mathbf{Y}[l,l]$ is the principal submatrix of $\mathbf{Y}$ indexed by the block of indices $\mathcal{L}^{(l)}$. Efficient calculation of all group elimination costs simultaneously, such as is performed in the ungrouped case (5.9), is not possible as a matrix inverse needs to be calculated for each group. In this case, the group downdate needs to be calculated for each active group separately, which can lead to many small matrix inverses being calculated.

In a similar fashion to the F–NS-OMP algorithm proposed in Algorithm 4.2, a simple strategy for accelerating this algorithm is proposed. Using the fact that the norm of the residual after Least Squares (LS) projection of a signal onto a set of atoms is less than the norm of the residual after a similar projection onto a subset of the atoms, it follows that the reduction in residual norm is greater for a group than for any of the individual atoms in the group, or otherwise stated

$$\begin{aligned}\frac{\mathbf{x}[l]^T\mathbf{x}[l]}{[(\mathbf{D}_\Gamma^T\mathbf{D}_\Gamma)^{-1}][l,l]} &\geq \max_{[l,i]}\frac{\mathbf{x}[l,i]^2}{[(\mathbf{D}_\Gamma^T\mathbf{D}_\Gamma)^{-1}][(l,i),(l,i)]}\\ &= \max_{n}\frac{x_n^2}{[(\mathbf{D}_\Gamma^T\mathbf{D}_\Gamma)^{-1}]_{n,n}} \quad where \quad n\in\mathcal{L}^{(l)}\end{aligned} \tag{5.17}$$

where $[(l,i)]$ refers to the ith member of the lth group, which is also referred to by the nth atom index. Using the fact (5.17), the atomwise elimination costs can be calculated simply using (5.9) and a maximum taken for each group. If the groupwise maximums are ordered in ascending magnitude an iterative process similar to that used in F-NS-OMP, outlined in Algorithm 4.2, is possible. Again, similar to F-NS-OMP calculation of a group coefficient (5.16) is not necessary if the elimination cost of any atom in a given group is larger than the current lowest group coefficient.

Similar to the standard atomic elimination case, the group sparse elimination criteria can be thought of as locally optimising the cost function

$$\mathcal{C}_G = ||\mathbf{s}-\mathbf{Dx}||_2^2+\lambda\,||x||_{\perp,0} \tag{5.18}$$

where $||x||_{\perp,0} = ||\mathbf{h}||_0$ and $h_l = ||\mathbf{D}[l]\mathbf{x}[l]||_2$. Other norms than the $\ell_\perp$-norm can be used for the group penalty, however the backwards elimination is not seen to necessarily optimise $\mathcal{C}_G$ in this case. As a modified cost function was seen to improve AMT performance in the standard sparse case, a *modified group sparse cost function* is proposed :

$$\mathcal{C}_{modG} = ||\mathbf{s}-\mathbf{Dx}||_2+\lambda\,||x||_{\perp,0} \tag{5.19}$$

again differing only in replacement of the least squares coefficient with the residual norm. Similar to the standard sparse case (5.14) the order of elimination is the same for $\mathcal{C}_G$ and $\mathcal{C}_{modG}$ and the modified downdate cost is simply expressed :

$$\hat{\Delta}_B\mathbf{r}[\hat{l}] = \sqrt{||\mathbf{r}^i||_2^2+\Delta_B\mathbf{r}[\hat{l}]}-||\mathbf{r}^i||_2. \tag{5.20}$$

k-sparse Experiments

Two sets of experiments are run to compare the effectiveness of GBF-NNLS against other stepwise methods for AMT. The first set of these use *k*-sparse decompositions to compare GBF-

Transform	T-NNLS	GT-NNLS	GBF-NNLS	NS-OMP
$\mathbb{S}1$	70.0	72.2 (5)	**75.0** (5)	69.3 (5)
$\mathbb{S}2$	70.0	71.9 (6)	**75.3** (6)	68.8 (5)
$\mathbb{E}1$	72.3	72.8 (5)	**76.8** (6)	70.4 (3)
$\mathbb{E}2$	74.5	76.9 (4)	**79.7** (7)	76.0 (6)
$\mathbb{E}3$	74.1	76.6 (4)	**79.4** (5)	75.8 (4)
$\mathbb{E}4$	75.6	78.7 (5)	**80.8** (5)	78.3 (5)

Table 5.3: Comparision of T-NNLS(P=1), Group T-NNLS, GBF-NNLS and NS-OMP in terms of precision in k-sparse experiments. Numbers in brackets denote optimal size of P for group decompositions

NNLS with GT-NNLS and NN-NS-OMP. The standard dataset (§3.1) was decomposed using the subspace dictionaries (§3.3.2) for all values of the groupsize $P \in \{2,...,7\}$. GBF-NNLS and NN-NS-OMP are set to select k atoms, while GT-NNLS is post-processed with a k-sparse thresholding (§3.4). The results for T-NNLS using the atomic pitch atom dictionaries are also retabulated for comparison. For the group sparse algorithms, the results given are for the optimal value of P.

The results for the experiments are shown in Table 5.3, where it seen that GBF-NNLS outperforms all other approaches. Compared to the T-NNLS approach an improvement of the order of 5% is seen for all transforms, while compared to the GT-NNLS an improvement of 2 to 4% is seen in each case, while the results using NN-NS-OMP are lower than those of GT-NNLS. The optimal groupsize is seen to be around $P = 5$ or 6 for GBF-NNLS, slightly larger than for GT-NNLS or NN-NS-OMP. Interestingly, the performance of NN-NS-OMP is observed to approach that of GT-NNLS in Transforms $\mathbb{E}$2-4 with the difference in $\mathcal{F}$-measure being seen to be less than 1% in each case. In these cases NN-NS-OMP is also seen to outperform T-NNLS.

The results validate the use of GBF-NNLS with the subspace dictionaries. In particular it should be noted that BF-NNLS and T-NNLS provide similar results in the k-sparse experiments for the standard sparse case in (§5.1), in which case it can be assumed that the ordering of relevant coefficients was similar. When the subspace dictionaries are used however, the ordering of the coefficients varies for GBF-NNLS relative to GT-NNLS.

δ-thresholded Experiments

Further experiments are run using the GBF-NNLS in a δ-thresholded experimental setup. It is expected that improvements in AMT performance should be seen using GBF-NNLS as this approach improves upon the GT-NNLS in k-sparse experiments. Meanwhile, the BF-NNLS

Transform	T-NNLS	GT-NNLS	BF-NNLS	GBF-NNLS
$\mathbb{S}1$	64.0	64.5 (5)	65.8	**70.4** (5)
$\mathbb{S}2$	64.0	64.0 (2)	65.8	**70.6** (6)
$\mathbb{E}1$	66.6	65.5 (5)	68.0	**71.4** (6)
$\mathbb{E}2$	68.4	69.2 (4)	70.0	**74.6** (5)
$\mathbb{E}3$	68.1	69.1 (4)	69.6	**74.9** (4)
$\mathbb{E}4$	68.6	70.0 (5)	70.6	**75.0** (5)

Table 5.4: Comparision of T-NNLS(P=1), Group T-NNLS, GBF-NNLS and NS-OMP in terms of $\mathcal{F}$-measure in δ-threshold experiments. Numbers in brackets denote optimal size of P for group decompositions

algorithm does not improve upon T-NNLS in k-sparse experiments, yet improvements are seen in δ-thresholding experiments when the modified sparse cost function is used. From these two observations it is expected that a large improvement may be seen using the GBF-NNLS approach.

In order to isolate the individual contributions of group sparsity and the BF-NNLS approach to see which might be most important, a comparison is made between GBF-NNLS and GT-NNLS for the group sparse approach and BF-NNLS and T-NNLS. For all experiments decompositions of all pieces in the standard dataset (§3.1) are performed using subspace dictionaries (§3.3.2) for $P \in \{1,...,7\}$. Again decomposition are performed using all transforms outlined in (§3.2). The T-NNLS and GT-NNLS are post-processed using δ-thresholding (§3.4.1). For the BF-NNLS approaches, the stopping condition, λ is calculated using the delta parameter (3.3) applied to the largest entry of the coefficient matrix $[H]_{l,t} = ||\mathbf{D}[l]\mathbf{x}_t[l]||_2$. Results for all algorithms are given in terms of the optimal $\mathcal{F}$-measure.

The results for these experiments are given in Table 5.4, where some slightly different patterns are seen relative to the k-sparse results. The GT-NNLS does not always improve on the T-NNLS results and any improvements observed are relatively small $(< 1.5\%)$. BF-NNLS using the atomic pitch dictionaries is seen to outperform the GT-NNLS approach in all cases. However, the GBF-NNLS shows significant improvements over the other approaches, showing improvements of at least 5% relative to the GT-NNLS, and of over 3.5% relative to BF-NNLS, in all transforms. While the k-sparse threshold experiments measure the relative magnitudes at each time frame of the decomposition, the δ-thresholded experiments measure the magnitudes globally over a complete transform, suggesting that the backwards elimination cost may be a more appropriate measure of the energy for a given atom, possibly due to the coherence-aware weighting placed on the coeffcients. In terms of the groupsize, the optimal results in terms of $\mathcal{F}$-measure are seen at $P \approx 5$, similar to the NS-OMP and slightly smaller than those observed for GBF-NNLS in the *k*-sparse experiments.

5.4 Discussion

In this chapter the use of stepwise methods in the context of AMT was extended beyond the perspective of greedy OMP-based algorithms previously seen in the AMT literature and in Chapter 4. The motivation for using stepwise methods was the observation, described in a simple three-atom toy problem, that OMP can easily select an inactive atom, due to the coherence between atoms representing notes that are harmonically related, even in the noiseless case.

Several sparse stepwise algorithms from the literature were introduced and some steps taken to derive non-negative variants of these methods were described. These modifications were seen to be tricky, leading to the proposal of a backwards elimination strategy, BF-NNLS, that used a NNLS solution vector as its initial state. Similar performance was noted for the stepwise algorithms and proposed BF-NNLS in *k*-sparse experiments, with NNLS itself seen to be optimal. However, improved AMT results were seen when a proposed modified sparse cost function was applied using the BF-NNLS approach.

Group sparsity was then introduced to the BF-NNLS using a similarly modified group sparse cost function. In this case a large increase in AMT performance was seen, again validating the use of the subspace dictionaries with algorithms that explicitly enforce group sparsity. The results seen for the GBF-NNLS were better than the state-of-the-art benchmark given by β-NMD as described in (§3.5).. However, the largest part of these improvements are seen to be due mostly due to the transforms used. When Transform $\mathbb{E}1$ is used, as in the benchmark experiments, the observed improvement is slight.

While the GBF-NNLS algorithm is seen to perform well in the given context, further improvements may be possible. GBF-NNLS, as outlined, only considers atoms that are active in the initial NNLS decomposition. Allowing atoms that themselves are inactive, but belong to groups that are considered active, to enter the sparse support when other eliminations take place may provide some further enhancement of AMT.

While an accelerated group elimination subprocedure was proposed, the initial NNLS decomposition is seen to be relatively slow when subspace dictionaries are used. A bi-directional approach may be more computationally attractive in this case. A fast pursuit algorithm such as F-NS-OMP, which is seen to perform similarly to GT-NNLS in the *k*-sparse experiments, could be used for forward selection in tandem with stepwise optimal backwards elimination steps. An exploration of different strategies to select the direction at each iteration is probably required.

The advantages of the BF-NNLS approach are seen in a structured sparse setting, and a further development on this approach, undertaken by a visiting student at C4DM, is described further in [88], where temporal structure is also considered. However, the most suitable application for this type of approach may be in a multi-instrument setting, where matching pursuit methods have previously been reported to be useful [76]. More complex elimination criteria may need to be considered in this case.

In the next chapter, the use of greedy and stepwise decompositions for AMT is further extended using a molecular sparse approach, allowing temporal structure to be considered in a simple manner.

Chapter 6

Molecular sparsity

In the previous two chapters the use of stepwise methods for Automatic Music Transcription (AMT) was described. The decompositions were performed framewise, with each spectrogram frame decomposed independently of other frames. Greedy methods were seen to be quick, with reasonable overall performance. However, selection errors were often observed to cause correctly detected notes to be represented in a temporally fractured manner, with spurious omissions. Time continuity is considered an important element in musical signals and much research in musical signal processing has tried to leverage the fact that signal elements tend to be time continuous. Contrary to the greedy methods, Non-Negative Least Squares (NNLS) decompositions were seen to derive representations in which the time continuity in signal elements tended to be maintained, a state that was unchanged using subsequent backwards elimination of pitch-time points. When NNLS or or other non-negative spectrogram decompositions are used for the purpose of AMT, thresholding is often performed to ascertain the active signal elements [133]. An alternative approach to simple thresholding is to perform tracking of signal elements, introducing dependency between spectrogram frames. This can be performed in a probabilistic framework using, for instance, Hidden Markov Models (HMMs) to post-process an activation matrix [5] or probability distribution [104]. Alternatively, temporal dependency can be introduced by incorporating appropriate penalty terms into a cost function used in a decomposition method [134] [106], or as a prior in a probabilistic approach [39] [9].

Several structured sparse approaches have considered time continuity in audio signals. Kowalski et al propose [71] [70] neighbourhood sparse systems for audio denoising, using an approach called Windowed Group Lasso (WG-Lasso) which can be solved using an iterative shrinkage approach. WG-Lasso generalises the Group Lasso [137] by allowing overlapping groups when orthogonal atoms are considered. Sprechmann et al [118] propose the Collaborative Hierarchical Lasso (CHi-Lasso) which uses group sparsity at the frame level, while incorporating correlation between frames by using hierarchically structured norms. CHi-Lasso is applied to many applications including an instrument recognition task [118]. However, a limitation of these optimisation based approaches is that the groups are required to be predefined.

Molecular sparsity, first proposed by Daudet, [30] refers to greedy algorithms which select several structurally related atoms at each iteration. In this approach, an initial atom is selected based on a correlation measure, in typical greedy fashion, and a local search is performed to gather related atoms to form a molecule. In this way, grouping is performed on the fly. While other possible applications of this approach may be considered, it is found in the literature that molecular sparsity has tended to refer to representations that introduce structure in audio signals, with temporal continuity most commonly considered. Originally, Molecular Matching Pursuit (MMP) [30] was proposed to create signal representations in which the tonal and transient elements of an audio signal were represented in a Modified Discrete Cosine Transform (MDCT) and a Discrete Wavelet Transform (DWT), respectively, for the purpose of audio coding. MMP performs this separation by extracting a molecule of either stationary tonal, or time-localised transient, atoms at each iteration. Structure is favoured in the atom selection stage through locality based coefficients, using either a time-smoothed MDCT coefficient matrix or summed coefficients of DWT branches. A Matching Pursuit type algorithm selects either a tonal or transient component at each iteration, based on the maximum localised coefficient and the selected component is used as the start point of a local search for other structurally related atoms. If a transient atom is selected, the corresponding wavelet tree is pruned using a threshold and connectivity criteria, and the remaining atoms form a molecule that is added to the sparse support. Of particular interest here, when a tonal atom is selected a search, or tracking, is performed backwards and forwards through time frames along a narrow frequency window, until an low energy threshold is reached. All atoms found during the tracking step are added to the molecule and enter the sparse support simultaneously.

Meta-Molecular Matching Pursuit (M3P) [72] employs a similar tracking methodology to that used for tonal tracking in MMP. The M3P approach was proposed for the purpose of pitch tracking and harmonic structure is also considered. Rather than selecting one Gabor atom at each time frame, several harmonically related frequency atoms are selected together, in a similar fashion to that used for Harmonic Matching Pursuit [53]. Similar to MMP, structure is favoured in the initial atom selection through the use of a harmonicity index, and the initially selected harmonic atom is then tracked backwards and forwards through time until an energy threshold is met. A similar approach is also used in [76] where a dictionary of pitch and instrument labelled harmonic atoms is used for the purpose of forming a mid-level representation similar to a piano roll, while also allowing for pitch variance, of a multi-instrument signal. Again a molecular tracking method is used, differing from that used in [30] by also searching through adjacent pitches, while the atoms found through tracking are constrained to be labelled similarly, in terms of instrument, to the initially selected atom.

In this chapter an alternative molecular sparse strategy for introducing temporal continuity to greedy methods for the purpose of AMT is proposed. Following an initial NNLS spectrogram decomposition, molecules are defined through a clustering step. A molecular variant of OMP is then proposed to decompose a spectrogram using these predefined molecules. Experiments using this molecular OMP are performed with frame-based and event-based analyses. While promising results are seen with this molecular approach, analysis of an oracle transcription is then proposed, providing further insight to spectrogram decomposition-based AMT. A molecular norm is then defined that affords easy adaptation of other approaches, including backwards elimination, to perform similar molecular decompositions. An experimental comparison of these approaches is undertaken before the chapter concludes with a discussion section.

6.1 Molecular methods with non-negative dictionaries for AMT

Molecular methods such as MMP [30] and M3P [72] introduce a temporal element to sparse representations through the use of tracking. Some initial experiments were run that sought to employ such a tracking approach for the purpose of AMT. However, tracking was seen to be problematic in this context as it was observed that many molecules were extended far beyond their natural length, either before a note commenced or after a note had ceased, and often in both directions. This extension problem was particularly common at early iterations of the tracking based molecular approach and can be explained in terms of dictionary coherence.

In the case of the MMP, a Modified Discrete Cosine Transform (MDCT) dictionary was used for capturing tonal elements of the decomposition. The MDCT is a Fourier-related transform in which atoms of different frequency that share the same time domain are orthogonal to each other. However, in the context of AMT, the dictionary is coherent due to the non-negativity and harmonic structure. In particular coherence exists between atoms representing notes that are consonant, and more likely to be played in proximity to each other (§2.1.2). In the context of tracking, this can lead to high signal projection values for selected atoms being observed beyond the ends of notes, resulting in the observed temporal overestimation. Hence, an alternative molecular approach is required.

It was previously observed that time continuity was reasonably maintained in NNLS decompositions. A greedy molecular approach is proposed that attempts to leverage this desirable feature by predefining the molecules. An initial piano roll, Γ, is derived by thresholding

$$[\Gamma]_{l,t} = \begin{cases} 1, \text{ if } [\mathbf{H}]_{l,t} > \lambda \\ 0, \text{ otherwise} \end{cases} \tag{6.1}$$

the (group) coefficient matrix $\mathbf{H}$ where λ is a defined threshold.

Clustering of temporally-adjacent pitch-similar active atoms in the piano roll Γ is performed to construct molecules. This clustering can be seen as similar in nature to the agglomerative clustering approach of [122] for molecular decompositions, where atoms supported in an OMP decomposition are clustered according to a similarity measure. The set of molecules $\mathcal{M} = \{\mathcal{M}^{(m)}\}$ is defined with each molecule represented by a tuple:

$$\mathcal{M}^{(m)} = (l^{(m)}, \tau_0^{(m)}, \tau_\infty^{(m)}) \tag{6.2}$$

where $l^{(m)}$ denotes the group membership of the molecule thereby relating the pitch of all atoms in the molecule while $\tau_0^{(m)}$ and $\tau_\infty^{(m)}$ represent the start and end points of the molecule, respectively, with all intermediate time-pitch points active. Otherwise put, the mth molecule contains only active points: $\Gamma_{l^{(m)},\tau} = 1$ where $\tau_0^{(m)} \leq \tau \leq \tau_\infty^{(m)}$, and points just beyond the molecule ends, $\Gamma_{l^{(m)},\tau_0^{(m)}-1}$ and $\Gamma_{l^{(m)},\tau_\infty^{(m)}+1}$ are inactive.

To decompose a spectrogram using the molecules defined in (6.2), an algorithm referred to as Molecular NN-NS-OMP (M-NS-OMP) is proposed. M-NS-OMP uses a similar atom selection criteria to MMP. Notable differences between the two algorithms include the backprojection step

Algorithm 6.1 Molecular NN-NS-OMP (M-NS-OMP)

```
**Input**
    $\mathbf{D} \in \mathbb{R}^{M \times N}, \quad \mathbf{S} \in \mathbb{R}^{M \times T}, \quad \Gamma, \quad \mathcal{M}, \quad \mathcal{L}, \quad \kappa$
**Initialise**
    $i = 0; \quad \mathcal{J} = \{\mathcal{J}^{(t)} = \{\}\}; \quad \hat{\Gamma} = 0^{L \times T}\mathbf{;} \quad \mathbf{R}^0 = \mathbf{S}\mathbf{;} \quad \mathbf{X} = 0^{N \times T}$
**repeat**
    $i = i + 1$
    $\mathbf{y}_t[l] = \arg\min_y \|\mathbf{r}_t^i - \mathbf{D}[l]\mathbf{y}\|_2^2 \quad \forall (l,t) \quad s.t. \quad l \in \Gamma_t$
    $\Phi_{l,t} = \|\mathbf{D}[l]\bar{\mathbf{y}}_t[l]\|_2$
    **Calculate** $\Theta$ (6.3) **and** $\xi_m \, \forall m$ (6.4)
    $\hat{m} = \arg\max_m \xi_m$
    **For** $\tau = \tau_0^{(\hat{m})} : \tau_\infty^{(\hat{m})}$
        $[\hat{\Gamma}]_{l^{(\hat{m})},\tau} = 1$
        $\mathcal{J}^{(\tau)} = \mathcal{J}^{(\tau)} \cup \mathcal{L}^{l^{(\hat{m})}}$
        $\mathbf{x}_{\mathcal{J}^{(\tau)},\tau} = \arg\min_t \|\mathbf{s}_\tau - \mathbf{D}_{\mathcal{J}^{(\tau)}}\mathbf{x}\|_2^2 \quad s.t. \quad \mathbf{x} \geq 0$
        $\mathbf{r}_\tau^i = \mathbf{s}_\tau - \mathbf{D}\mathbf{x}_\tau$
    **End For**
**until Stopping condition**
**Output** $\hat{\Gamma}$(pianoroll)
```

of OMP and the use of group sparsity at the frame level, both seen to improve AMT performance using greedy methods (§4.1) and the use of predefined molecules instead of a tracking-based approach. M-NS-OMP uses a straightforward OMP methodology, selecting one molecule at a time, adding this to the sparse support and backprojecting the support onto the signal. As well as the dictionary and signal, M-NS-OMP, outlined in Algorithm 6.1, accepts as input the piano roll, or (group) sparse support, Γ , the molecular structure, $\mathcal{M}$ (6.2), the group structure, $\mathcal{L}$ (§2.2.2), and a smoothing factor κ.

Initially, the NN-NS-OMP projection coefficient, Φ is calculated at each pitch-time point of each molecule, noting that $\Phi = \mathbf{D}^T \mathbf{r}$ when the atomic pitch dictionaries are used. Similar to the tonal molecule selection criteria used in MMP [30], a smoothing filter of length κ is applied to the coefficient matrix Φ to encourage structure in the decomposition :

$$[\Theta]_{l,t} = \sum_{t'=t-\frac{\kappa-1}{2}}^{t+\frac{\kappa-1}{2}} [\Phi]_{l,t'}/\kappa \tag{6.3}$$

and hence the vector of molecular coefficients

$$\xi_m = \|[\Theta]^{l^{(m)}}_{\tau_0^{(m)}:\tau_\infty^{(m)}}\|_\infty \tag{6.4}$$

is derived from the smoothed coefficient matrix, where $\mathbf{x}_{a:b}$ represents a subvector of $\mathbf{x}$ consisting of all entries in the range $[a,b]$. The molecule displaying the maximum value of ξ_m is then selected and its pitch index, $l^{(\hat{m})}$, is added to the sparse support, $\mathcal{J}^{(t)}$, at each time point τ in the time domain of the molecule. Backprojections need only be calculated in the time domain of the selected molecule, and similarly the group coefficients need only be recalculated within this domain, while the molecular coefficients (6.4) are calculated for molecules which temporally overlap with the currently selected molecule.

Experiments

Experiments were performed in order to test if M-NS-OMP resulted in improved AMT performance results relative to (G)T-NNLS. It is observed in the previous chapter that the best δ-thresholding performance for AMT using NNLS decompositions is achieved at $\delta_{opt} \approx 30$dB. A potential advantage of the proposed two step molecular approach is that a lower value of δ may be used prior to clustering. This should result in higher Recall, while it is hoped that the M-NS-OMP method may eject false positives introduced by the use of a lower threshold, thereby leading to improved AMT results.

AMT experiments were performed on the standard MAPS dataset (§3.1) using STFT spectrograms of dimension 1024, similar to Transform $\mathbb{S}1$ (§3.2). However in this set of experiments an overlap of 50% in the signal windows is used, rather than the 75% overlap used previously with Transform $\mathbb{S}1$. The subspace dictionaries (§3.3.2) of size $P \leq 5$ are used, and NNLS is used to perform the initial decomposition, with group coefficients calculated when $P > 1$. This resultant coefficient matrix is δ-thresholded with a parameter value $\delta = 0.01$, lower than the previously observed δ_{opt} by around 10dB, leading to the derivation of the initial piano roll Γ (6.1) on which molecular clustering is performed. For the M-NS-OMP algorithm a persistence factor of $\kappa = 5$ was used, and the stopping condition was set to $\xi_{\hat{m}} < 1$, parameters that were seen from initial experiments to provide good results across all piano pieces.

A frame-based analysis is performed, for which the $\mathcal{P},\mathcal{R},\mathcal{F}$ ensemble of metrics are recorded. For comparison, the results for the frame-based analysis are compared with the δ-thresholded NNLS using δ_{opt}, giving the optimum $\mathcal{F}$-measure and $\delta = 0.01$, the value used in the initial pre-clustering thresholding. In addition, an event-based analysis is performed. Note events are detected using a simple threshold-based event detector, similar to that used in [9] [8], in which an onset is detected when a threshold, δ^o, is exceeded and subsequently sustained for a minimum duration of 2 time bins. A true positive event is detected when an onset for a note is found within

P	Metric	(G)T-NNLS (δ_{opt})	(G)T-NNLS ($\delta = 0.01$)	M-NS-OMP
1	$\mathcal{P}$	**69.0**	44.7	**69.0**
	$\mathcal{R}$	62.0	**78.5**	73.4
	$\mathcal{F}$	65.7	57.0	**71.1**
2	$\mathcal{P}$	68.0	39.6	**68.8**
	$\mathcal{R}$	63.5	**81.4**	75.5
	$\mathcal{F}$	65.7	51.4	**73.5**
3	$\mathcal{P}$	67.5	37.1	**69.9**
	$\mathcal{R}$	62.6	**83.9**	77.8
	$\mathcal{F}$	65.0	51.4	**73.5**
4	$\mathcal{P}$	68.4	37.1	**71.8**
	$\mathcal{R}$	63.0	**84.4**	78.1
	$\mathcal{F}$	65.6	51.6	**74.8**
5	$\mathcal{P}$	69.9	37.3	**74.7**
	$\mathcal{R}$	63.3	**84.8**	78.8
	$\mathcal{F}$	66.4	51.9	**76.7**

Table 6.1: Frame-based transcription results for (G)T-NNLS and M-NS-OMP in terms of $\mathcal{P}, \mathcal{R}, \mathcal{F}$ metrics for different group sizes P.

1 time frame of the ground truth piano roll. As the spectrogram frame size is $46ms$ this results in a worst case tolerance of $92ms$. The threshold for the event detector was set to $\delta^o = 5.5$, which was seen from initial experiments to be a reasonable value across all group sizes. The event-based analysis is performed on the (G)T-NNLS decompositions with δ_{opt} and also the M-NS-OMP coefficient matrix.

The frame-based results of the experiments are presented in Table 6.1. Here it is seen that the M-NS-OMP algorithm outperforms the other approaches by a considerable margin, in terms of $\mathcal{F}$-measure. When the atomic pitch dictionaries are used, the T-NNLS is outperformed by M-NS-OMP by approximately 5%. When the subspace dictionaries are used, GT-NNLS transcription results are only seen to improve upon those of T-NNLS once, and by a small margin of less than 1%. Conversely M-NS-OMP performance is seen to improve with each increment in groupsize. An improvement of 5.6% in $\mathcal{F}$-measure is recorded when $P = 5$ relative to when the atomic pitch dictionary is used. Similarly, when $P = 5$ the improvement in $\mathcal{F}$-measure relative to the (G)T-NNLS is over 10%.

The results for (G)T-NNLS, with $\delta = 0.01$ are also recorded. In these experiments, M-NS-OMP can be considered a post-processing of these thresholded NNLS decompositions, on which the molecular clustering is performed. In this sense, comparing the results for M-NS-OMP and (G)T-NNLS demonstrates the effect of M-NS-OMP in isolation. In particular, the Recall using

	P				
	1	2	3	4	5
(G)T-NNLS	**74.7**	74.0	73.0	73.3	73.9
M-NS-OMP	75.6	76.7	77.1	77.7	**78.0**

Table 6.2: $\mathcal{F}$-measure for event-based analysis with (G)T-NNLS and M-NS-OMP decompositions used to perform AMT for different groupsizes P.

M-NS-OMP is limited by the Recall of the (G)T-NNLS decomposition, as no new atoms are activated. It is seen in Table 6.1 that Recall using M-NS-OMP comes within 6% of that of the (G)T-NNLS for all group sizes. Meanwhile Precision is greatly increased, leading to superior $\mathcal{F}$-measure in these frame-based AMT results. M-NS-OMP can be regarded to perform well, selecting many of the true positives contained in the presented molecules, while ignoring many of the false positives introduced by the lower threshold related to δ.

Results for the onset-based analysis are given in Table 6.2, where M-NS-OMP is seen to outperform (G)T-NNLS by 0.9% in terms of $\mathcal{F}$-measure when the atomic pitch dictionaries are used. The performance difference becomes greater when the subspace dictionaries are used, in which case event detection deteriorates for the (G)T-NNLS. For M-NS-OMP, a small increment in $\mathcal{F}$-measure was observed with each increase in the groupsize.

6.2 AMT analysis using oracle decomposition

Many possible sources of error exist when spectrogram decompositions are used to perform AMT, some of which are innate to the signal and the transform. The dataset used here is recorded live in a studio, with audible echoic effects in the recordings. The frequency response of both the room and the microphone used characterise the recording. However the possibility of error resulting from these sources is limited in the experimental setup used in this thesis as the dictionaries were learnt from isolated notes recorded in the same environment piano pieces in the dataset. Further introductions to err exist in the use of STFT and ERBT time-frequency representations which produce sidelobes due to spectral leakage and have limited time-frequency resolution, while the use of a non-negative spectrogram leads to loss of phase information. However, the approximations assumed in spectrogram decompositions are considered reasonable, in the context of this thesis, and the signal transforms are considered a black box operation.

Indeed, the focus of this thesis is to improve such spectrogram decompositions, and some progress has been seen by using a group sparse model with dictionaries formed from a union of

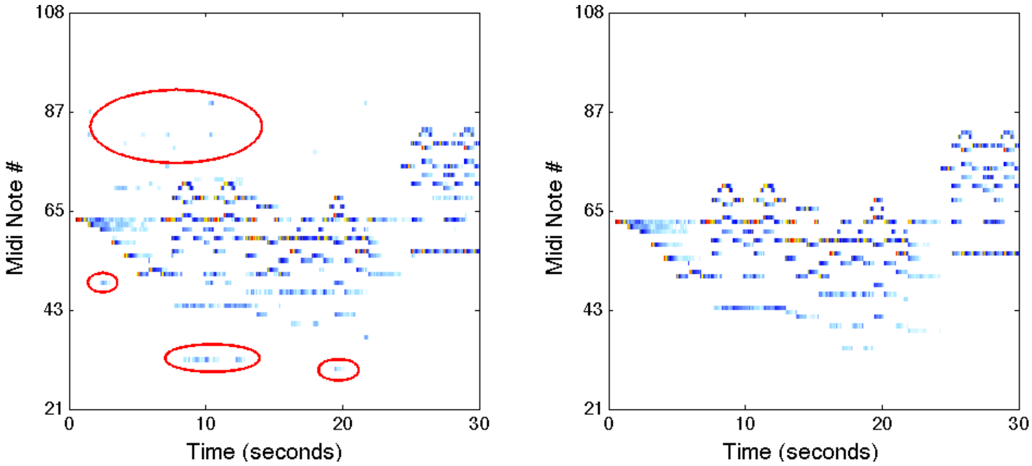

Figure 6.1: Comparision of AMT output using M-NS-OMP method (left) and Oracle(right) with some errors found using M-NS-OMP circled.

pitched subspaces. It is considered that the subspace dictionaries provide better modelling of the evolving note spectra, and several stepwise methods, each of which give sparse subset NNLS solutions, have been proposed that leverage this advantage. While some improvement have been noted, several problems are still obvious. Low energy in the tails of long notes is difficult to recognise leading to false negatives, while harmonic overlaps still seem to cause most of the false recognitions in terms of a frame-based analysis. Furthermore, the simple threshold-based onset detection approach described is perceived to be suboptimal.

It is proposed to perform an oracle decomposition, thereby omitting incorrect detections, in order to try to better understand the problem at hand. The MAPS dataset (§3.1) is provided with a ground truth, from which $\Gamma_t^{\mathcal{O}}$, the ground truth atomwise sparse support at a time frame indexed by t can be derived. The oracle decomposition, $\mathbf{X}^{\mathcal{O}}$, is then simply calculated using NNLS:

$$[X^{\mathcal{O}}]_{\Gamma_t^{\mathcal{O}},t} = \min_{\mathbf{x}} \|\mathbf{s}_t - \mathbf{D}_{\Gamma_t^{\mathcal{O}}}\mathbf{x}\| \quad s.t \quad \mathbf{x} > 0 \tag{6.5}$$

from which the group coefficients $[H^{\mathcal{O}}]_{l,t} = \|\mathbf{D}[l]\mathbf{x}_t^{\mathcal{O}}[l]\|_2$ are calculated when a subspace dictionary is used. This oracle decomposition can then be compared with the output from a decomposition algorithm.

A graphical comparison of an oracle decomposition and the corresponding output from the M-NS-OMP algorithm is shown in Figure 6.1. Here, the two decompositions are seen to be quite similar in terms of coincident energy levels, while some false positive errors are observed in the output of the M-NS-OMP, a pattern which is seen across many pieces. Such an oracle decom-

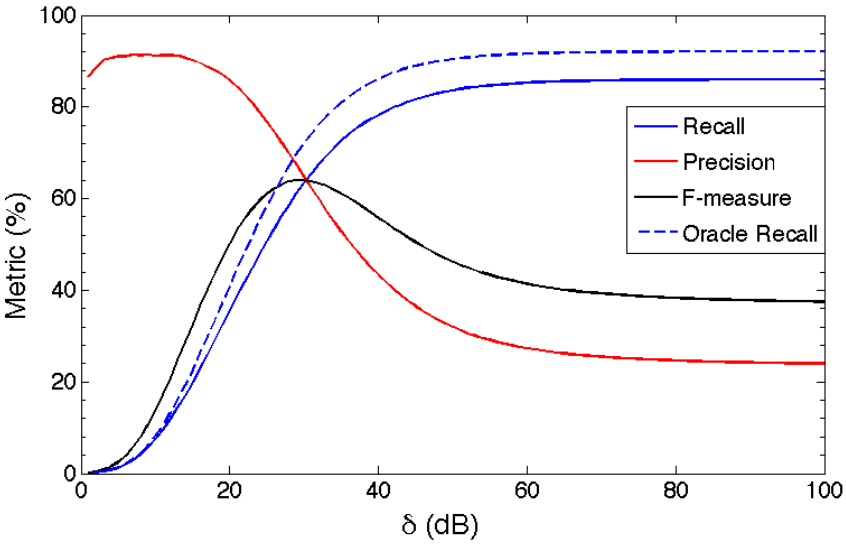

Figure 6.2: Precision, Recall and F-measure for AMT experiments using T-NNLS relative to δ thresholding parameter for Transform $\mathbb{S}1$ with atomic pitch dictionary. Recall for oracle transcription using same setup compared.

position provides a best case transcription using NNLS-based algorithms, a class that includes stepwise methods. Further analysis is undertaken, corresponding to the two forms of analysis performed on the spectrogram decompositions. First, the effects of δ-thresholding on Recall are observed, before the effectiveness of the onset detection system employed is examined.

Energy Based Thresholding

In NNLS-based AMT experiments described in the previous chapter δ-thresholding (§3.4.1) of the coefficient matrix was used to produce a piano roll. An optimum value of $\delta \approx 27-32$dB empirically asserted in [133] for AMT spectrogram decomposition methods was confirmed in the experiments undertaken. The δ-thresholding approach is also employed in the proposed molecular approach, prior to the clustering stage. The Recall after this thresholding places an upper bound on the Recall of M-NS-OMP as no new time-pitch points are activated. A lower value of $\delta = 40dB$ was used in the M-NS-OMP experiments described in the previous section, affording greater Recall, while the greedy approach was seen to omit many of the false elements detected at this low threshold.

It is worth considering the possible Recall relative to the thresholding parameter δ as a general investigation in the context of AMT. To this end, Figure 6.2 shows the Recall of oracle spectrogram decompositions of all pieces in the standard dataset. These decompositions were performed with Transform $\mathbb{S}1$, using the atomic pitch dictionary and δ-thresholding was per-

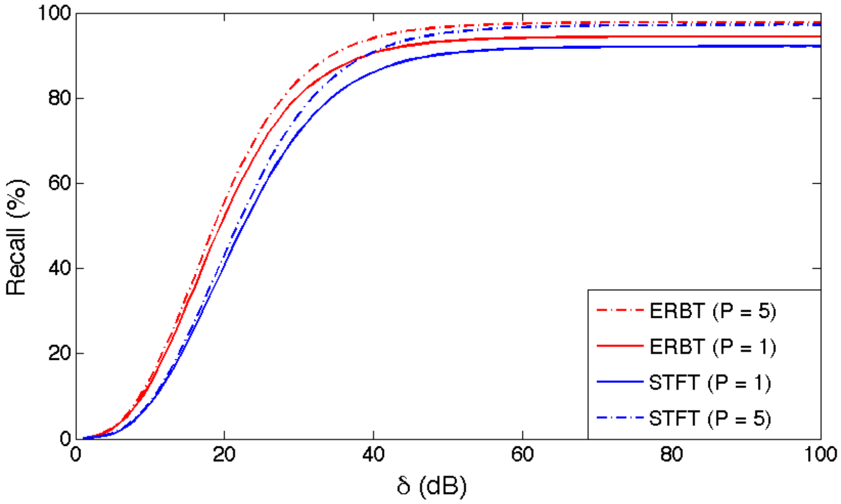

Figure 6.3: Recall for oracle decompositions using STFT, $\mathbb{S}1$, and ERBT, $\mathbb{E}2$, with atomic pitch dictionaries and subspace dictionaries with $P = 5$ relative to thresholding parameter δ.

formed for a range of values of $\delta \in 1,...,100\}$dB in steps of 1dB. The $\mathcal{P},\mathcal{R},\mathcal{F}$ measures are also shown in Figure 6.2 for the equivalent NNLS decompositions, for which the $\mathcal{F}$-measure is seen to peak at 30dB at a value of 64.0% with similar values of Recall and Precision. At higher values of δ the Recall increases to a maximum of 86% while Precision decreases at a faster rate. The oracle Recall is seen to be larger than that of the NNLS decompositions at all points and reaches a maximum of 92.1%. This maximum Recall is achieved at $\delta = 100$dB, while a flattening is observed at around 70dB. This can be considered a very low relative threshold, and it is seen that Precision in the NNLS decompositions is below 30% at such values of δ. Considering that this is an oracle decomposition suggests a limitation in the methodology for the purpose, in that the maximum achievable recall falls short of 100% by what would seem a considerable amount.

Varying AMT results have been observed in previous chapters when different spectrogram transforms have been employed, while improvements have been recorded with the use of the subspace dictionaries. A similar comparison is performed on oracle decompositions to test if higher Recall than that observable in Figure 6.2 may be possible with spectrogram decompositions. Oracle decompositions are performed on the standard dataset with both the $\mathbb{S}1$ and $\mathbb{E}2$ transforms using the atomic pitch atom dictionaries and subspace dictionaries with $P = 5$. Group coefficients are calculated for the subspace dictionary-based decompositions and δ-thresholding is performed on all coefficient matrices, again for the range $\delta \in \{1,...,100\}$dB. The Recall is recorded at each value of δ and plotted in Figure 6.3 where the Recall using the ERBT is seen to rise faster than with the STFT for the atomic pitch dictionaries, and achieve a slightly higher maximum, with an increase in possible Recall up to 94.4%. Meanwhile, Recall improves significantly when the subspace dictionaries are used. In the case of the STFT, a Recall of 97.1% is achieved, which increases to 97.9% for the ERBT. If the vagaries of the pitch-time tiling of the spectrograms and ground truth piano rolls (§3.1) are considered, the Recall in these cases may be considered very close to optimal. This observation, coupled with the improved results seen in group sparse AMT experiments in previous chapters, validates the subspace approach for note representation in decomposition-based AMT. Indeed, it can be conjectured that a subspace dictionary may be necessary if polyphonic piano transcription is to be considered a solved problem using decomposition-based methods.

Onset Analysis

An event-based analysis of AMT using M-NS-OMP and (G)T-NNLS was performed in experiments presented earlier in this chapter. Note events were detected using a simple threshold-based onset detection system that detected an onset when a threshold value was surpassed and sustained for a minimum duration. A true positive was flagged when this trigger happened within one time frame of a ground truth onset of the same note, giving a worst-case temporal resolution of 92*ms*. This onset detector was proposed in [8] [9] for the purpose of event detection. However, the effectiveness, or otherwise, of this simple approach has not previously been tested. An oracle transcription affords the opportunity to simply test the effectiveness of the onset detection system itself. This analysis was performed on oracle decompositions of the standard dataset, again using a spectrogram similar to Transform $\mathbb{S}1$ with a 50% overlap between time frames. The atomic pitch dictionary and a subspace dictionary with $P = 5$ were used, while the threshold used to detect an onset, δ^o, was varied between 0 and 7 in steps of 0.1 for both dictionaries. In order to trigger a note detection, the threshold δ^o had to be surpassed for a minimum of two time frames, similar to the experiments described earlier in this chapter.

The results of the onset detection on the oracle decomposition using the subspace dictionary are shown in terms of Precision, Recall and $\mathcal{F}$-measure in Figure 6.4 for a range of values of δ^o. The $\mathcal{F}$-measure from the AMT experiments is also shown, for comparison. The $\mathcal{F}$-measure for the oracle is seen to be highest at very low values of δ^o, where the molecular decomposition is seen to be low. This is to be expected due to low-value false detections in the coefficient matrix of the AMT decomposition triggering onsets. Overall, an increase of 7% in $\mathcal{F}$-measure is seen using the oracle, which reduces to around 4% in the locality of δ^o_{opt}, where the peak performance

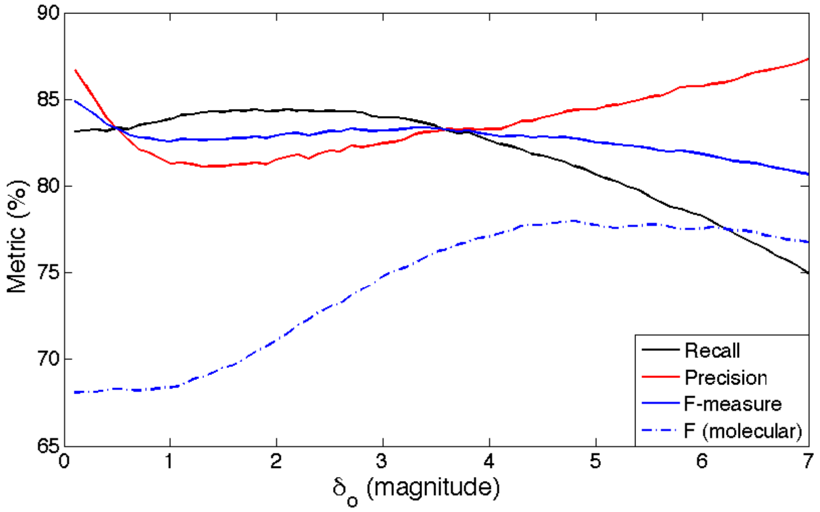

Figure 6.4: Evaluation of onset detector using oracle decomposition of transform $\mathbb{S}1$ with subspace dictionary ($P = 5$). $\mathcal{P},\mathcal{R},\mathcal{F}$ values plotted against values of the onset threshold δ^o ranging from [0.1, 7.0] in steps of 0.1. $\mathcal{F}$-measure of event detection applied to M-NS-OMP decomposition shown for comparison.

of the molecular method is recorded, suggesting that the molecular approach performs quite well relative to what can be expected. Further comparison is offered in Table 6.3 where the event-based results of the oracle and molecular transcriptions in terms of the optimal $\mathcal{F}$-measure and corresponding $\mathcal{P},\mathcal{R}$ and δ^o_{opt} are given for both dictionaries employed. Here it can be seen that the use of the subspace dictionaries is seen to afford improved performance in both oracle and molecular approaches, while the better performance recorded for the oracle decomposition is observed to result mostly from improved Precision.

The event detection results for the oracle transcription suggest the performance of the onset detector is far from optimal. However, the difficulty of the Disklavier datasets from the MAPS database in terms of onset detection has previously been observed in [12], where state-of-the art onset detection results for polyphonic piano recordings are given using a recurrent neural

	Oracle		Molecular	
	$P = 1$	5	1	5
$\mathcal{P}$	86.9	**92.4**	77.8	78.7
$\mathcal{R}$	76.6	**77.6**	73.5	77.2
$\mathcal{F}$	81.4	**84.4**	75.6	78.0
δ^o_{opt}	0.1	0.1	5.5	4.8

Table 6.3: Optimum event-based AMT results using M-NS-OMP and oracle decompositions for dictionaries with groupsizes $P \in \{1,5\}$.

network. The onset detection results using the molecular and oracle methods with subspace dictionaries are seen to be similar to those given in [12] for the MAPS Disklavier, while noting that direct comparison cannot be performed as a smaller time resolution and a larger dataset are used in [12].

However, careful consideration of the onset detection problem in spectrogram decompositions could yet yield improvements. Close inspection of the individual oracle decompositions reveal repetitive systematic flaws in the onset detection, as graphically described in Figure 6.5. False positives are often found when a sustained note is retriggered by oscillation around the threshold value, behaviour which is often found in the presence of other note onsets and may be due to transient signal elements effecting the smoothness of the decomposition across time. Several common types of false negative were found. It is observed that a note that is replayed, with minimal time between the offset of the original event and the onset of the following event, often produces a false negative as the observed coefficient may not have fallen below the threshold value. When several notes onset simultaneously, onsets may not be detected for all of these notes. A tendency for lower pitched notes not to trigger an onset event in the detection system is also noticed. Further to this some timing errors are found, where a false negative and a false positive are closely spaced.

6.3 Further molecular algorithms

In the first section of this chapter molecular decompositions were performed using an OMP-based algorithm with a selection criteria based upon that used for tonal elements in the MMP [30]. This selection criteria incorporated a smoothing of the projection matrix (6.3), which is important in the context of MMP, which seeks to capture temporally localised transient elements with a wavelet transform. Smoothing the tonal elements in this scenario encourages spurious tonal elements found around transients to be ignored. However, it may be possible that smoothing is not particularly relevant in the case of AMT, and employing a different selection criteria in a molecular OMP-based approach will reduce the number of input parameters, and may possibly yield improved performance.

An alternative molecular selection criteria is proposed using the norm of the projection coef-

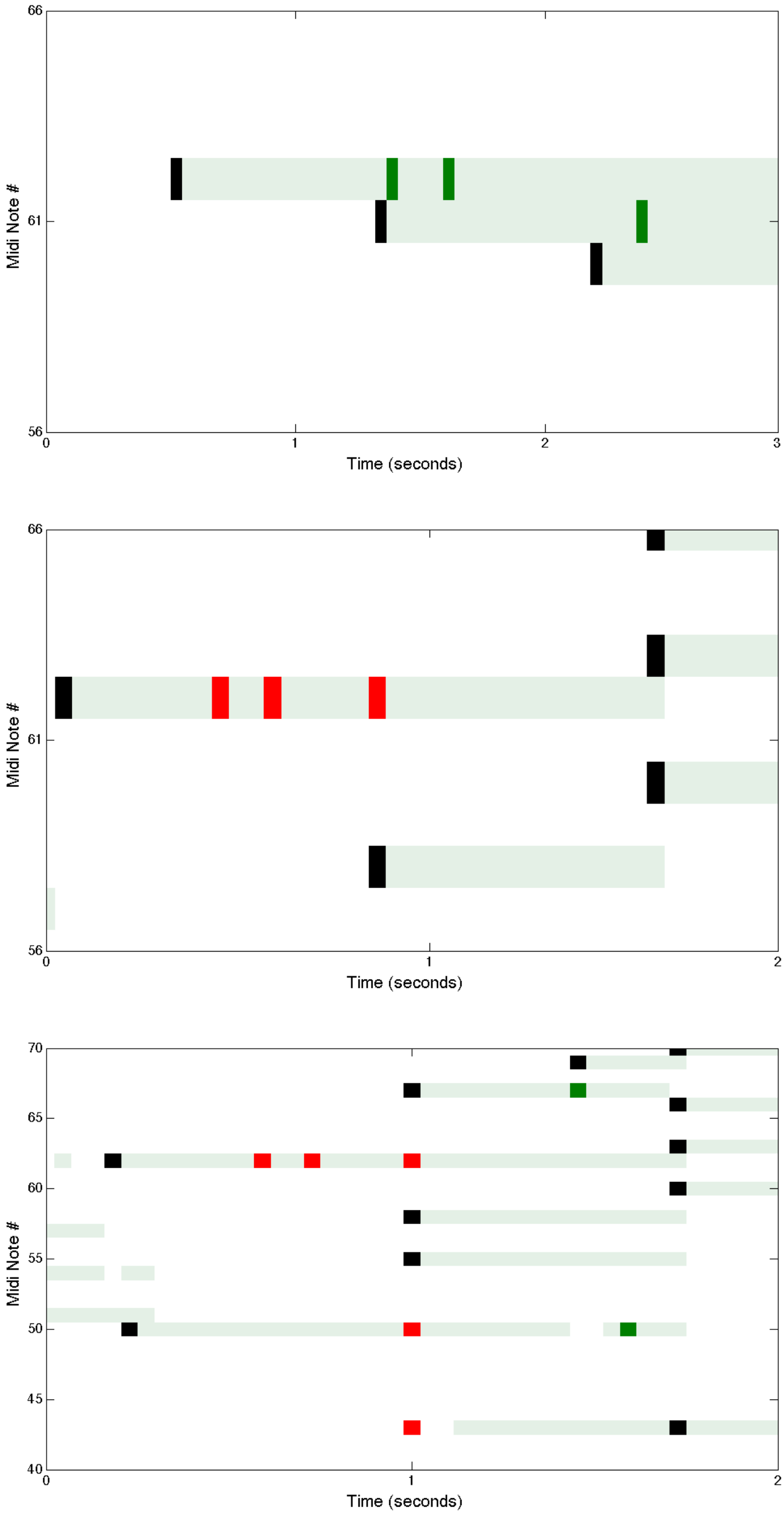

Figure 6.5: Some common problems with onset detection system. True positives in black, false positives in green, false negatives in red. Retriggered onsets (top), repeated notes not triggering onsets (middle) and untriggered onsets when many notes onset simultaneously (bottom).

Algorithm 6.2 Molecular Hard Thresholding (MHT)

```
Input
  $\mathbf{H} \in \mathbb{R}^{N \times T}, \quad p, \quad \delta_{\mathcal{M}}, \quad \mathcal{M}$
Initialise
  $\Gamma = 0^{N \times T}$
Perform thresholding
  $\xi_m = \|\mathcal{M}^{(m)}\|_{\mathcal{M}p} = \|\mathbf{H}^{l^{(m)}}_{\tau_0^{(m)}:\tau_\infty^{(m)}}\|_p \ \forall m$
  $\hat{m} = \arg\max_m \xi_m$
  $\lambda_m = \xi_{\hat{m}} \times \delta_{\mathcal{M}}$
  $\mathcal{J} = \{m | \xi_m > \lambda_{\mathcal{M}}\}$
Assign sparse support / pianoroll
For i = 1: $|\mathcal{J}|$
  For $\tau = \tau_0^{(\mathcal{J}^{(i)})} : \tau_\infty^{(\mathcal{J}^{(i)})}$
    $\Gamma_\tau^{l(\mathcal{J}^{(i)})} = 1$
  End For
End For
Output   $\Gamma$(pianoroll)
```

ficients at each time point of a molecule :

$$\xi_m = \|\mathcal{M}^{(m)}\|_{\mathcal{M}_p} = \|\Phi^{l^{(m)}}_{\tau_0^{(m)}:\tau_\infty^{(m)}}\|_p \tag{6.6}$$

where Φ is any matrix of projections onto the residual signal. When a subspace dictionary is employed (6.6) could also be considered in mixed-norm format such as $\|\mathcal{M}^{(m)}\|_{\mathcal{M}_{p,q}}$, where an ℓ_q-norm forms the group coefficient. However considering Φ in (6.6) as a group coefficient matrix has the same effect, whilst being simpler notation-wise. A new algorithm, Molecular-OMP (M-OMP) is proposed by simple replacement of the molecular selection criteria using the smoothed coeffcient matrix (6.3) in M-NS-OMP, by the molecular norm (6.6). Indeed, apart from the smoothing of the coefficient matrix, M-NS-OMP can be considered equivalent to M-OMP with a $\mathcal{M}_\infty$-norm used, as all other steps in the two algorithms are equal.

Other algorithms are easily derived using this molecular norm. A Molecular Hard Thresholding (MHT) algorithm described in Algorithm 6.3 is now proposed. MHT accepts as input a set of molecules, $\mathcal{M}$, a coefficient matrix $\mathbf{H}$, the molecular norm parameter, p, and a molecular threshold parameter, $\delta_{\mathcal{M}}$. MHT proceeds by calculating the molecular norm coeffcient for each molecule and calculating a molecular threshold $\lambda_{\mathcal{M}}$ using a δ-thresholding approach with the parameter $\delta_{\mathcal{M}}$. The set of active molecules $\mathcal{J}$ is then determined by thresholding out molecules such that $\xi_m < \lambda_{\mathcal{M}}$, before finally the piano roll, Γ, is assigned by activating all pitch-time points

Algorithm 6.3 Molecular Backwards Elimination (MBE)

```
Input
  $\mathbf{D} \in \mathbb{R}^{M \times N}, \quad \mathbf{S} \in \mathbb{R}^{M \times T}, \quad \mathbf{X} \in \mathbb{R}^{N \times T}, \quad \Gamma \in \{0,1\}^{L \times T}, \quad \mathcal{M}, \quad \mathcal{L}, \quad \lambda_{\mathcal{M}}, \quad p$
Initialise
  $\mathcal{J} = \{1, ..., |\mathcal{M}|\}; \quad \mathbf{R} = \mathbf{S} - \mathbf{D}\mathbf{X}$
  $\Delta_B \mathbf{r}_t^{[l]} = \frac{\mathbf{x}[l]^T \mathbf{x}[l]}{[(\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1}][l,l]} \quad \forall (l,t) \quad s.t. \quad \Gamma_{l,t} = 1$
  $\Phi_{l,t} = \sqrt{\|\mathbf{r}_t\|_2^2 + \Delta_B \mathbf{r}_t^{[l]}} - \|\mathbf{r}_t\|_2^2$
  $\xi_m = \|\Phi^{l(m)}_{\tau_0^{(m)}:\tau_\infty^{(0)}}\|_p \quad \forall m \in \mathcal{J}$
While $\lambda_{\mathcal{M}} > \min \xi$
  $\hat{m} = \arg\min_m \xi_m$
  $\mathcal{J} = \mathcal{J} \backslash \hat{m}; \quad \xi_{\hat{m}} = \infty$
  For $\tau = \tau_0^{(\hat{m})} : \tau_\infty^{(\hat{m})}$
    $[\Gamma]_{\tau, l^{(\hat{m})}} = 0$
    $\mathbf{x}_{\mathcal{J}^{(\tau)}, \tau} = \arg\min_t \|\mathbf{s}_\tau - \mathbf{D}_{\mathcal{J}^{(\tau)}} \mathbf{x}\|_2^2 \quad s.t. \quad \mathbf{x} \geq 0$
    $\mathbf{r}_\tau^i = \mathbf{s}_\tau - \mathbf{D}\mathbf{x}_\tau$
    $\Delta_B \mathbf{r}_\tau^{([l])} = \frac{\mathbf{x}[l]^T \mathbf{x}[l]}{[(\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1}][l,l]} \quad \forall\,(l,t) \quad s.t. \quad \Gamma_{l,\tau} = 1$
    $\Phi_{l,t} = \sqrt{\|\mathbf{r}_\tau\|_2^2 + \Delta_B \mathbf{r}_\tau^{[l]}} - \|\mathbf{r}_\tau\|_2^2$
  End For
  $\xi_m = \|\Phi^{l(m)}_{\tau_0^{(m)}:\tau_\infty^{(0)}}\|_p \forall m \in \mathcal{J}$
EndWhile
Output   $\bar{\Gamma}$(pianoroll)
```

contained in the set of active molecules.

The molecular norm also suggests that an elimination cost for a molecule can easily be derived. Hence a Molecular Backward Elimination (MBE) algorithm is now proposed. MBE is a stepwise algorithm, similar in a manner to M-OMP. However, in typical backwards elimination fashion, MBE starts with a sparse support in which all molecules, and their constituent pitch-time points, are active. Different molecular backwards elimination costs can be considered. However, the stepwise optimal approach using the ℓ_0-penalised least squares error norm was seen to perform well in the context of AMT in the previous chapter, and is adapted for the MBE algorithm, outlined in Algorithm 6.3.

MBE accepts as input the spectrogram, **S**, dictionary, **D**, the molecular and group structures $\mathcal{M}, \mathcal{L}$, the current binary support matrix, Γ, the NNLS coefficient matrix, **X** given the sparse support, a molecular threshold $\lambda_{\mathcal{M}}$ that is used as a stopping condition, and p, relating the value of the molecular norm $\mathcal{M}_p$ to be used. The residual, **R**, and the set of indices of supported molecules $\mathcal{J}$ are initialised. The backwards elimination cost $\Delta_B \mathbf{r}$ for the squared error norm is calculated, from which Φ, the elimination cost using the error norm is derived. Finally the molecular elimination cost ξ is calculated using the molecular norm applied to Φ. MBE then enters an iterative loop. At each iteration elimination of the molecule, indexed by $\hat{m}$, displaying the smallest elimination cost is performed and the index is removed from the set $\mathcal{J}$. At each pitch-time point in the eliminated molecule, the corresponding pitch-time point in the piano roll, Γ, is set to zero, and a new coefficient vector and residual are calculated with the downdated support. The elimination costs for each pitch-time point active at that time frame are calculated. When the pointwise elimination costs are calculated for all time frames contained in the molecule, the molecular elimination costs are recalculated. This iteration continues until the elimination cost of all molecules is greater than the molecular threshold $\lambda_{\mathcal{M}}$.

Experiments with MHT

Before a comparison of these newly proposed molecular algorithms is undertaken, some initial experiments are run to examine the comparative performance of the MHT algorithm for AMT using different transforms and dictionary subspace sizes. Improvements were observed when ERBTs and subspace dictionaries were used with other algorithms. While similar improvements may be expected here, it can also be hypothesised that the introduction of temporal continuity may attenuate the differences in performance, as a leveraging effect may be introduced when structured collections of pitch-time points are considered together. Using MHT, or any of the other proposed molecular approaches, an extra parameter, $\delta^{\mathcal{M}}$, is introduced. The initial NNLS coeffcient matrix is δ-thresholded using a parameter, δ_F, before clustering of molecules is performed as described earlier in the context of M-NS-OMP. The subsequent molecular decomposition is then performed using a second threshold, or stopping conditon, $\delta^{\mathcal{M}}$. An examination of the potential robustness, or otherwise, of the molecular approach through a thorough exploration of the $(\delta^F, \delta^{\mathcal{M}})$ threshold parameter space is undertaken in tandem with the relative comparison of different transforms.

Similar to the experiments in §6.1, a NNLS decomposition is (group) thresholded using δ^F prior to clustering molecules. This thresholding is performed for a range of values of $\delta^F \in \{6, ..., 60\}$dB in steps of 1dB. At each value of δ_F a molecular structure, $\mathcal{M}$, and corresponding sparse support Γ are derived and input to MHT, along with a value of $\delta^{\mathcal{M}}$, the molecular thresholding parameter. Experiments are run for a range of δ^M similar to that used for δ^F, and results are recorded for every $(\delta^F, \delta^{\mathcal{M}})$ pair. These experiments are performed for all transforms used in this thesis (§3.2), and using subspace dictionaries of groupsize $P \in \{1, ..., 7\}$. Group coefficients

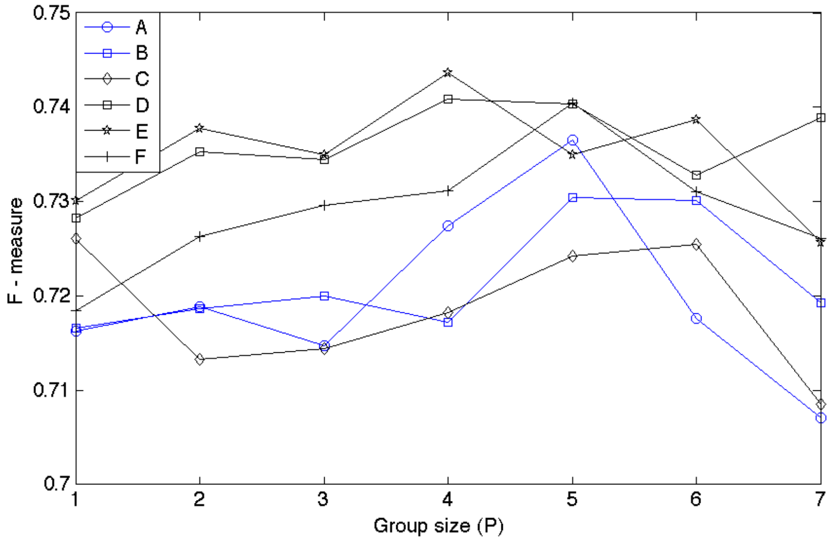

Figure 6.6: Maximum $\mathcal{F}$-measure achieved for each transform and groupsize using MHT.

when $P > 1$ are determined by the $\ell_{\perp}$-norm of the groupwise non-negative solution vector. The molecular norm coefficient is calculated using a $\mathcal{M}_1$-norm, summing all pitch-time coefficients in the molecule. The results are given for the tuple $(\delta^F, \delta^{\mathcal{M}})_{opt}$ giving the optimal $\mathcal{F}$-measure for each transform/groupsize pair.

The performance relative to groupsize and transform is graphically represented in Figure 6.6. Here, it is seen that performance using the different transforms is quite similar, with around 2% difference in $\mathcal{F}$-measure recorded between the best results for each transform, compared to around 5% in the case of the NNLS decompositions. Further to this the effect of the groupsize is similarly attenuated using MHT. The benefits of MHT are mostly observed for the STFTs, which are seen to perform better relative to the GBF-NNLS approach proposed in the previous chapter, while performance of the ERBTs deteriorate in a similar comparison. This flattening of relative performance may be very useful, as the computation of an STFT is faster by a large factor than that of an ERBT.

Contour maps of the $\mathcal{F}, \mathcal{P}, \mathcal{R}$ ensemble of metrics in terms of $(\delta^F, \delta^{\mathcal{M}})$ using MHT with Transform $\mathbb{E}2$ and groupsize $P = 5$. are shown in Figure 6.7, As can be expected, the Recall and Precision metrics display relatively monotonic relationships with the thresholding parameters. The $\mathcal{F}$-measure is seen to peak in a region which extends from 31dB to 54dB in terms of δ_F and from 24dB to 34dB in terms of $\delta_{\mathcal{M}}$, while a high value is maintained on a large L-shaped ridge. It is noted that the difference between contours in this diagram is 2% in which case any point in the

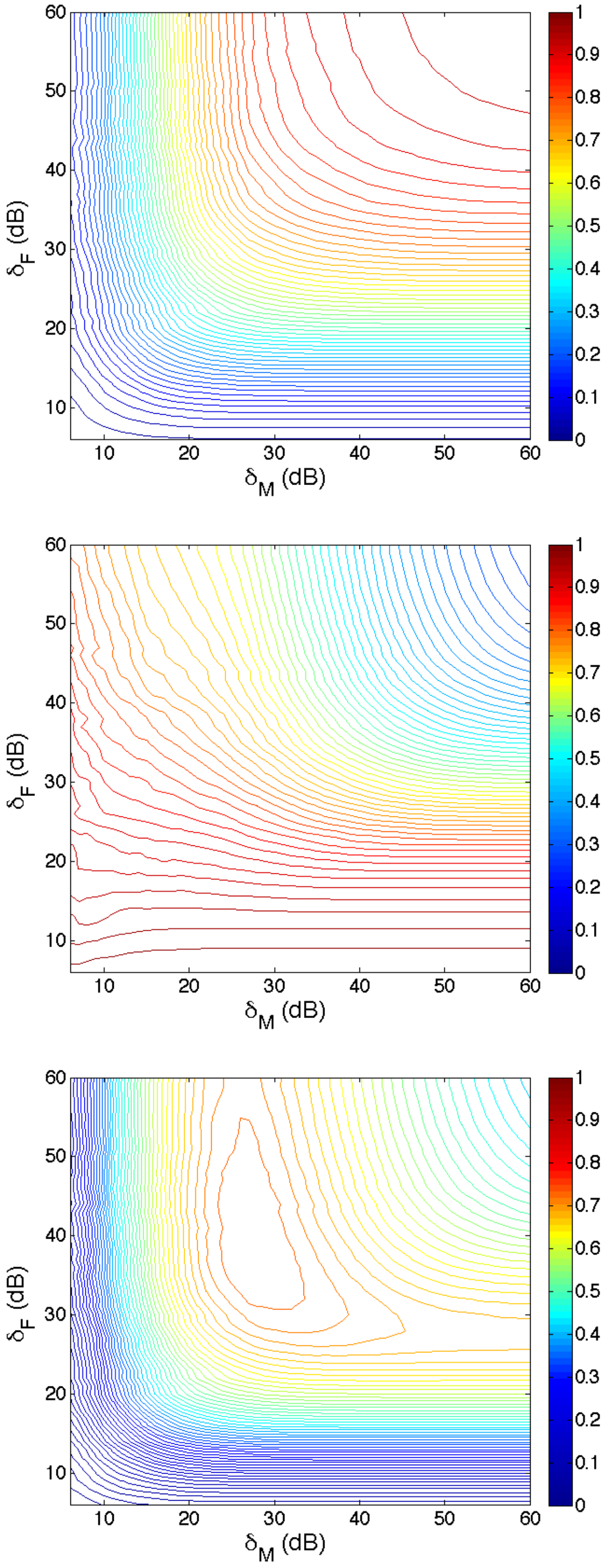

Figure 6.7: Contour maps showing Recall (top), Precision (middle) and $\mathcal{F}$-measure (bottom) using molecular thresholding with transform $\mathbb{E}2$ and groupsize $P = 5$, relative to the framewise and molecular thresholding parameters δ_F and δ_M.

	$P=1$		$P=5$	
	$\mathbb{S}1$	$\mathbb{E}2$	$\mathbb{S}1$	$\mathbb{E}2$
MHT (T)	71.6	72.8	73.5	74.0
M-OMP (T)	71.7	72.7	76.3	75.8
MBE (T)	**73.6**	**74.5**	**77.3**	**77.1**
MHT (BF)	73.0	74.0	**77.7**	77.5
M-OMP (BF)	70.8	72.6	76.8	77.4
MBE (BF)	**73.4**	**74.3**	77.3	**77.8**

Table 6.4: $\mathcal{F}$-measure for various molecular algorithms in Transforms $\mathbb{S}1, \mathbb{E}2$ for single atom dictionary and subspace dictionary (P=5). Coefficient matrices from two different decompositions are input to the molecular method: (T) denotes NNLS: (BF) denotes backwards elimination.

large peak region returns an $\mathcal{F}$-measure less than 2% below the optimum. This would indicate a potential robustness of MHT in terms of the thresholding parameters.

Further Experiments

Some further experiments are undertaken to compare the proposed M-OMP, MHT and MBE algorithms. In previous molecular experiments clustering was performed on a NNLS decomposition, and this is repeated here for all algorithms. Furthermore, clustering is also performed on a (G)BF-NNLS coefficient matrix. In the case of MHT, modifications are made to (G)BF-NNLS, omitting the stopping condition and iterating at each spectrogram frame until all pitch-time points are eliminated, while the elimination cost for each pitch-time point is recorded in the output coefficient matrix, which is input to MHT.

Experiments are run on the standard dataset (§3.1) using Transforms $\mathbb{S}1\&\mathbb{E}2$. Experiments are limited to the atomic pitch dictionaries and to subspace dictionaries with $P=5$. For the MHT, a similar setup to the previous set of experiments is used with decompositions performed at all values of $(\delta^F, \delta^{\mathcal{M}}) \in \{6, ..., 60\}$dB and the optimal $\mathcal{F}$-measure is recorded. For the M-OMP and MBE, a similar approach is taken, with decompositions performed for a similar wide range of values. However, in this case δ^F-thresholding is performed at intervals of 5dB.

The results of these experiments are given in Table 6.4. Here, little difference is observed in the results relative to the transform used for the decompositions, a scenario that, again, favours STFT. The use of the subspace dictionaries is effective giving an improvement of around 4% in all cases except for MHT with the NNLS coefficient matrix input. When the subspace dictionaries are used, inputting the GBF-NNLS coefficient matrix is seen to be superior to inputting the NNLS coeffcient matrix for all algorithms. Furthermore little difference is seen in the performance of the algorithms when the GBF-NNLS coefficients are input. However when the NNLS coeffcients are used there is a difference of over 3%, between the superior MBE and the results of MHT, with the M-OMP algorithm displaying performance intermediate to these two algorithms.

In the case of the atomic pitch dictionaries, some different effects are observed. The consistency observed between algorithms with the subspace dictionaries when the elimination coefficients are input is not observed in this case. A higher $\mathcal{F}$-measure is recorded for the stepwise algorithms when the NNLS coefficient matrix is input, while the elimination coeffcients lead to better performance when MHT is used. The best results are seen with the MBE algorithm for both types of input coefficient matrix. To summarise it would seem that the use of a subspace dictionary with backwards elimination performed, either pre- or post-clustering, results in the best performance.

6.4 Discussion

In this chapter, structured sparse methods such as neighbourhood sparsity and molecular sparsity, that introduce time continuity to sparse representations, were briefly introduced. Problems adapting such approaches to the AMT problem were noted, before an alternative molecular approach was proposed that took advantage of the time continuity often observed in NNLS decompositions.

Initially the M-NS-OMP algorithm was proposed, which performed greedy selection from a dictionary of predefined molecules. M-NS-OMP was seen to perform well both in terms of frame-based and onset-based analyses. An analysis of oracle decompositions was then performed, giving some insights to spectrogram decomposition-based AMT. In particular the use of the subspace dictionaries was validated, as a larger recall was seen to be possible using the oracle. Further to this the onset detector, while seen to perform reasonably with M-NS-OMP, was seen to reveal systematic errors when used on the oracle transcription, consideration of which may lead to improved onset detection.

Finally a molecular norm, affording simple adaptation of other common sparse approximation algorithms to a molecular context was defined. Experiments showed improved AMT, particularly when subspace dictionaries and backwards elimination were used. The performance enhancements were most notable in the STFT, where improvements relative to previous best AMT using GBF-NNLS of the order of 6% were seen. In the case of the ERBT, the observed performance enhancement was around 2%. Indeed the molecular approach is seen to equalise the transforms in terms of performance, which is advantageous due to the computational inexpensiveness of the STFT relative to the ERBT. It can probably be considered that the introduction of a temporal element to the spectrogram decompositions suppresses some large individual errors observable when the spectrogram frames are considered independently. As the STFT performs the worst in the case of independently considered frames, it benefits most from the introduction of temporal structure.

While improved results are seen in general, some large individual errors are occasionally observed, such as over-extended notes, that the molecular approach sought to counter. Ideally each molecule would represent one note-event. However, the molecular clustering approach used is very coarse, with little consideration given to the concept of a note-event, and several repeated notes may be contained in one molecule. This problem may be enhanced when a lower threshold is used pre-clustering, and ultimately limits the potential effectiveness of the molecular approach. It may be worth considering other molecular approaches. One such approach is seen in [88], where the molecules are ultimately delineated by an onset detector. Another possible approach might be to start with a higher threshold applied to the initial decomposition, to encourage separability of different note-events in a molecule. A molecular decomposition could then be performed with subsequent lowering of the framewise threshold, allowing active molecules to dilate whilst not merging. This dilation could be performed by further molecular clustering or alternatively, by using a framewise forward selection method constrained by connectivity to active molecules. This is an avenue of research that may be investigated at a later date.

However, in the context of this thesis, a change in direction is now taken. Until now, sparse and structured sparse decompositions have been performed in the context of AMT using stepwise approaches which perform ℓ_0-penalised NNLS approximations. However, cost functions other than the Euclidean are generally considered more effective in the context of musical signal processing. The following chapter considers some of the cost functions commonly used, and introduces some new cost functions to musical signal processing, while the focus on sparsity is maintained through consideration of penalised cost functions including group sparse approximations.

Chapter 7

Non-Negative Matrix Decompositions

Previous chapters have seen the exploration of stepwise methods for Automatic Music Transcription (AMT), with a particular focus on group sparse decompositions. In particular it was seen that using an NNLS decomposition with backwards elimination improved AMT, which is seen as an ℓ_0-, or $\ell_{\perp,0}$-penalised NNLS approximation. Typically in the sparse representations literature an ℓ_1 penalty is used, as this affords a convex relaxation of the problem [23]. The ℓ_1 penalty is also often applied in Non-negative Matrix Factorisation (NMF) approaches [56] [107].

The Basis Pursuit Denoising / LASSO problem has been adapted for the group sparse case. A mixed norm penalty term for grouped atoms is introduced in the Group Lasso [137] :

$$\mathbf{x} = \arg\min_{x} \|\mathbf{s} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_{p,q} \tag{7.1}$$

where the mixed norm is given by

$$\|\mathbf{x}\|_{p,q} = \left(\sum_{l} \left(\sum_{i \in \mathcal{L}(l)} |\mathbf{x}_i|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \tag{7.2}$$

with varying values of the tuple (p,q) used to effect different properties on the decomposition. For instance in the Group Lasso paper [137] an $\ell_{1,2}$ norm is used to effect sparsity within a group, while sparsity between groups is not enforced. Conversely, an $\ell_{2,1}$ norm is used [37] [71] when it is expected that atoms within a given group are active together, while few groups are active in a given coefficient vector. An Iterative Soft Thresholding algorithm is used to perform this optimisation in [71] using a proximal, or shrinkage operator

$$x_i \leftarrow x_i \times \max\left(0, 1 - \frac{\lambda}{\|\mathbf{x}[l]\|_2}\right) \quad s.t.\, i \in \mathcal{L}^{(l)} \tag{7.3}$$

for the thresholding step, with the authors noting that convergence is guaranteed when each group is internally orthonormal. The use of groups which are not orthonormal is considered in [38], who propose the use of a mixed norm using the orthogonal projection of the group

$$\mathbf{x} = \arg\min_x \|\mathbf{s} - \mathbf{D}\mathbf{x}\|_2^2 + \|\mathbf{h}\|_1 \tag{7.4}$$

where $h_l = \|\mathbf{D}[l]\mathbf{x}[l]\|_2$, thereby considering the correlations within each group.

While the penalised least squares approach is relatively well studied, it has previously been seen in the AMT and musical signal processing literature that cost functions other than the Euclidean distance result in better performance [115] [40] [133] when using Non-negative Matrix Factorisation (NMF) [74] algorithms. Typically, NMF refers to an unsupervised learning algorithm, in which both the dictionary and activation matrix are learnt using an alternating projection strategy. However, the NMF coefficient matrix updates can be used with a fixed dictionary, an approach referred to as supervised NMF [4], or alternatively Non-negative Matrix Decomposition (NMD) [32]. Indeed, when a fixed pitch-labelled dictionary with atoms which captures well the spectral shape of the instrument being played is used, superior results to those found using unsupervised NMF are to be expected [32] [133].

NMF using multiplicative updates was first proposed by Lee & Seung [74] who propose multiplicative update algorithms for the Euclidean distance and Kullback-Leibler (KL) divergence cost function:

$$\mathcal{C}_{KL}(\mathbf{s}|\mathbf{z}) = \sum_i s_i \log\frac{s_i}{z_i} - s_i + z_i. \tag{7.5}$$

Smaragdis and Brown [115] first proposed the use of NMF as a tool for AMT, and experiments with the cost functions given in [74] demonstrate superior performance when using the KL-divergence. Later the Itakuro-Saito divergence

$$\mathcal{C}_{IS}(\mathbf{s}|\mathbf{z}) = \sum_i \frac{s_i}{z_i} - \log\left(\frac{s_i}{z_i}\right) - 1 \tag{7.6}$$

was proposed for use with musical signals [40], in particular for use with the power spectrogram.

Notably, this divergence had inadvertently been used at an earlier date for AMT in [1] where it was derived by the authors who sought to accomodate a multiplicative noise model, again in the case of a power spectrogram. It is worth noting that these three cost functions can all be seen to perform maximum likelihood estimation with different distributions on the noise [40], with the Euclidean distance assuming a Gaussian distribution, while the KL and IS divergences assume a Poisson and Gamma distribution, respectively.

NMF algorithms using multiplicative updates have been proposed for many different cost functions, and a pattern is seen in the literature towards generalised divergences that allow the same algorithm to be used with varying parameters in order to effect different cost functions. For example the generalised β-divergence [24] [27] given by

$$\mathcal{C}_\beta(\mathbf{s}|\mathbf{z}) = \sum_i \frac{s_i^\beta}{\beta(\beta-1)} + \frac{z_i^\beta}{\beta} - \frac{s_i z_i^{\beta-1}}{\beta-1} \tag{7.7}$$

is seen to include the Euclidean distance ($\beta = 2$), while the Itakuro-Saito (IS) and Kullback-Leibler (KL) divergences are seen as limiting cases when $\beta \rightarrow \{0,1\}$ respectively. Considering that $\mathbf{z} = \mathbf{Dx}$ the multiplicative update [24] for β-NMD derived by using a fixed stepsize in a similar fashion to that used in [74] is given by

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes [\mathbf{D}^T(\mathbf{S} \otimes [\mathbf{DX}]^{[\beta-2]})] \oslash [\mathbf{D}^T[\mathbf{DX}]^{[\beta-1]}] \tag{7.8}$$

where $\otimes$ and $\oslash$ denote elementwise multiplication and division, respectively and $\mathbf{x}^{[a]}$ denotes elementwise exponentiation of $\mathbf{x}$ to the power of a. Indeed this update (7.8) reduces to the multiplicative updates given for Euclidean distance and KL-divergence in [74] and IS-divergence given in [1] [40] with the appropriate values of β. An exploration of the use of NMD with the β-divergence for the purpose of AMT was undertaken in [133] where the authors report superior results using the value of $\beta = 0.5$, having compared with values in the range of $\beta = \{0, ..., 2\}$ in steps of 0.1, thereby including the popular Euclidean and KL and IS cost functions.

The Amari α-divergence [24]

$$\mathcal{C}_\alpha(\mathbf{s}|\mathbf{z}) = \frac{1}{\alpha(\alpha-1)} \sum_i s_i^\alpha z_i^{1-\alpha} - \alpha s_i + (\alpha-1) z_i \tag{7.9}$$

is a generalised divergence which also includes the KL-divergence as a special case. Similar to the generalised β-divergence, the α-divergence is seen to encompass several other well known cost functions, such as the Pearson chi-squared distance

$$\mathcal{C}_P(\mathbf{s}|\mathbf{z}) = \sum_i \frac{(s_i - z_i)^2}{z_i} \tag{7.10}$$

when $\alpha = 2$; the Hellinger distance

$$\mathcal{C}_H(\mathbf{s}|\mathbf{z}) = \|\mathbf{s}^{[0.5]} - \mathbf{z}^{[0.5]}\|_2^2 \tag{7.11}$$

when $\alpha = 0.5$, and the Neyman chi-squared distance

$$\mathcal{C}_N(\mathbf{s}|\mathbf{z}) = \sum_i \frac{(s_i - z_i)^2}{s_i} \tag{7.12}$$

when $\alpha = -1$, while the KL-divergence is effected when $\alpha = 1$. The multiplicative update for the Amari α-divergence is given as

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes \left(\mathbf{D}^T \left[\frac{\mathbf{S}}{\mathbf{DX}} \right]^{[\alpha]} \right)^{\frac{1}{\alpha}} \tag{7.13}$$

in [27] where the division is elementwise and the atoms of $\mathbf{D}$ are required to have unit sum. Further to this the same authors recently connected both α- and β- divergences into a larger framework known as the generalised alpha-beta divergence [25], given by

$$\mathcal{C}_{\alpha\beta}(\mathbf{s}|\mathbf{z}) = -\frac{1}{\alpha\beta} \sum_i s_i^{\alpha} z_i^{\beta} - \frac{\alpha}{\alpha+\beta} s_i^{\alpha+\beta} - \frac{\beta}{\alpha+\beta} z_i^{\alpha+\beta} \tag{7.14}$$

which includes all α- and β- divergences mentioned above in a two-dimensional divergence parameter space. The definition of the α- and β-divergences as part of the $\alpha\beta$-divergence should be noted. The β divergence (7.7) is given by

$$\mathcal{C}_{\beta}^{(b)}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_{\alpha\beta}^{(1,b-1)}(\mathbf{s}|\mathbf{z}) \tag{7.15}$$

where $\mathcal{C}_{\alpha\beta}^{(a,b)}$ relates the $\alpha\beta$-divergence with $\alpha = a; \beta = b$. Similarly the the α-divergence (7.9) is given by

$$\mathcal{C}_{\alpha}^{(a)}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_{\alpha\beta}^{(a,1-a)}(\mathbf{s}|\mathbf{z}). \tag{7.16}$$

A NMD multiplicative update for the $\alpha\beta$-divergence is given [25] by

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T(\mathbf{S}^{[\alpha]} \otimes [\mathbf{DX}]^{[\beta-1]})}{\mathbf{D}^T[\mathbf{DX}]^{[\alpha+\beta-1]}} \right)^{\left[\frac{1}{\alpha}\right]} \tag{7.17}$$

when $\alpha \neq 0$.

Application-specific enhancement is often sought in NMF, both supervised and otherwise. Penalty terms afford one possibility for effecting such enhancements, and many different penalty terms have been used to augment NMF [134] [106]. Very often a sparsity-inducing penalty is desirable. While a ℓ_1 penalty has been applied [55] [107] other strategies such as logarithmic penalties [134] [1] have also been employed. Group sparsity has recently been introduced to the NMF problem. For example, it is proposed in [40] to use the IS divergence with a log-based group sparse penalty

$$\mathcal{C}_{GIS}(\mathbf{s}|\mathbf{Dx}) = \mathcal{C}_{IS}(\mathbf{s}|\mathbf{Dx}) + \lambda\Phi(\mathbf{x}) \tag{7.18}$$

for the purpose of source separation, where $\Phi(\mathbf{x})$ is a group sparse operation. In particular, the authors propose that $\Phi(\mathbf{x}) = \sum_l \log(a + g_l)$ where $g_l = \|\mathbf{x}[l]\|_1$ where $\sum \mathbf{d}_i = 1 \forall\, i$.

Monotonic decreases

The seminal NMF paper [74] proposed multiplicative updates for the Euclidean and KL cost functions, and it was shown, by employing the auxiliary function methodology, that both cost functions were non-decreasing under the actions of the proposed updates. This monotonic decrease in a cost function is a much desirable trait for any descent algorithm. Indeed part of the reason for the popularity of the multiplicative update methodology might well be the fact that cost functions which are otherwise seen to display badly scaled gradients such as the KL-divergence [135] and Itakuro-Saito divergence [1] are solvable in an uncomplicated manner, where other methods can be quite inefficient.

An interesting area of research has sought to find provably monotonic decreasing multiplicative updates. For instance a monotonic variant of the β-divergence multiplicative update

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T(\mathbf{S} \otimes [\mathbf{DX}]^{[\beta-2]})}{\mathbf{D}^T[\mathbf{DX}]^{[\beta-1]}} \right)^{[\varphi(\beta)]} \tag{7.19}$$

is proposed by Nakano et al [85], which is seen to differ from the original heuristic β-divergence algorithm (2.35) only through the exponential factor $\varphi(\beta)$:

$$\varphi(\beta) = \begin{cases} \frac{1}{2-\beta} & \text{if } \beta < 1 \\ 1 & \text{if } 1 \leq \beta \leq 2 \\ \frac{1}{\beta-1} & \text{otherwise.} \end{cases} \tag{7.20}$$

This update was derived by using an auxiliary function separately for each case shown as the β-divergence (7.7) typically consists of separate convex and concave elements. While the monotonicity is often observed for the unexponentiated updates, it is often not guaranteed [4] [41]. The results from [85], where the exponential factors (7.20) are proposed are extended by Fevotte and Idier in [41] where it is shown that $\varphi(\beta) = 1$ guarantees a monotonic descent when $0 < \beta \leq 1$.

The $\alpha\beta$ divergence is also equipped with a monotonic variant when $\alpha \neq 0$ using a similar exponential factor :

$$\varphi(\alpha\beta) = \begin{cases} \frac{1}{1-\beta} & \text{if } \frac{\beta}{\alpha} < \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} & \text{if } \frac{1}{\alpha} - 1 \leq \frac{\beta}{\alpha} \leq \frac{1}{\alpha} \\ \frac{1}{\alpha+\beta-1} & \text{if } \frac{\beta}{\alpha} > \frac{1}{\alpha}. \end{cases} \tag{7.21}$$

which correspond to $\varphi(\beta)$ in [85] when the β-divergence is generalised as in (7.15). When the α-divergence is generalised as in (7.15), $\varphi(\alpha\beta) = 1/\alpha$ guarantees monotonic descent, as proposed in the original α-divergence update (7.13).

In the rest of this chapter, the use of $\alpha\beta$-divergence for AMT is first explored, starting with experiments with some well-known cost functions. Based on these observations another generalised divergence is proposed, that itself is a special case of the $\alpha\beta$-divergence, and leads to improved AMT results with the STFT. While the $\alpha\beta$-divergence comes with guaranteed monotonicity with exponentiated multiplicative updates, it is found that the new divergence is a special case, similar to the β-divergence when $0 \leq \beta \leq 1$, where a larger stepsize can be used while maintaining montonic descent, as shown in [41]. Finally, some experiments are described, using the mixed-norm group sparse penalty with NMF.

7.1 AMT using $\alpha\beta$ divergence

The use of the β-divergences is well explored for the purpose of AMT, with $\beta = 0.5$ previously seen to give optimal results [133] [32]. In particular, the authors of [133] performed extensive

	$\mathbb{S}1$	$\mathbb{S}2$	$\mathbb{E}1$	$\mathbb{E}2$	$\mathbb{E}3$	$\mathbb{E}4$
$\mathcal{C}_N$	61.6	61.5	64.6	67.3	66.5	67.7
$\mathcal{C}_{EUCL}$	64.0	64.1	66.7	68.4	68.1	68.6
$\mathcal{C}_P$	68.5	68.7	70.1	71.2	70.9	72.8
$\mathcal{C}_H$	70.1	70.1	71.3	73.0	72.6	73.7
$\mathcal{C}_{KL}$	70.9	71.0	70.9	73.5	73.1	73.9
$\mathcal{C}_\beta^{(0.5)}$	**71.1**	**71.5**	**72.0**	**74.9**	**74.7**	**75.3**

Table 7.1: Comparision of NMD using several different $\alpha\beta$ divergence cost functions for AMT experiments on standard dataset across various transforms using single atom dictionaries. Results given in terms of optimal $\mathcal{F}$-measure.

experiments similar to those presented here, for a range of values of β. However the use of the α-divergence is relatively unexplored, with the exception of the KL-divergence which is both an α- and a β- divergence. The KL-divergence is a commonly used cost function in musical signal processing [134] [115], for which good results are generally reported, and it may be worthwhile to compare the use of some other α-divergences.

Experiments

In light of this, some experiments were run to compare the relative performance of some well known cost functions that are generalised by the $\alpha\beta$-divergence. AMT decompositions were performed on the standard dataset (§3.1), using single atom dictionaries with all transforms outlined in Table 3.1. The cost functions used were the Euclidean, $\mathcal{C}_E$, Hellinger, $\mathcal{C}_H$, (7.11), Neyman, $\mathcal{C}_N$, (7.12) and Pearson, $\mathcal{C}_P$ (7.10), distances, the KL-divergence, $\mathcal{C}_{KL}$ (7.5) and generalised β-divergence with $\beta = 0.5$ denoted by $\mathcal{C}_\beta^{(0.5)}$.

The results are shown in Table 7.1. It is seen here that, from the selected cost functions, $\mathcal{C}_\beta^{(0.5)}$ performs best, followed by the Kullback-Leibler divergence, Hellinger distance and the Pearson distance, while the Neyman distance is seen to perform worst for all transforms, followed by the Euclidean distance. Across the different transforms, a variation in performance is seen, with a similar pattern to when greedy and stepwise methods were used in Chapters 4 and 5, with the larger dimension ERBTs performing best, and STFTs seen to perform the worst. However, the difference in performance relative to the transform used is small in comparision to the results seen for the OMP based methods, being approximately $3-5\%$, similar to results seen with the (G)BF-NNLS methods. It is notable that the performance of the β- and KL-divergences are quite similar in the case of the STFTs while those of the Hellinger distance and KL-divergence are closely matched in the case of the ERBs.

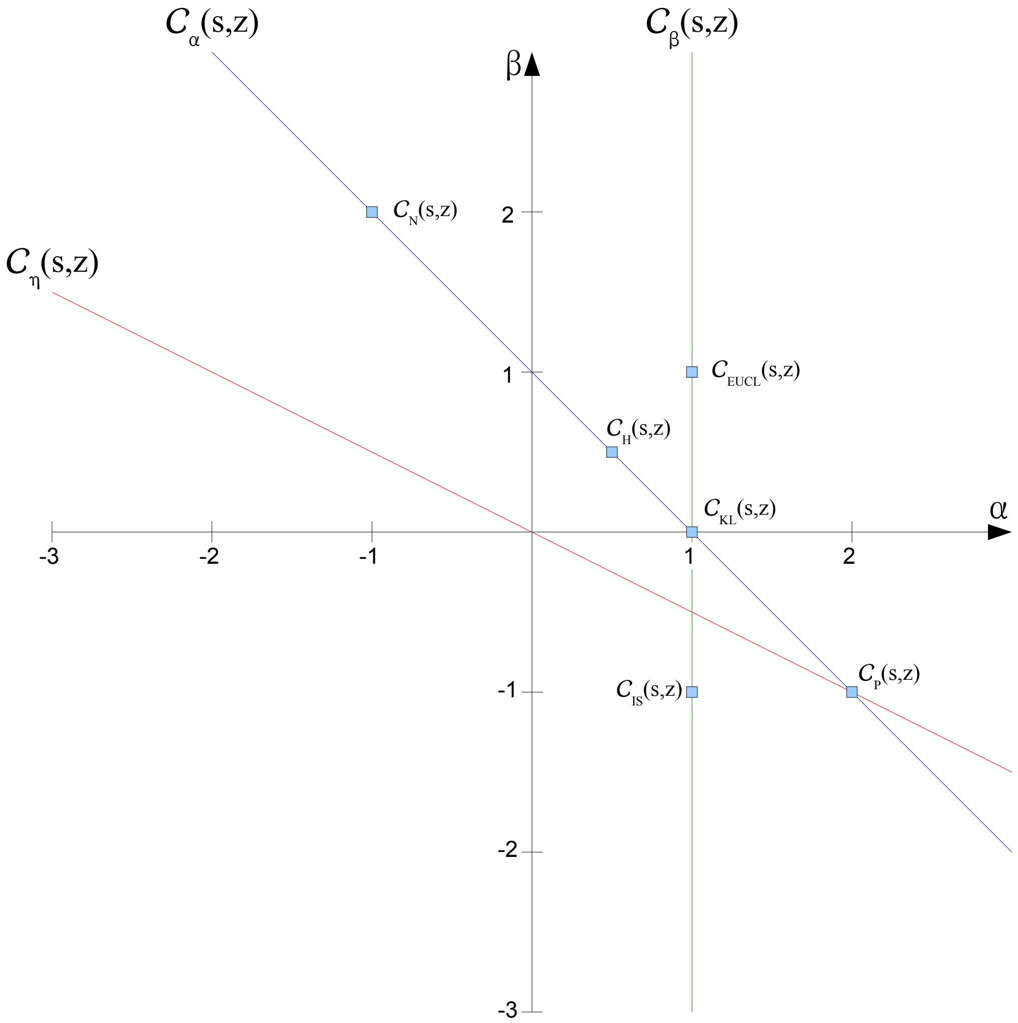

Figure 7.1: Diagram showing the η-, α- and β- divergences as generalised $\alpha\beta$-divergences. Popular cost functions generalised by $\alpha\beta$-divergence also indicated

7.1.1 Another generalised divergence

In light of the results seen in the previous experiments, it is worth noting the similar form of the generalised β-divergence with $\beta = 0.5$:

$$\mathcal{C}_\beta^{(0.5)}(\mathbf{s}|\mathbf{z}) = \frac{1}{0.5}\sum_i \frac{(s_i^{0.5} - z_i^{0.5})^2}{z_i^{0.5}} \tag{7.22}$$

and the Pearson distance (7.10). Of particular interest is the fact that both are seen to be weighted by a denominator, based upon the current estimate $\mathbf{z} = \mathbf{Dx}$. Otherwise stated, $\mathcal{C}_P$ and $\mathcal{C}_\beta^{(0.5)}$ are *model weighted* versions of the Euclidean distance and Hellinger distance (7.11), respectively.

Furthermore, these model weighted cost functions outperform the corresponding unweighted cost functions. Conversely, the Neyman distance (7.12) is seen to be weighted by the signal, and is seen to perform the worst of all cost functions used. This resonates with results given in [24] where the Neyman distance is seen to perform well when the signal matrix is dense, and it is assumed that presence of small values in the signal matrix leads to instabilities in the signal representations leading to possibly large errors [24].

As the model weighting approach is seen to be useful, it may be worthy of further exploration. Hence , it is proposed to generalise the two model weighted cost functions, $\mathcal{C}_P$ and $\mathcal{C}_\beta^{(0.5)}$, and a parametric cost function which is referred to as the η-divergence is proposed :

$$\mathcal{C}_\eta(\mathbf{s}|\mathbf{z}) = \frac{1}{\eta}\sum_i \frac{(s_i^\eta - z_i^\eta)^2}{z_i^\eta} = \frac{1}{\eta}\sum_i s_i^{2\eta} z_i^{-\eta} + z_i^\eta - 2s_i^\eta. \tag{7.23}$$

In order to derive a multiplicative update for (7.23), the gradient is first taken :

$$\frac{d\mathcal{C}_\eta(\mathbf{s}|\mathbf{Dx})}{d\mathbf{x}} = \mathbf{D}^T[\mathbf{Dx}^{[\eta-1]}] - \mathbf{D}^T\left[\mathbf{s}^{[2\eta]} \otimes [\mathbf{Dx}]^{[-\eta-1]}\right]. \tag{7.24}$$

In typical NMF fashion [74] [24] a fixed stepsize

$$v = -\frac{\mathbf{x}}{\mathbf{D}^T[\mathbf{Dx}]_i^{[\eta-1]}} \tag{7.25}$$

can be defined which leads to additive gradient descent reducing to the multiplicative update:

$$\begin{aligned} \mathbf{X} &\longleftarrow \mathbf{X} + v\frac{d\mathcal{C}_\eta(\mathbf{s}|\mathbf{Dx})}{d\mathbf{x}} \\ &= \mathbf{X} \otimes \frac{\mathbf{D}^T\left[\mathbf{S}^{[2\eta]} \otimes [\mathbf{DX}]^{[-\eta-1]}\right]}{\mathbf{D}^T[\mathbf{DX}]^{[\eta-1]}}. \end{aligned} \tag{7.26}$$

This is equivalent to the heuristic β-NMF update (2.35) when $\beta = \eta = 0.5$. Indeed the update (7.26) bears resemblance to the update for the $\alpha\beta$-divergence . As might be expected, due to its derivation from linking two separate $\alpha\beta$-divergences, close inspection reveals that the η-divergence is a special case of the $\alpha\beta$-divergence, where $\alpha = -2\beta = 2\eta$. This can also be expressed by

$$\mathcal{C}_\eta^{(y)}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_{\alpha\beta}^{(2y,-y)}. \tag{7.27}$$

where y is any value of η, and is graphically shown in Figure 7.1. An advantage of the relation-

	Transform					
	$\mathbb{S}1$	$\mathbb{S}2$	$\mathbb{E}1$	$\mathbb{E}2$	$\mathbb{E}3$	$\mathbb{E}4$
$\eta = 0.1$	49.8	39.6	59.5	65.7	64.5	66.9
0.2	57.9	53.7	66.2	70.7	70.3	71.5
0.3	63.7	63.3	69.6	73.3	73.0	73.8
0.4	68.1	68.5	71.3	74.5	74.3	75.0
0.5	71.0	71.4	**72.1**	**74.9**	**74.7**	**75.3**
0.6	72.4	72.6	**72.1**	74.8	74.5	75.2
0.7	**72.6**	**72.8**	71.8	74.4	74.1	74.9
0.8	72.1	72.2	71.4	74.0	73.6	74.4
0.9	71.3	71.4	71.0	73.5	73.1	74.0
1.0	69.0	69.0	69.6	72.1	71.7	72.7

Table 7.2: AMT performance for η-divergence expressed in F-measure for various transforms over a range of values of η.

ship with the $\alpha\beta$-divergence is the admission of a monotonic descent algorithm

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T \left[\mathbf{S}^{[2\eta]} \otimes [\mathbf{DX}]^{[-\eta-1]} \right]}{\mathbf{D}^T [\mathbf{DX}]^{[\eta-1]}} \right)^{\varphi(\eta)} \tag{7.28}$$

where

$$\varphi(\eta) = \begin{cases} \frac{1}{\eta+1} & \text{if } 0 < \eta \leq 1 \\ \frac{1}{\eta-1} & \text{if } -1 \leq \eta < 0 \\ \frac{1}{2\eta} & \text{otherwise.} \end{cases} \tag{7.29}$$

is derived from the definition of $\varphi(\alpha\beta)$ (7.21)

Experiments

Experiments were run using the η-divergence with varying values of η in order to test if further AMT improvements are possible. The experiments are similar to the previous AMT experiments for $\alpha\beta$-divergence, with decompositions of all pieces in the standard dataset (§3.1) performed in all transforms (§3.2) using the atomic pitch dictionaries (§3.3.2). Decompositions were performed for values of $\eta \in \{0.1, ...1.0\}$ in steps of 0.1.

Results are shown in Table 7.2 where two different patterns are seen. For the STFT transforms, the best results are seen with $\eta = 0.7$, noting that previously the best results for these transforms were seen with $\eta = \beta = 0.5$. The improvement seen is mild at around 1.5%; however improvements over $\eta = 0.5$ are seen over the range $\eta \in [0.6, 0.9]$ in both cases. However, these results with the STFT are better than the results seen for the STFT using all other cost functions. For the ERB transforms, the best results are seen for $\eta = 0.5 - 0.6$, and little separates the results for these values.

The authors of [25] describe how the alpha parameter can be seen as a ”zoom” parameter on a β-divergence. When $\alpha \neq 0$ this relationship is expressed by

$$\mathcal{C}_{\alpha,\beta}^{(a,b)}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_{\beta}^{(b+1)}(\mathbf{s}^{[a]}|\mathbf{z}^{[a]}) \tag{7.30}$$

where the change in the variable b in $\alpha\beta$-divergence to $b+1$ in β-divergence is an expression of (7.15). In this manner, the η-divergence can be seen as the β-divergence with fixed $\beta = 0.5$, and an α-zoom on the signal and model. For example, when $\eta = 0.7$, as was seen to produce superior results for the STFTs, the η-divergence can also be expressed as :

$$\mathcal{C}_{\eta}^{(0.7)}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_{\beta}^{(0.5)}(\mathbf{s}^{[1.4]}|\mathbf{z}^{[1.4]}). \tag{7.31}$$

7.1.2 Improved monotonic update for η-divergence

In the last section results were given showing that the η-divergence may provide useful cost functions, particularly in the range $\eta \in [0.5, 1.0]$, with improved results observed in the case of the STFTs. Indeed, these can be considered state-of-the art decompositions for AMT using the STFT transform.

The multiplicative update used (7.28) included an exponential factor $\varphi(\eta)$ (7.29), the use of which guarantees monotonic decreases in the cost function. This factor $\varphi(\eta)$ is derived from $\varphi(\alpha\beta)$ given as part of the $\alpha\beta$-NMF framework (7.21). $\varphi(\alpha\beta)$ also generalises $\varphi(\beta)$ [85] which gives monotonic updates for the β-divergence, and the monotonic descent algorithm given for the α-divergence (7.13) [24].

While the exponentiated updates given for the $\alpha\beta$-divergence have proven monotonicity, special cases also exist where a larger step than that proposed in (7.21) may be taken. For example, Fevotte and Idier [41] show that for β-divergence, the unexponentiated multiplicative update (7.8) originally proposed in [24] is monotonic for values of $\beta \in [0, 1]$, where previously this was only shown to be the case for $\beta \in [1, 2]$ (7.20) as originally given by Nakano et al [85] and also generalised by $\varphi(\alpha\beta)$ (7.21). Similarly, it may be shown that the η-divergence in the range $\eta \in [0.5, 1]$ is a similar special case that can accommodate a larger step size while maintaining monotonicity.

	$\mathbf{s}^{[2\eta]}[\mathbf{Dx}]^{[-\eta]}$	$[\mathbf{Dx}]^{[\eta]}$	$-2\mathbf{s}^{[\eta]}$
$0 < \eta \leq 1$	$\smile$	$\frown$	-
$-1 \leq \eta < 0$	$\frown$	$\smile$	-
$\lvert\eta\rvert > 1$	$\smile$	$\smile$	-

Table 7.3: Convexity/ concavity of separate terms of η-divergence in terms of $\mathbf{Dx}$ relative to value of η. Convexity denoted by $\smile$; concavity denoted by $\frown$ and constant denoted by $-$

Theorem 1. *For $\eta \in [0.5, 1.0]$ the cost function* (7.23) *is non-increasing under the multiplicative update*

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T \left[\mathbf{S}^{[2\eta]} \otimes [\mathbf{DX}]^{[\eta+1]} \right]}{\mathbf{D}^T [\mathbf{DX}]^{[\eta-1]}} \right)^{\frac{1}{2\eta}} \tag{7.32}$$

where $\mathbf{X}, \mathbf{D}, \mathbf{S} \geq 0$.

The proof of this theorem is delayed while the requisite tools are described. A similar methodology to that used in [41] is taken. This requires the use of an auxiliary function, as used in [74][85][41][25] to prove monotonicity for NMF algorithms. Given a cost function $\mathcal{C}(\mathbf{s}|\mathbf{z}) = \mathcal{C}[\mathbf{x}]$ where $\mathbf{z} = \mathbf{Dx}$ an auxiliary function is given by $F[\mathbf{x}, \hat{\mathbf{x}}]$ where $\hat{\mathbf{x}}$ is referred to as an auxiliary variable. An auxiliary function has by definition [41], the following properties

$$\begin{aligned} \mathcal{C}[\mathbf{x}] &= F[\mathbf{x}|\mathbf{x}]\, \forall \mathbf{x} \\ \mathcal{C}[\mathbf{x}] &\leq F[\mathbf{x}|\hat{\mathbf{x}}]\, \forall (\mathbf{x}, \hat{\mathbf{x}}) \end{aligned} \tag{7.33}$$

which show that the auxiliary function provides an upper bound to the cost function at $\mathbf{x}$. In most cases of the $\alpha\beta$-divergence the cost function is seen to consist of separate convex, concave and linear elements [25], and the η-divergence is similar. Multiplying out the $\mathcal{C}_\eta$ cost function (7.23) and displaying the curvature of each term, when $0 \leq \eta \leq 1$ gives:

$$\begin{aligned} \mathcal{C}_\eta(\mathbf{s}|\mathbf{z}) &= \mathbf{s}^{[2\eta]}[\mathbf{z}]^{[-\eta]} + [\mathbf{z}]^{[\eta]} - 2\mathbf{s}^{[\eta]} \\ &= \overset{\smile}{\mathcal{C}}_\eta(\mathbf{s}|\mathbf{z}) + \overset{\frown}{\mathcal{C}}_\eta(\mathbf{s}|\mathbf{z}) + \bar{\mathcal{C}}_\eta(\mathbf{s}|\mathbf{z}) \end{aligned} \tag{7.34}$$

where $\overset{\smile}{\mathcal{C}}$ denotes a convex term, $\overset{\frown}{\mathcal{C}}$ denotes a concave term and $\bar{\mathcal{C}}$ denotes a constant term. These curvature definitions do not hold for values of η outside the range $(0, 1)$, and the possible combinations of convex and concave functions are outlined in Table 7.3, leading to three possible combinations relative to the value of η. However, the focus here is solely on the first row of Table 7.3 as given in (7.34), and particularly a subset of that range, when $0.5 \leq \eta \leq 1$.

In this case of mixed curvature, an auxilary function is similarly split into concave, convex and linear parts, and the following theorem gives a general auxiliary function for a mixed curvature cost function.

Theorem 2 (Fevotte and Idier [41]). *Given $\hat{\mathbf{z}} = \mathbf{D}\hat{\mathbf{x}}$, the function*

$$F[\mathbf{x},\hat{\mathbf{x}}] = \sum_m \left[\left(\sum_n \frac{d_{mn}\hat{x}_n}{\hat{z}} \breve{\mathcal{C}}(s_m|\hat{z}_m \frac{x_n}{\hat{x}_n}) \right) + \left(\overset{\frown}{\mathcal{C}}(s_m|\hat{z}_m) + \sum_n d_{mn}(x_n - \hat{x}_n)\overset{\frown}{\mathcal{C}}'(s_m|\hat{z}_m) \right) + \bar{\mathcal{C}}(s_m|\hat{z}_m) \right] \quad (7.35)$$

is an auxiliary function to $\mathcal{C}(\mathbf{s}|\mathbf{z}) = \breve{\mathcal{C}}(\mathbf{s}|\mathbf{z}) + \overset{\frown}{\mathcal{C}}(\mathbf{s}|\mathbf{z}) + \bar{\mathcal{C}}(\mathbf{s}|\mathbf{z})$ where $\mathbf{z} = \mathbf{D}\mathbf{x}$.

Proof. Defining $\mathcal{C}_m[\mathbf{x}] = \mathcal{C}(s_m|z_m)$ it is noted that $\mathcal{C}[\mathbf{x}] = \sum_m \mathcal{C}_m[\mathbf{x}]$. The auxiliary function can be constructed:

$$\begin{aligned} F[\mathbf{x},\hat{\mathbf{x}}] &= \sum_m F_m[\mathbf{x},\hat{\mathbf{x}}] \\ &= \sum_m \breve{F}_m[\mathbf{x},\hat{\mathbf{x}}] + \overset{\frown}{F}_m[\mathbf{x},\hat{\mathbf{x}}] + \bar{F}_m[\mathbf{x},\hat{\mathbf{x}}] \end{aligned} \quad (7.36)$$

such that $\breve{F}_m[\mathbf{x},\hat{\mathbf{x}}] \geq \breve{\mathcal{C}}_m[\mathbf{x}]$ and $\overset{\frown}{F}_m[\mathbf{x},\hat{\mathbf{x}}] \geq \overset{\frown}{\mathcal{C}}_m[\mathbf{x}]$ while $\bar{F}_m[\mathbf{x},\hat{\mathbf{x}}] = \bar{\mathcal{C}}_m[\mathbf{x}]$, thereby considering separate auxiliary functions for the convex and concave parts at each element of the signal.

The convex auxiliary function, $\breve{F}_m[\mathbf{x},\hat{\mathbf{x}}]$, is given using Jensen's inequality and equates to the first term of (7.35) while the first term of the Taylor expansion is used to give the concave auxiliary function, $\overset{\frown}{F}_m[\mathbf{x},\hat{\mathbf{x}}]$, which is equal to the second term of (7.35). A full proof is given in [41]. □

It is noted that Theorem 2 generalises the approach taken by Nakano et al [85] who provide a similar proof for the specific case of the β-divergence. An important property of the auxiliary function (7.35) is that it is separable in each variable. In particular the following relationships are noted in [41]:

$$F[\mathbf{x},\hat{\mathbf{x}}] = \sum_n F_n[x_n,\hat{\mathbf{x}}] + K \quad (7.37)$$

where K is a constant with respect to $\hat{\mathbf{x}}$, and

$$F_n[x_n,\hat{\mathbf{x}}] = \hat{x}_n \left[\sum_m \frac{d_{mn}}{\hat{z}_m} \breve{\mathcal{C}}\left(s_m|\hat{z}_m \frac{x_n}{\hat{x}_n} \right) \right] + x_n \left[\sum_m d_{mn} \overset{\frown}{\mathcal{C}}'(s_m|\hat{z}_m) \right] \quad (7.38)$$

which in the particular case of the η-divergence in the specified range reduces to

$$F_n[x_n,\hat{\mathbf{x}}] = \hat{x}_n \left(\frac{\hat{x}_n}{x_n} \right)^{\eta} \left[\mathbf{d}_k^T (\mathbf{s}^{[2\eta]} \otimes [\mathbf{D}\hat{\mathbf{x}}]^{[-\eta-1]}) \right] + \eta x_n \left[\mathbf{d}_k^T ([\mathbf{D}\hat{\mathbf{x}}]^{[\eta-1]}) \right] \quad (7.39)$$

using the fact that

$$\frac{d\overset{\frown}{\mathcal{C}}_\eta(\mathbf{s}|\mathbf{z})}{d\mathbf{z}} = \eta z^{\eta-1}. \tag{7.40}$$

Another necessary element of the proof of Theorem 1 is the introduction of a scalar auxiliary function, as defined in [41]:

$$f(a|b,c) = \breve{\mathcal{C}}(c|a) + \overset{\frown}{\mathcal{C}}(c|b) + (a-b)\breve{\mathcal{C}}'(c|b) + \bar{\mathcal{C}}(c) \tag{7.41}$$

where $f(a|b,b) = \mathcal{C}(a|b)$. The η-divergence update can be rewritten in the univariate case with $\varphi(\alpha\beta) = 1/\alpha$

$$\bar{x}_n \longleftarrow \hat{x}_n \left(\frac{\mathbf{d}_n^T(\mathbf{s}^{[2\eta]} \otimes \hat{\mathbf{z}}^{[-\eta-1]})}{\mathbf{d}_n^T \hat{\mathbf{z}}^{[\eta-1]}} \right)^{\frac{1}{\alpha}} = \hat{x}_n J^{[\frac{1}{\alpha}]}$$

Further to this, the following Lemma is proposed, which is similar to that used for the case of the β-divergence in [41].

Lemma 1. *For η in the range $(0,1]$*

$$F_n[x_n,\hat{\mathbf{x}}] = \hat{x}_n^{(1-\eta)} \left(\mathbf{d}_n^T(\hat{\mathbf{z}}^{[\eta-1]}) \right) f(x_n|\hat{x}_n,\bar{x}_n) + K \tag{7.42}$$

where K is constant in terms of $\hat{x}_n and \bar{x}_n$.

Proof. First, writing out the scalar auxiliary function (7.41) for the case of the η-divergence in the range $(0,1)$ gives

$$f(x_n|\hat{x}_n,\bar{x}_n) = \bar{x}_n^{2\eta} x_n^{-\eta} + \eta x_n \hat{x}_n^{\eta-1} + (1-\eta)\hat{x}_n^{\eta} - 2\bar{x}_n. \tag{7.43}$$

The different terms of $f(x|\hat{x},\bar{x})$ (7.43) can be multiplied out separately. First, consider the first term:

$$\begin{aligned}
\hat{x}_n^{(1-\eta)} \left(\mathbf{d}_n^T(\hat{\mathbf{z}}^{[\eta-1]}) \right) \bar{x}^{2\eta} x^{-\eta} &= \hat{x}_n^{(1-\eta)} \left(\mathbf{d}_n^T(\hat{\mathbf{z}}^{[\eta-1]}) \right) (\bar{x}_n J^{[\frac{1}{\alpha}]})^{2\eta} x^{-\eta} \\
&= \hat{x}_n \left(\frac{\hat{x}_n}{x_n} \right)^{\eta} \left(\mathbf{d}_n^T(\hat{\mathbf{z}}^{[\eta-1]}) \right) J \\
&= \hat{x}_n \left(\frac{\hat{x}_n}{x_n} \right)^{\eta} \left(\mathbf{d}_n^T(\mathbf{s}^{[2\eta]} \otimes \hat{\mathbf{z}}^{[-\eta-1]}) \right) \qquad (7.44)
\end{aligned}$$

which is seen to be the same as the first term of $F_n[x_n, \hat{\mathbf{x}}]$ in (7.39). Similarly, the second term

$$\hat{x}_n^{(1-\eta)}\left(\mathbf{d}_n^T[\hat{\mathbf{z}}]^{[\eta-1]}\right)\eta x_n \hat{x}_n^{[\eta-1]} = \eta x_n \mathbf{d}_k^T[\hat{\mathbf{z}}]^{[\eta-1]} \tag{7.45}$$

is seen to be equal to the second term of $F_n[x_n, \hat{\mathbf{x}}]$ in (7.39). The third and fourth terms of (7.43) are seen to be constants in terms of $\bar{x}$ or $\hat{x}$, thereby proving the Lemma. □

Now that the necessary framework is in place, proof of Theorem 1 is given.

Proof of Theorem 1. Ultimately, the proof requires demonstration that

$$\mathcal{C}_\eta(\mathbf{s}|\mathbf{D}\bar{\mathbf{x}}) < \mathcal{C}_\eta(\mathbf{s}|\mathbf{D}\hat{\mathbf{x}}) \tag{7.46}$$

for $\eta \in (0.5, 1)$ where each $\hat{x}_n$ is given by (7.42). Due to the definition of an auxiliary function (7.33) the condition (7.46) is guaranteed when $F[\bar{\mathbf{x}}, \hat{\mathbf{x}}] \le F[\hat{\mathbf{x}}, \hat{\mathbf{x}}]$. Furthermore, separability of the auxiliary function guarantees (7.46) when

$$F_n[\bar{x}_n, \hat{\mathbf{x}}] \le F_n[\hat{x}_n, \hat{\mathbf{x}}]$$

which reduces to

$$f(\bar{x}|\hat{x}, \bar{x}) \le f(\hat{x}|\hat{x}, \bar{x}) \tag{7.47}$$

following Lemma 1.

Using dummy variables, rearranging and making cancellations leads to (7.47) being rewritten

$$f(a|a,b) - f(b|a,b) = \breve{\mathcal{C}}(a|b) - \breve{\mathcal{C}}(a|a) - (a-b)\overset{\frown}{\mathcal{C}}'(a|b) \ge 0 \tag{7.48}$$

t which can be stated in terms of η-divergence

$$\begin{aligned} f(a|a,b) - f(b|a,b) &= a^{[2\eta]}b^{-\eta} - a^{[2\eta]}a^{-\eta} - (a-b)\eta b^{\eta-1} \\ &= a^{[2\eta]}b^{-\eta} - a^{\eta} - a\eta b^{\eta-1} + \eta b^{\eta} \\ &= b^{\eta}(\Theta^{2\eta} - \Theta^{\eta} - \eta\Theta + \eta) \end{aligned} \tag{7.49}$$

where $\Theta = a/b$. It follows from non-negativity that $b^{\eta} \ge 0$. Given that $\Theta^{\eta} \le 1 + (\Theta - 1)\eta$ from concavity of Θ^{η} as $\eta \le 1$ the proof only requires that

$$\begin{aligned}\Theta^{2\eta} &\geq \eta\Theta-\eta+1+(\Theta-1)\eta\\ &= 1+2\eta(\Theta-1)\end{aligned} \tag{7.50}$$

and recalling that $\alpha = 2\eta$ gives

$$\Theta^{\alpha} \geq 1+\alpha(\Theta-1) \tag{7.51}$$

which is true in all cases, again due to the Taylor expansion with convexity of Θ^{α} which is convex as $\alpha \geq 1$. □

7.2 Applying sparse penalties to Alpha-beta divergence

In order to make a signal representation sparse, application of an ℓ_1-norm penalty is a common procedure in the sparse representations methodology [23] [124]. The ℓ_1 penalty is also often considered in NMF [56] [107]. Application of an ℓ_1 penalty to the β-divergence is proposed in [41] leading to the cost function

$$\mathcal{C}_{S\beta}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_{\beta}(\mathbf{s}|\mathbf{z}) + \lambda\|\mathbf{x}\|_1 \tag{7.52}$$

where $\mathbf{z} = \mathbf{D}\mathbf{x}$. Monotonic descent multiplicative updates for (7.52) are given in [41] by

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T(\mathbf{S}\otimes[\mathbf{D}\mathbf{X}]^{[\beta-2]})}{\mathbf{D}^T[\mathbf{D}\mathbf{X}]^{[\beta-1]}+\lambda}\right)^{[\frac{1}{2-\beta}]} \tag{7.53}$$

when $\beta \in [0,1]$, and similarly when $\beta \geq 2$:

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T(\mathbf{S}\otimes[\mathbf{D}\mathbf{X}]^{[\beta-2]})-\lambda}{\mathbf{D}^T[\mathbf{D}\mathbf{X}]^{[\beta-1]}}\right)^{[\frac{1}{\beta-1}]}. \tag{7.54}$$

It is now shown that the α-divergence also accommodates an ℓ_1 penalty term in a straightforward manner

Lemma 2. *The cost function*

$$\mathcal{C}_{S\alpha} = \frac{1}{\alpha(\alpha-1)}\left(\sum_i s_m^{\alpha} z_m^{1-\alpha} - \alpha s_m + (\alpha-1)z_m\right) + \frac{\lambda\|\mathbf{x}\|_1}{\alpha} \tag{7.55}$$

where $\mathbf{z} = \mathbf{D}\mathbf{x}$ *is non-increasing with the multiplicative update*

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T [\frac{\mathbf{S}}{\mathbf{DX}}]^{[\alpha]}}{\mathbf{D}^T \mathbf{1}_M + \lambda} \right)^{[\frac{1}{\alpha}]}. \tag{7.56}$$

Proof. First an auxiliary function for (7.55) is formed using (7.35). This is performed by bounding only the first element of the summed term in (7.55) which is convex as all other terms are linear, or constant:

$$F[\mathbf{x}, \hat{\mathbf{x}}] = \sum_m \left[\sum_n \frac{d_{mn}\hat{x}_n}{\hat{z}_m} \breve{\mathcal{C}}(s_m | \hat{z}_m \frac{x_n}{\hat{x}_n}) - \frac{1}{(\alpha-1)} s_m + \frac{1}{\alpha}\hat{z}_m \right] + \frac{1}{\alpha}\lambda \|\mathbf{x}\|_1 \tag{7.57}$$

Separability of (7.57) is given by (7.38) [41] and the fact that the ℓ_1 norm is separable:

$$F_n[x_n | \hat{\mathbf{x}}] = \hat{x}_n \left[\sum_m \frac{d_{mn}}{\hat{z}_m} \breve{\mathcal{C}}(s_m | \hat{z}_m \frac{x_n}{\hat{x}_n}) \right] + \frac{1}{\alpha}\mathbf{d}_n x_n + \frac{1}{\alpha}\lambda x_n \tag{7.58}$$

the gradient of which, with respect to $\mathbf{x}$ is given by

$$-\frac{1}{\alpha}\mathbf{d}_n^T (\mathbf{s}^{[a]} \otimes \hat{\mathbf{z}}^{[-a]}) \left(\frac{x_n}{\hat{x}_n} \right)^{-\alpha} + \frac{1}{\alpha}\mathbf{d}_n^T \mathbf{1}_M + \frac{1}{\alpha}\lambda. \tag{7.59}$$

Setting the gradient (7.59) to zero gives

$$\left(\frac{x_n}{\hat{x}_n} \right)^{-\alpha} = \frac{\mathbf{d}_n^T \mathbf{1}_M + \lambda}{\mathbf{d}_n^T (\mathbf{s}^{[a]} \otimes \hat{\mathbf{z}}^{[-a]})} \tag{7.60}$$

and rearranging:

$$x_n = \hat{x}_n \otimes \left(\frac{\mathbf{d}_n^T (\mathbf{s}^{[a]} \otimes \mathbf{D}\hat{\mathbf{x}}^{[-a]})}{\mathbf{d}_n^T \mathbf{1}_M + \lambda} \right)^{[\frac{1}{\alpha}]} \tag{7.61}$$

from which (7.56) follows. □

While the ℓ_1-penalty is commonly used, other approaches are possible. It is proposed to use a ℓ_p^p penalty term, described in [43] [19]. It is worth noting that the ℓ_p-norm exponentiated to p is also the ℓ_1 norm of the likewise exponentiated vector :

$$\|\mathbf{x}\|_p^p = \|\mathbf{x}^{[p]}\|_1. \tag{7.62}$$

In particular this affords separability of the penalty term in each variable, a property that is seen in Lemma 2 to allow incorporation of penalty terms to monotonic NMF updates in a straightforward manner. It is now shown that a monotonic update is available for the ℓ_p^p-penalised $\alpha\beta$-divergence in the range $\alpha+\beta < 1, \beta < 0$.

Lemma 3. *The cost function*

$$\mathcal{C}_{S\alpha\beta} = -\frac{1}{\alpha\beta}\left(\sum_i s_i^{\alpha} z_i^{\beta} - \frac{\alpha}{\alpha+\beta} s_i^{\alpha+\beta} - \frac{\beta}{\alpha+\beta} z_i^{\alpha+\beta}\right) + \frac{\lambda}{\alpha p}\|\mathbf{x}\|_p^p \tag{7.63}$$

where $\mathbf{z} = \mathbf{D}\mathbf{x}$ *is non-increasing with the multiplicative update*

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes \left(\frac{\mathbf{D}^T[\mathbf{S}^{[\alpha]} \otimes [\mathbf{D}\mathbf{X}]^{[\beta-1]}}{\mathbf{D}^T[\mathbf{D}\mathbf{X}]^{[\alpha+\beta-1]} + \lambda \mathbf{X}^{[p-1]}}\right)^{\left[\frac{1}{1-\beta}\right]}. \tag{7.64}$$

when $0 < \alpha+\beta < 1$*;* $0 < p < 1$ *and* $\beta < 0$*.*

Proof. A similar proof to that given for Lemma 2 is used, relying on the auxiliary function defined in Theorem 2. The ℓ_p^p norm is separable, as can be seen from its statement as a $\ell_1$norm (7.62) and concave in each element as $p < 1$, leading to an auxiliary function for the penalty term

$$\begin{aligned} F_n^{\|\cdot\|}[x_n|\hat{\mathbf{x}}] &= \frac{\lambda}{\alpha p}\hat{x_n}^p + (x_n - \hat{x}_n)\frac{d\hat{x}_n^{[p]}}{d\hat{x}_n} \\ &= \frac{\lambda}{\alpha p}\left[(1-p)\hat{x}_n^p + p\hat{x}^{(p-1)}x_n\right]. \end{aligned} \tag{7.65}$$

The general auxiliary function (7.35) is then augmented by (7.65). The gradient for the combined auxiliary function with relation to x_n for the $\alpha\beta$-divergence in the given range is then given as

$$\frac{1}{\alpha}\left[-\mathbf{d}_n^T(\mathbf{s}^{[\alpha]} \otimes \hat{\mathbf{z}}^{[\beta-1]})\left(\frac{x_n}{\hat{x}_n}\right)^{[\beta-1]} + \mathbf{d}_n^T\hat{\mathbf{z}}^{[\alpha+\beta-1]} + \lambda\hat{x}_n^{[p-1]}\right] \tag{7.66}$$

Setting the gradient to zero gives

$$-\mathbf{d}_n^T(\mathbf{s}^{[\alpha]} \otimes \hat{\mathbf{z}}^{[\beta-1]})\left(\frac{x_n}{\hat{x}_n}\right)^{[\beta-1]} = \mathbf{d}_n^T\hat{\mathbf{z}}^{[\alpha+\beta-1]} + \lambda\hat{x}_n^{[p-1]} \tag{7.67}$$

and rearranging leads to

$$x_n = \hat{x}_n\left(\frac{\mathbf{d}_n^T(\mathbf{s}^{[\alpha]} \otimes \hat{\mathbf{z}}^{[\beta-1]})}{\mathbf{d}_n^T\hat{\mathbf{z}}^{[\alpha+\beta-1]} + \lambda\hat{x}_n^{[p-1]}}\right)^{\left[\frac{1}{1-\beta}\right]} \tag{7.68}$$

	Penalty term		
	0	ℓ_1	$\ell_{0.5}^{0.5}$
$\mathcal{C}_{KL}$	73.9	75.0	**75.6**
$\mathcal{C}_{\beta}^{(0.5)}$	75.2	75.5	**76.4**

Table 7.4: AMT results in $\mathcal{F}$-measure for two β-divergences, without a sparse penalty, and with different sparse penalty strategies applied.

from which (7.64) follows. □

Experiments

Some decomposition experiments were run to compare the two penalisation approaches. The cost functions used were limited for these experiments to $\mathcal{C}_{KL}$ and $\mathcal{C}_{\beta}^{(0.5)}$. Decompositions were performed on the standard dataset with spectrograms from Transform $\mathbb{E}4$, the best performing transform. Only the atomic pitch dictionaries were used. After some initial experiments a value of $\lambda = 0.5$ was selected for $\mathcal{C}_{\beta}^{(0.5)}$ and in the case of $\mathcal{C}_{KL}$; $\lambda = 2$. These values were seen to be good for the individual transforms and to maintain that state regardless of the penalty norm employed. For both cost functions, a $\ell_{0.5}^{0.5}$ norm is used.

The results of the experiments are shown in Table 7.5. Here it is seen that the $\ell_{0.5}^{0.5}$ penalty results in better performance than the ℓ_1 penalty. The improvements are mild, less than 2% better than the results for the unpenalised approach.

7.2.1 Group sparse penalisation

A particular focus of this thesis is the incorporation of subspace modelling using group sparse methods. Group sparse NMF has previously been proposed [75] using the Itakuro-Saito divergence with a log-penalised group penalty with an ℓ_1 norm for the group coefficient.

While proofs of monotonicity are not offered, it is proposed to perform group sparse penalisation using mixed norm approaches, in particular penalties of the form $\ell_{p,q}^{q}$. Two different approaches are undertaken, an $\ell_{2,0.5}^{0.5}$ and $\ell_{\perp,0.5}^{0.5}$ penalty. The gradient has to calculated offline, i.e. not during the multiplicative update.

Experiments were undertaken to compare the different approaches mentioned above. Similar to the last set of experiments, only Transform $\mathbb{E}4$ is used for this set of spectrogram decompositions. Group sparse penalised variants of the KL-divergence and $\mathcal{C}_{\beta}^{(0.5)}$ are used, with subspace dictionaries of size $P \in \{3,5\}$ employed. δ-thresholding (§3.4.1) is employed and the optimum $\mathcal{F}$-measure for each approach is recorded.

	P = 3		P = 5	
	$\ell^{0.5}_{\perp,0.5}$	$\ell^{0.5}_{2,0.5}$	$\ell^{0.5}_{\perp,0.5}$	$\ell^{0.5}_{2,0.5}$
$\mathcal{C}_{KL}$	77.9	**78.4**	77.8	**78.0**
$\mathcal{C}^{(0.5)}_{\beta}$	77.7	78.3	76.8	77.5

Table 7.5: AMT results in $\mathcal{F}$-measure for different group sparse penalty approaches with two cost functions, $\mathcal{C}^{(0.5)}_{\beta}$ and $\mathcal{C}_{KL}$. Two different groupsizes P also applied.

The results are shown in Table 7.5. A tendency for the performance of G-β-NMD to diminish when $P = 5$ is observable. Similar results were obtained using the G-KL-NMD as the G-β-NMD. G-KL-NMD also performed relatively consistently with respect to group size In terms of the penalty terms employed, both cost function performed slightly better with the $\ell^{0.5}_{2,0.5}$. Nonetheless, both algorithms show improved performance relative to the standard atomic pitch setup, and also in comparison to the ℓ^{p}_{p} penalised approaches described above.

7.3 Discussion

In this chapter an outline of some recent developments in NMF was given. In particular generalised divergences for which multiplicative update algorithms are available were described, before focussing on the forms of these algorithms that have guaranteed monotonic descent properties. Some initial AMT experiments were performed to compare the use of α-divergences for this purpose with the already popular β-divergences. Results showed that the β-divergence, with $\beta = 0.5$ to provide superior results, as previously described in the literature [133] [32].

However, multiplying out the $\mathcal{C}^{(0.5)}_{\beta}$ cost function revealed its similar form to the Pearson distance, affording a simple generalisation, referred to as the η-divergence, of the two cost functions. Further experiments yielded improved AMT results using η-divergence for the STFT, while improved monotonic descent was shown.

Following this, the use of sparse penalties for NMF was explored, and it was found that an ℓ^{p}_{p}-norm penalty was more effective than the standard ℓ_1 penalty for AMT, with a small improvement in results demonstrated, while monotonicity of the penalised approaches was shown. Further to this some further experiments using group sparse penalties were also described, again with improved results. Indeed, these results are superior to those of the stepwise and molecular approaches previously described, and exceed the results of the benchmark experiments by over 6%. However, monotonicity of these algorithms was not shown, and further exploration may be required. A comparison with the use of logarithmic penalties should also be undertaken. While the $\ell_{0.5}^{0.5}$ penalty performed well, it will be worthwhile further exploring the use of other penalty terms of such form.

Chapter 8

Considering Coherence

In the sparse representations literature, coherence is considered a fundamental property of a dictionary, being simple to calculate and yet affording theoretical conditions under which signal recovery can be guaranteed using ℓ_1 minimisation or greedy algorithms such as OMP [128]. The coherence of a dictionary measures the maximum correlation between dictionary atoms and is simply defined as

$$\mu = \max_{i \neq j} |\mathbf{d}_i^T \mathbf{d}_j| \tag{8.1}$$

where $\|\mathbf{d}_n\|_2 = 1 \, \forall n$. A similar measure, the cumulative coherence

$$\mu(k) = \max_i \max_{|\mathcal{J}|=k, i \notin \mathcal{J}} \sum_{j \in \mathcal{J}} |\mathbf{d}_i^T \mathbf{d}_j| \tag{8.2}$$

relates the maximum sum of k correlations relative to one specific atom, and it is apparent from this definition of cumulative coherence that $\mu(k) \leq k\mu$. The classic condition for recoverability of a sparse representation of a signal is the Exact Recovery Condition (ERC) (2.14). Tropp [128] shows that the ERC is always met under the following condition on the dictionary coherence :

$$k < \frac{1}{2}(\mu^{-1} + 1) \tag{8.3}$$

where k is the number of active atoms in the decomposition. Similarly, the ERC is guaranteed when the following condition, using the cumulative coherence, is met

$$\mu(k) + \mu(k-1) < 1 \tag{8.4}$$

which can also be stated in a slightly looser fashion as $\mu(k) < 1/2$. Recently it has been shown that while these conditions may be necessary at the first iteration of OMP, they may be relaxed at later iterations [117]. However it is noted that research in sparse representations is mostly concerned with dictionaries where the coherence is assumed to be reasonably low, consisting of a union of orthogonal bases [128] [23] or using random dictionaries such as in Compressed Sensing [18].

Even in the case of relatively incoherent dictionaries, coherence may be problematic. This has led some researchers to propose preconditioning methods which seek to counter the effects of dictionary coherence, in particular for use with OMP-based methods. This was first considered by Schnass and Vandergheynst [112] who proposed the use of a sensing dictionary, Θ, with a modified OMP algorithm. The authors define a cross-coherence measure $\hat{\mu} = \max_{i \neq j} |\theta_i^T d_j|$ in a similar manner to coherence (8.1), while incorporating the sensing dictionary. Optimisation is then performed to derive a sensing dictionary that reduces the cross-coherence to the lowest possible value. The sensing dictionary is then used in a modified OMP algorithm, in which an atom index is selected according to the maximum correlation between a sensing dictionary atom and the residual, and added to the sparse support. The backprojection step is performed in the normal manner using atoms from the original dictionary $\mathbf{D}$. Improved recovery results using this approach were reported in experiments using a variety of dictionaries, such as Gaussian matrices, and unions of orthogonal bases. Improvements were also reported in terms of the number of atoms that were guaranteed to be recovered using a modified version of the ERC (2.14) that incorporates the sensing dictionary, in a manner similar to the cross-coherence property. A similar sensing dictionary approach, which also uses the modified OMP, is proposed in [59], where the authors propose learning a data-adaptive sensing dictionary. Superior results shown to those given in [112] are reported in [59]. A similar work from the same research group [58] uses a similar methodology in the context of block sparsity, with a modified BOMP algorithm.

The methods described in the paragraph above focus on fixed dictionaries which are relatively incoherent and use quasi-orthogonalisation to decohere the dictionaries. Other researchers have considered coherence in different scenarios. A sparse dictionary learning method, based on the K-SVD algorithm [2] and referred to as IN-KSVD, is proposed in [77]. IN-KSVD seeks to learn dictionaries which are relatively incoherent by augmenting the K-SVD algorithm with an extra decorrelation step which selects, in a greedy manner, pairs of highly correlated atoms and decorrelates them by quasi-orthogonalisation. A large improvement in sparse approximations in audio signals is reported using the IN-KSVD method [77].

A non-negative version of OMP (NN-OMP) is proposed in [16] where the authors note that problems with dictionary coherence are innate to non-negative dictionaries. In light of this the authors propose to use NN-OMP to solve a preconditioned approximation :

$$\mathbf{Ps} \approx \mathbf{PDx} \tag{8.5}$$

where $\mathbf{P}$ can be any invertible matrix. In [16] the authors select $\mathbf{P}$ such that multiplication with a vector $\hat{\mathbf{y}} = \mathbf{Py}$ is equivalent to subtraction of the mean coefficient of the vector from each element $\hat{y}_i = y_i - (\mathbf{1}^T\mathbf{y})/|\mathbf{y}|$. In this case the preconditioner, $\mathbf{P}$, performs a centring of the dictionary and data that introduces negative elements to both and reduces the dictionary coherence. The alternative approximation (8.5) is then performed using the NN-OMP algorithm, and improved performance for non-negative sparse approximation is reported.

The concept of dictionary coherence has not previously been explicitly considered in a musical signal processing context. In a musical dictionary, coherence and harmonic overlap can be considered synonymous, as the non-negativity enforces the summing of all overlapping elements. This equivalence can be seen in Figure 8.1, a graphical representation of the Gram matrix of the single atom dictionary for Transform $\mathbb{E}4$ in which high correlations are seen mostly in lines parallel to the diagonal. Close inspection reveals these lines to be located at consonant musical intervals, for example the octave and the fifth note, in the scale. While the problem of harmonic overlap is often noted in musical signal processing research when decomposition-based methods are used, little research has attempted to explicitly counter the problem, with researchers tending to prefer to focus their attention on the use of different cost functions or incorporating prior information such as temporal structure. In an extensive literature search only one published paper [106] was found that regards the correlation of harmonically related atoms in decomposition-based musical signal processing. In this work [106], a harmonically constrained unsupervised NMF algorithm is used for the purpose of AMT. The dictionary is initialised with 88 atoms, each

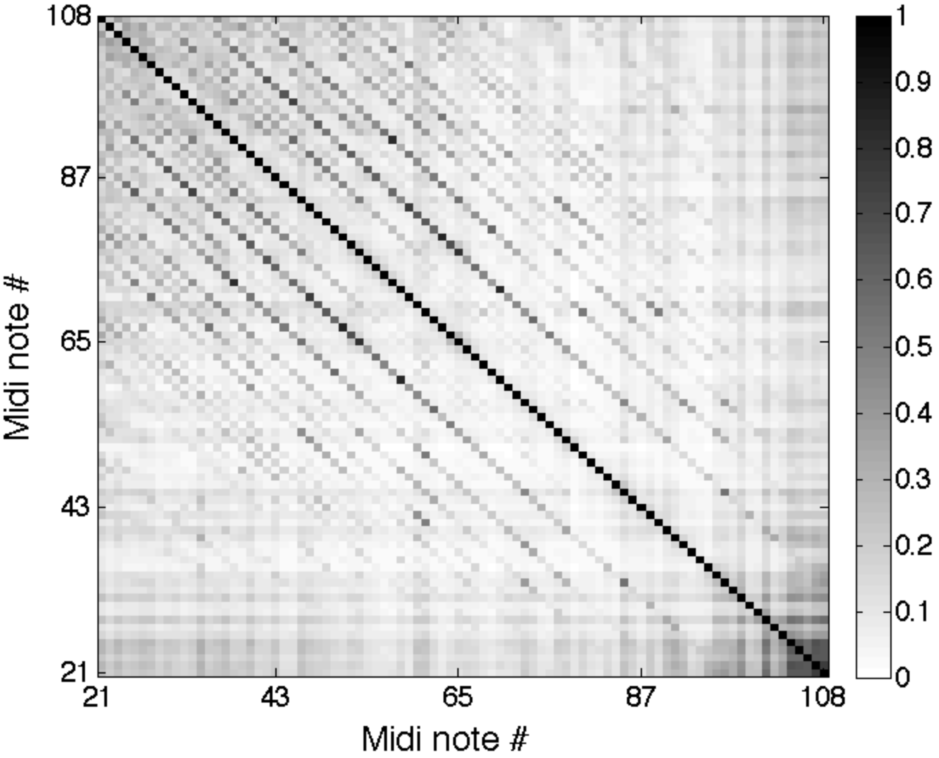

Figure 8.1: Gram matrix of single atom dictionary of Transform $\mathbb{E}4$. Note lines parallel to the diagonal indicating structured high correlations corresponding to musical structure.

representing a note on the piano. All points of each atom which are not multiples of the fundamental frequency of the note the atom represents are initialised to zero. The authors noted that atoms representing pitches that were not present in the signal were likely to become active, whilst their shape became more similar to that of harmonically related pitched elements. Therefore a penalty term $\Phi = \lambda \mathbf{W}\mathbf{X}$ was added to the denominator of the NMF coefficient update where $\mathbf{W}$ is a circulant matrix with high values where dictionary correlation is expected, similar to those seen in Figure 8.1. This penalisation seeks to lower coefficients of atoms that co-occur with other harmonically related atoms. While this method affects only the coefficients of the update it can be understood to implicitly consider the harmonic overlap as this is where the dictionary correlations tend to reside.

In the rest of this chapter the atomic pitch dictionaries of the different transforms (§3.2) used to perform spectrogram decompositions throughout this thesis are analysed in terms of coherence. This coherence analysis is then related to the varying AMT performance seen for these different transforms. The use of row-weighting is then proposed for conditioning of non-negative harmonic dictionaries, with a novel effective coherence measure used to learn a different weighting at each time frame. Experimental results are given which validate the approach, while some further insight into the problem is given by considering row weighting in a noiseless signal, before concluding.

8.1 Considering dictionary coherence for AMT

It would seem that consideration of dictionary coherence might offer little in the context of AMT, at least in the conventional sense of how coherence is generally applied, as a condition on signal recovery. Non-negative musical dictionaries are highly coherent. The inner product of two atoms separated by a musical octave is expected to be of the order 0.7 under a fixed spectral envelope, due to overlap at every other partial of the lower pitched note. Indeed, higher values of coherence are recorded for every dictionary used in this thesis, as can be seen in Table 8.1. In terms of ERC, this would suggest that recovery is never guaranteed when more than one atom is active, even in a noiseless signal. An example of this behaviour was given at the start of Chapter 5.

Different transforms have been compared, using a variety of algorithms, for the purpose of AMT throughout this thesis. Regardless of the algorithm used, be it greedy or gradient descent based, a variation in AMT performance relative to the transform used was observed, with a relatively distinct ordering in the varying performance. The worst AMT performance, for each algorithm, has generally been observed when the STFT transforms have been used. A small improvement is usually observed with the smallest ERBT, Transform $\mathbb{E}1$. Meanwhile the larger dimension ERBT transforms $\mathbb{E}2$-4 provide the best AMT performance with transform $\mathbb{E}4$, the largest dimension ERBT used, generally seen to achieve the best performance. The relative performance difference is particularly enhanced in the OMP algorithms, where a difference in $\mathcal{F}$-measure of 10% was often seen. Using NNLS-based algorithms, the performance differential was milder, of the order of 5%. Transform $\mathbb{E}1$, the ERBT with atom dimension 250, was proposed for AMT in [133] and was designed specifically such that the fundamental frequencies were disjoint on the frequency scale of the transform. This counters a problem seen in the STFT where lower pitched notes have fundamental frequencies which share the same frequency frame in the case of reasonable time resolution being applied. However, the introduction of further larger dimension ERBTs in this thesis has seen a further increase in AMT performance. While separability of fundamental frequency partials is maintained, observation of experimental results leads us to suspect that somehow these larger ERBTs result in dictionaries that are better conditioned. To

Transform	$\kappa(\mathbf{D})$	μ	μ_F	μ_Σ	T-NNLS ($\mathcal{F}$)	NS-OMP ($\mathcal{F}$)
$\mathbb{S}1$	42.87	0.8695	19.82	1356	64.0	69.3
$\mathbb{S}2$	43.13	0.8693	19.93	1367	64.1	68.8
$\mathbb{E}1$	55.35	0.8723	20.11	1434	66.7	70.2
$\mathbb{E}2$	24.36	0.8619	15.74	1082	68.4	75.9
$\mathbb{E}3$	28.27	0.8641	16.49	1142	68.1	74.8
$\mathbb{E}4$	**17.30**	**0.8513**	**13.85**	**934**	**68.6**	**78.3**

Table 8.1: Comparison of matrix condition $\kappa(\mathbf{D})$ (8.6), coherence μ (8.1), global coherence μ_F(8.7) and summed coherence, μ_Σ, (8.8) of dictionaries learnt from the same signals for different transforms and corresponding AMT results for T-NNLS and NS-OMP in terms of $\mathcal{F}$ -measure

this end, a relative comparison in terms of dictionary coherence is now offered.

In order to compare the transform-specific dictionaries, various matrix metrics were calculated for the atomic pitch dictionaries. In this context, it is important to recall that each atom in a given dictionary should be similar, at some level, with the correspondingly pitched atom from another dictionary as all dictionaries are learnt from the same set of isolated note signals. The first measure considered is the matrix condition number:

$$\kappa(\mathbf{D}) = \sigma_{max}(\mathbf{D})/\sigma_{min}(\mathbf{D}) \tag{8.6}$$

which gives the ratio of the largest and smallest singular values of the matrix. The matrix condition, being an eigenvalue measure can be considered to have some correspondence with the Restricted Isometry Property (RIP) condition (2.17) [18], which measures the maximum deviation from unity of the eigenvalues of a submatrix of size k, when all atoms have unit ℓ_2 norm. While the matrix condition considers the eigenvalues of the matrix, the RIP considers the eigenvalues of a submatrix. However, the eigenvalues of a matrix are known to bound those of a submatrix [83]. Other matrix measures are then applied to the Gram matrix $\mathbf{G} = \mathbf{D}^T\mathbf{D}$. The coherence value μ (8.1) and the cumulative coherence $\mu(k)$ (8.2) for $k \in \{1,...,87\}$ are also calculated. While these measures are used to prove the theoretical performance bounds of sparse methods with incoherent dictionaries, at some level they may not be totally indicative of the coherence in the whole dictionary as the coherence relates to one pair of atoms, while cumulative coherence relates the correlation of one atom with a few other atoms. To give some measure of all coherences in the dictionary a *global coherence measure*

$$\mu_F = ||\mathbf{G} - \mathbf{I}||_F \tag{8.7}$$

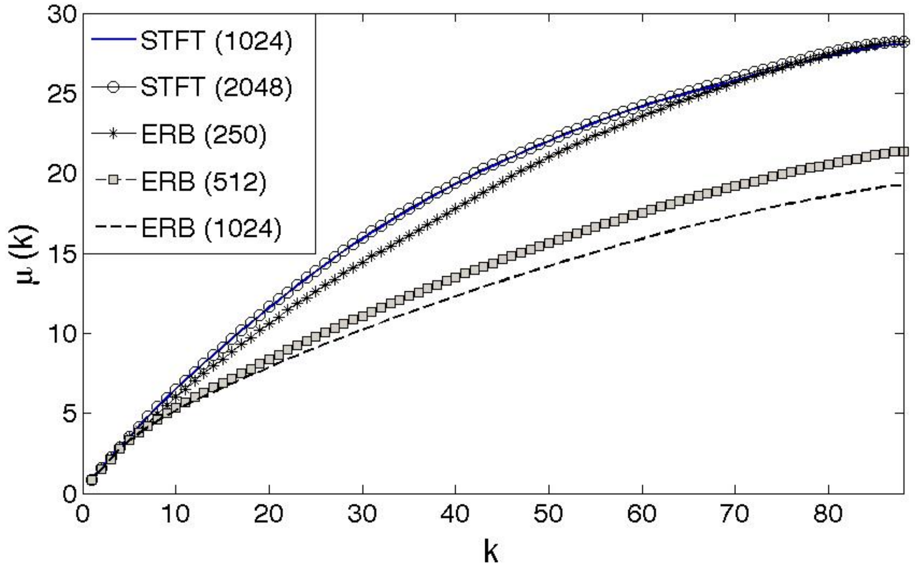

Figure 8.2: Cumulative coherence $\mu(k)$ plotted against k for dictionaries learned from the same dataset in several transforms

and *summed coherence measure*

$$\mu_{\Sigma} = \sum_{j \neq i} [\mathbf{G}]_{i,j} \tag{8.8}$$

are defined for comparison, where $\|\mathbf{X}\|_F$ is the Frobenius norm of $\mathbf{X}$.

In Table 8.1 the AMT results in terms of $\mathcal{F}$-measure for T-NNLS and for NS-OMP are given for each transform alongside the matrix measures, described in the previous paragraph, of the corresponding dictionaries. While the NN-NS-OMP is a group sparse algorithm, using a dictionary formed from a union of pitched subspaces, the same ordering of performance is seen as for the T-NNLS, as has been observed for other group sparse algorithms throughout the thesis. The cumulative coherence values relating to the dictionaries in each transform are plotted in Figure 8.2. Some distinctive patterns are obvious in these results. While the coherence value μ seems relatively uninformative in this context, being quite similar for all algorithms, a pattern in the ordering of the other tabulated matrix measures is seen that is matched in the coherence parameter. The smallest ERBT, Transform $\mathbb{E}1$, has the highest measures in all cases, followed by the two STFTs, which are seen to be very similar to each other, both in terms of AMT results and matrix measures. In terms of AMT performance, Transform $\mathbb{E}1$ performs better than the STFTs, in contradiction to the ordering of the tabulated measures; however $\mathbb{E}1$ is seen in Figure 8.2 to have a lower cumulative coherence than the STFTs. Transforms $\mathbb{E}2$-4, the larger dimension ERBTs, are seen to perform significantly better in terms of AMT, and the corresponding matrix measures are also significantly lower than those of the first three transforms. A strict ordering is also seen amongst these ERBTs with $\mathbb{E}4$ performing best and having the lowest matrix measures for all measures, followed by $\mathbb{E}2$ and $\mathbb{E}3$. It is interesting that the STFTs display better tabulated matrix measures than the small dimension ERBT, Transform $\mathbb{E}1$, when AMT performance is seen to suffer in relation, while a strict correspondence is seen amongst the ERBTs themselves. Possible explanations may be given by the fact Transform $\mathbb{E}1$ displays lower cumulative coherence than the STFTs, while the dictionary of Transform $\mathbb{E}1$ has disjoint fundamental frequencies, unlike the STFT dictionaries. From this comparison of dictionaries it may be concluded that AMT performance may be somewhat related to coherence measures. If this is the case it would seem appropriate to attempt to leverage the dictionary coherence in order to further improve AMT. In the next section, an attempt to condition a harmonic dictionary to improve AMT performance is described.

8.2 Conditioning a Harmonic Dictionary

The relationship of dictionary coherence to AMT performance has been established in the previous section, and it would seem worthwhile to manipulate the dictionary coherence, if possible, to improve performance. The AMT problem has distinct characteristics that discourage the use of general methods for coherence-based preconditioning. Each atom in a musical dictionary used for AMT is semantically meaningful, representing a particular note, and the dictionaries used are often undercomplete. Several coherence-based preconditioning approaches in the sparse representations literature [112] [59] [77] use quasi-orthogonalisation to reduce the coherence. However, in the case of undercomplete dictionaries, such as musical dictionaries, it is trivial to form an equivalent orthogonal dictionary. Observation of the orthogonalisation of a musical dictionary quickly indicates the unsuitability of such an approach. Atoms in the transformed dictionary are far removed from corresponding atoms in the original dictionary. Many negative elements, some of which display very large coefficients, are introduced to the dictionary particularly in areas of harmonic overlap, while some very large coefficients also appear in high frequency elements where there is little energy. The high dictionary coefficients can introduce instability to the problem, effecting the representation capability of the dictionary. From these observations it would seem that the structure inherent in the dictionary should be maintained as much as possible.

Another candidate preconditioning approach, specifically proposed for non-negative dictionaries, is the centring approach of [16], which can be considered to maintain the structure of the dictionary. However, this approach was observed not to be effective in the context of the problem at hand. Some experiments using NN-OMP with this preconditioning approach were performed. While this method was seen to be effective using randomly generated dictionaries, detrimental performance was observed in the context of AMT. It can be hypothesised that the harmonic overlap is not effected by such a centring, or alternatively that the sparsity of the musical dictionary itself may not be amenable to a centring approach. In light of these observations, it would seem sensible to employ an approach that conserved the non-negativity and structure of the dictionary, while aiming to counter the effect of harmonic overlaps.

It is proposed to use a row weighting, or scaling approach. Row weighting applies a different scale to each dimension of a vector, and can be effected through multiplication with a diagonal matrix. Row weighting is known to effect a least squares solution [50], and is a commonly used approach in methods such as Total Least Squares [81]. Typically, the goal of row-weighting is to better condition a problem. Many row weighting schemes are possible, such as simple row weighting [50] in which each row of a matrix is scaled such that all rows contain the same largest entry. A further example is given in [135] where the authors propose to use projected gradient descent to perform unsupervised NMF with the Kullback-Leibler divergence cost function. A row-weighting that equalises the sum of each row of is employed to condition the matrix to be decomposed thereby discouraging problems with gradient scaling. It is noted in [50] that while row scaling strategies may be effective, a generic approach to scaling is not available and the approach taken should vary as per application.

For the specific problem at hand it is proposed that the row weighting is determined by coherence. In particular, a row weighting that lessens the effective coherence in a decomposition is sought. Using a row weighting leads to the modified approximation

$$\mathbf{Ws} \approx \mathbf{WD}\hat{\mathbf{x}} \tag{8.9}$$

where $\mathbf{W}$ is a diagonal matrix, with all diagonal entries $w_{m,m} > 0$, which scales each row of $\mathbf{D}$ and $\mathbf{s}$, and $\hat{\mathbf{x}}$ is the solution vector to the weighted problem. The form of this approximation is similar to that employed in the centring approach used in tandem with the NN-OMP in [16]; however the form of the preconditioning matrix $\mathbf{W}$ differs.

While it would be desirable to find a single weighting matrix, $\mathbf{W}$, that would enhance many

Algorithm 8.1 Post-preconditioning approach

```
Input
    $\mathbf{D} \in \mathbb{R}^{M \times N}$,   $\mathbf{s} \in \mathbb{R}^{M}$
Decompose
    $\mathbf{x} = \arg\min_x ||\mathbf{s} - \mathbf{Dx}||_2^2 \quad s.t. \quad \mathbf{x} \geq 0$
Find row-weighting matrix $\mathbf{W}$ using (8.11)
Apply weighting and decompose
    $\hat{\mathbf{x}} = \arg\min_x ||\mathbf{Ws} - \mathbf{WDx}||_2^2 \quad s.t. \quad \mathbf{x} \geq 0$
```

decompositions, prior experiments indicated that this would not be a very effective approach. While a less coherent dictionary is derivable by scaling in a manner that reduces various coherence measures seen previously in this chapter (8.1) (8.2) (8.7) (8.8), ultimately the connectedness of the musical scale limits the effectiveness of such an approach, as a local improvement may be detrimental in other areas.

Therefore, it is proposed to take an alternative approach, in order to adapt to signal vectors individually. Hence a *post-preconditioning* approach is proposed, as outlined in Algorithm 8.1. The term post-preconditioning refers to the fact that an initial decomposition is made using NNLS as outlined. Based on the solution to this decomposition, $\mathbf{x}$, a diagonal conditioning matrix $\mathbf{W}$ is derived, and a subsequent row weighted decomposition (8.9) is performed, giving the final solution vector $\hat{\mathbf{x}}$. While NNLS is proposed for performing the decompositions in Algorithm 8.1, different decomposition methods may also be used. In terms of a spectrogram decomposition these steps are performed at each time frame, deriving a different weighting matrix, $\mathbf{W}^t$, based on each individual solution vector $\mathbf{x}_t$, leading to a new approximation $\mathbf{W}^t\mathbf{s}_t \approx \hat{\mathbf{D}}\hat{\mathbf{x}}_t$ being performed, where $\hat{\mathbf{D}}^t = \mathbf{W}^t\mathbf{D}$.

Effective Coherence

The proposed post-preconditioning approach requires that the conditioning matrix, $\mathbf{W}$, is learnt at each time frame, and it is proposed to emphasise the coherence of atoms that have reasonably high coefficients. Before further description of the learning approach is given, it is necessary to consider that multiplication of a dictionary atom, such that $\hat{\mathbf{d}}_n = \mathbf{W}\mathbf{d}_n$, will generally cause the norm of the atom to change such that $||\hat{\mathbf{d}}_n||_2 \neq 1$. This transformation results in the Gram matrix $\Phi = \hat{\mathbf{D}}^T\hat{\mathbf{D}}$. As coherence measures presume that $\mathbf{d}_n = 1 \forall n$, it is necessary to consider the Gram matrix of the equivalent normalised dictionary Θ, which can be conveniently derived from Φ:

$$\Theta^{[2]} = \Phi^{[2]} \oslash [\mathbf{h}\mathbf{h}^T] \tag{8.10}$$

where $\oslash$ denotes elementwise division, $\mathbf{X}^{[\cdot]}$ indicates elementwise exponentiation of $\mathbf{X}$ and $\mathbf{h} = \text{diag}(\Phi)$. With Θ now defined, a new *effective coherence measure* is proposed:

$$\mu^e = \mathbf{x}^T[\Theta^{[2]} - \mathbf{I}_N]\mathbf{x} \tag{8.11}$$

where $\mathbf{I}_N$ is an identity matrix of dimension $N \times N$. This effective coherence measure emphasises coherence in atoms that are active relative to the product of their coefficients and is used in order to derive a suitable value for $\mathbf{W}^t$ such that

$$\mathbf{W} = \arg\min_{\mathbf{W}} \mu^e \quad s.t. [W]_{m,n \neq m} = 0 \tag{8.12}$$

which can be performed using descent based methods. While it is possible to update $\mathbf{x}$ through iterations of the descent method, the choice is made to keep $\mathbf{x}$ fixed which necessitates the use of the normalised Gram matrix Θ in the effective coherence measure (8.11). The possible values of $[W]_{m,m}$ are bounded above and below in order to prevent trivial solutions and to maintain the structure of the original problem. Hence the projected gradient descent method is used to estimate $\mathbf{W}$. As $\mathbf{x}$ is kept constant the gradient of the effective coherence measure relative to any given dimension, m is given by :

$$\frac{\partial \mu^e}{\partial w_m} = \sum_{i \neq j} \mathbf{x}_i \mathbf{x}_j \frac{\partial [\Theta^{[2]}]_{i,j}}{\partial w_m} \tag{8.13}$$

where $w_m = [W]_{m,m}$, $[X]_{i,j} = x_{i,j}$ denotes the element in the ith row and jth column of $\mathbf{X}$. The gradient term $\frac{\partial [\Theta^{[2]}]_{i,j}}{\partial w_m}$ can be expressed in terms of Φ, individual dictionary elements and the norms of the modified atoms:

$$\begin{aligned}\frac{\partial [\Theta^{[2]}]_{i,j}}{\partial w_m} = \frac{2w_m \Phi_{i,j}}{\|\hat{\mathbf{d}}_i\|_2^4 \|\hat{\mathbf{d}}_j\|_2^4} \times \quad & \{2\|\hat{\mathbf{d}}_i\|_2^2 \|\hat{\mathbf{d}}_j\|_2^2 d_{m,i} d_{m,j} - \\ & \Phi_{i,j}(\|\hat{\mathbf{d}}_i\|_2^2 d_{m,j}^2 + \|\hat{\mathbf{d}}_j\|_2^2 d_{m,i}^2)\}\end{aligned} \tag{8.14}$$

or alternatively, in matrix form

$$\frac{\partial \Theta^{[2]}}{\partial w_m} = 2w_m \Phi \otimes \mathbf{X} \otimes \left[[2\mathbf{A}^m \otimes \mathbf{A}^{mT}] - \Phi \otimes [\mathbf{Z}^m + \mathbf{Z}^{mT}]\right] \tag{8.15}$$

where $\otimes$ denotes the Hadamard elementwise multiplication, $\mathbf{X} = [\mathbf{h}^{[2]}\mathbf{h}^{[2]T}]^{[-1]}$; $\mathbf{A}^m = \mathbf{d}^m\mathbf{h}^T$ and $\mathbf{Z}^m = \mathbf{d}^{m[2]}\mathbf{h}^T$, where $\mathbf{d}^m$ is the mth row of $\mathbf{D}$. After $\mathbf{W}$ is estimated, a solution to the weighted approximation (8.9) is calculated, using NNLS or another decomposition method, giving $\hat{\mathbf{x}}$, the new coefficient matrix, from which the piano roll can be derived.

Some insight into how this method is proposed to work is offered. Firstly, consider a noiseless signal formed from a non-negative superposition of atoms from a non-negative dictionary, which is full rank and overdetermined. It is then assumed [114] that NNLS will correctly recover the support and coefficients. Furthermore the effect of row weighting in this noiseless case does not effect the solution as shown in the proceeding Lemma.

Fact 3. *Given* $\mathbf{s} = \mathbf{D}\mathbf{x}$*, and* $\mathbf{W}$ *is any diagonal matrix,* $\mathbf{x} = \arg\min_x ||\mathbf{W}\mathbf{s} - \mathbf{W}\mathbf{D}\mathbf{x}||_2^2$.

Proof. This is simply proved by considering the scalar case where $s = \sum_i x_i$. Applying a scalar weighting gives $ws = \sum_i wx_i$ which is easily generalised to the vector case. □

However, when the signal is noisy error may be introduced into the support selection, and even if the support is fixed two solution vectors $\mathbf{x}$ and $\hat{\mathbf{x}}$ can differ. In musical spectrogram decompositions the error is often introduced as false detections, many of which are observed to be of atoms of consonant pitches. This can be explained simply in terms of a noisy signal with one active pitch. If the correspondingly pitched atom does not exactly fit the signal, a residual signal can be expected to be left after projection onto the correctly pitched atom. Most of the energy in this residual can be expected to reside in the harmonic partials of the correctly pitched atom, and further atoms may be selected in the decomposition. It is likely that extra atoms will contain energy in the largest elements of the residual signal, which are likely to reside in the harmonic partials of the correctly pitched atom. Hence the incorrectly selected atoms are likely to have harmonic overlap with the correctly selected atom.

The proposed effective coherence measure (8.11) is seen to include the Gram matrix, emphasising the coherence of active atoms and the covariance of the coefficient vector from the initial decomposition. By containing both these elements in the one measure, it is hoped to learn a weighting that targets atoms that are both active in the decomposition and correlated to one another. By scaling the individual dimensions of these supported atoms in a coherence-aware manner, it is hoped to lessen the importance of dimensions which overlap in atoms in the support, and thereby reducing the coefficients of falsely detected atoms.

Transform	NNLS	WNNLS	β-NMD	Wβ-NMD
$\mathbb{E}1$	66.7	69.7	71.9	73.7
$\mathbb{E}2$	68.4	72.6	74.9	77.0
$\mathbb{E}4$	68.6	73.7	75.2	77.9

Table 8.2: Results from AMT experiments in terms of $\mathcal{F}$-measure (%) comparing the weighted methods (WNNLS and Wβ-NMF) against the original algorithms.

8.2.1 Experiments

AMT experiments were run using the standard dataset (§3.1) to assess the effect of the proposed coherence reducing row-weighting scheme. The experiments were limited to the ERB transforms $\mathbb{E}1,2\&4$. For the considered transforms, NNLS was used to perform the initial spectrogram decomposition, and a subset of the atoms was selected by thresholding the spectrogram decomposition based on its maximum value with $\delta = 0.01$ (3.3). The threshold is applied to lessen the computational expense, and is far below the optimum value optimum value of δ observed in previous AMT experiments, and it is assumed that its use will have little effect on either the weighting or the final transcription output. This active set of atoms after thresholding was used to calculate the weighting matrix $\mathbf{W}^t$ at each spectrogram frame, and $\mathbf{W}$ was bounded to have values in the interval [0.4, 1.6] giving an extremal weighting factor of 4.

Projected gradient descent was implemented to calculate $\mathbf{W}^t$. After the gradient was calculated at each dimension using (8.13), a line search was performed, starting with a small initial stepsize which was doubled at each iteration until improvements in the cost function ceased to be produced. The gradient was recalculated before returning to the linestep procedure, and the algorithm stopped when the first step of the linesearch failed to reduce the relative cost function by a factor of 10^{-9}, returning the weighting matrix $\mathbf{W}^t$.

A NNLS decomposition was performed on the transformed signal $\mathbf{W}^t\mathbf{s}_t$ using the transformed dictionary $\hat{\mathbf{D}}$ to derive the Weighted NNLS (WNNLS) coefficient matrix $\hat{\mathbf{X}}$. δ-thresholding (§3.4.1) was performed on $\hat{\mathbf{X}}$ for a range of values of $\delta \in \{10,...,50\}$dB and the recorded results described the optimum $\mathcal{F}$-measure at δ_{opt}. As mentioned earlier, other algorithms can be used to perform either the pre- or post-weighting decomposition. Similar weighted decompositions were also performed using β-NMD [27] using the weightings derived from the initial NNLS decomposition and referred to Wβ-NMD. The value of $\beta = 0.5$ was used as this setting was seen to be the optimum value for AMT on a similar dataset [133], and also in experiments on the ERBTs in Chapter 7.

The results for these experiments are shown in Table 8.2, where it can be observed that the weighted methods outperform their unweighted counterparts. In particular, the weighted NNLS method, WNNLS, shows improvements of up to 5.1% relative to NNLS, with improvements more marked in the transforms that already perform better in the unweighted case. In the case of the β-divergence, the improvements of the weighted approach, Wβ-NMD, relative to β-NMD are smaller, reaching 2.7% in the case of transform $\mathbb{E}4$. It is noted that the results for β-NMD without weighting are seen to improve more significantly than NNLS as the dimension of the ERBT is increased. Overall it is seen that employing the row weighting with transform $\mathbb{E}4$ results in an improvement of 7% and 6%, for the NNLS and the β-NMD respectively, relative to using an unweighted decomposition with transform $\mathbb{E}1$, as in the benchmark experiments (§3.5).

A graphical demonstration of the evolution of Precision, Recall and $\mathcal{F}$-measure across a range of values of δ is shown for NNLS and WNNLS for Transform $\mathbb{E}4$. Inspection of comparative performance of the two approaches shows that Recall is slightly improved in the weighted case. Meanwhile a large variation in Precision using these two approaches is observed, with the row weighted approach performing better, with a maximum improvement of 10.4% seen at 32dB. This validates the approach taken, which sought to eliminate false positives induced through harmonic overlapping. This increase in Precision also results in an increase in $\mathcal{F}$-measure, and δ_{opt} is seen to be 2dB lower for the WNNLS. Further to this the improvement using WNNLS relative to NNLS are seen to be statistically significant. In all individual songs an increase in $\mathcal{F}$-measure is seen at the value of δ_{opt} selected for all songs. The mean improvement of $\mathcal{F}$-measure is 5.0% with a standard deviation of 2.1%.

An example data weighting is seen in Figure 8.4 where the weightings are seen to be often set to extremes, which is common. While significant downwards scaling of the signal can be seen in this example, this is not necessarily indicative, as upwards scaling is also likely. However the relative flattening effect seen in Figure 8.4 is a general phenomenon, as the correlation between two atoms is most significantly reduced when coincidentally large elements are scaled down.

8.3 Discussion

In this chapter, an analysis of the AMT problem in terms of coherence was given. While the use of different transforms has previously been proposed in AMT and musical signal processing [133] [48], in particular to counter the problems of close spacing of low-pitched fundamental fre-

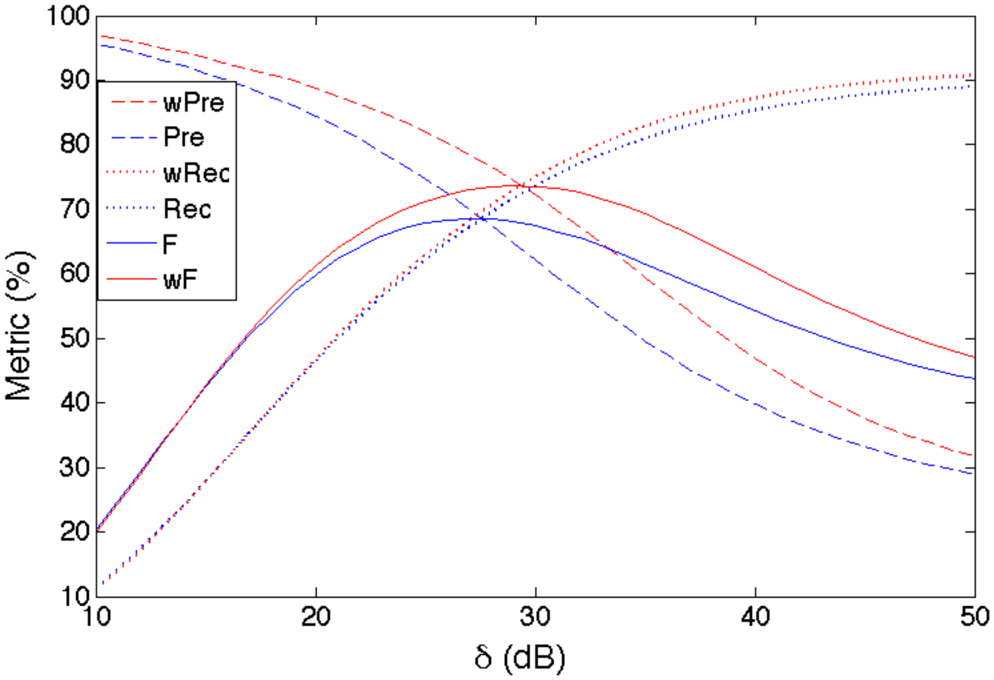

Figure 8.3: Evolution of $\mathcal{P}, \mathcal{R}, \mathcal{F}$ measures for NNLS and W-NNLS relative to thresholding parameter δ.

quencies, the perspective of analysing these transforms in terms of dictionary coherence is new. Indeed, this can be considered an unconventional perspective to take, considering that coherence is usually associated with recovery conditions for quasi-incoherent dictionaries. However, taking this perspective was demonstrated to be a fruitful endeavour. Throughout the previous chapters in this thesis, a consistent variation in AMT results relative to the time-frequency transforms used was observed, and this coherence analysis was seen to go some way to explaining this variation. The better performing transforms, the larger dimension ERBTs, were observed to be less coherent than the other transforms.

The realisation that improved dictionary coherence led to enhanced AMT performance motivated an attempt to further enhance performance by conditioning the dictionary in a coherence-aware manner. Previous methods for conditioning dictionaries were described and observations of their unsuitability to the problem at hand were related. In light of this, a new coherence-based row-weighting method was proposed. It is noted how coherence in a musical dictionary is related to harmonic overlap, which is often mentioned as problematic in the AMT problem [125] [106]. However in terms of decomposition-based methods little research has been performed with the stated aim of countering this problem. Conversely, a lot of research in decomposition based methods has focussed on temporal evolution of the signal [5] [134] [9], enforcing continuity in the decomposition. Here an alternative philosophy is originally proposed. Time continuity and smoothness may exist in the signal. However, it can be considered that the numeric instability of the decomposition is actually the problem, perturbing the underlying smoothness. Coherent, or equivalently ill-conditioned dictionaries, coupled with temporal evolution of note spectra result in the appearance of falsely detected notes, many with harmonic sympathy to the actual notes played. The work presented here demonstrates that it may be possible to tackle the problem at hand, in a more direct fashion, and in a reasonably principled manner.

While the work presented here is a solid progression over the state-of-the-art decomposition methods, it can also be viewed as a first step in a new direction for tackling the AMT problem, and possibly other problems which display similar characteristics such as non-negativity and structured overlapping. Coherence has been used as a parameter to perform the row-weighting, and while this has been seen to be effective, alternative strategies may be envisaged. It may be worthwhile to further consider the multiplicative noise model, previously assumed in [1] [40] in terms of NMF models. The coherence-based method presented may inadvertently be adept at attenuating the effects of multiplicative noise in the matrix decompositions, by weighting the frequency elements where such noise is likely to occur.

This completes the exploration of decomposition methods in this thesis and the next chapter focusses on unsupervised and semi-supervised learning using Non-negative Matrix Factorisation.

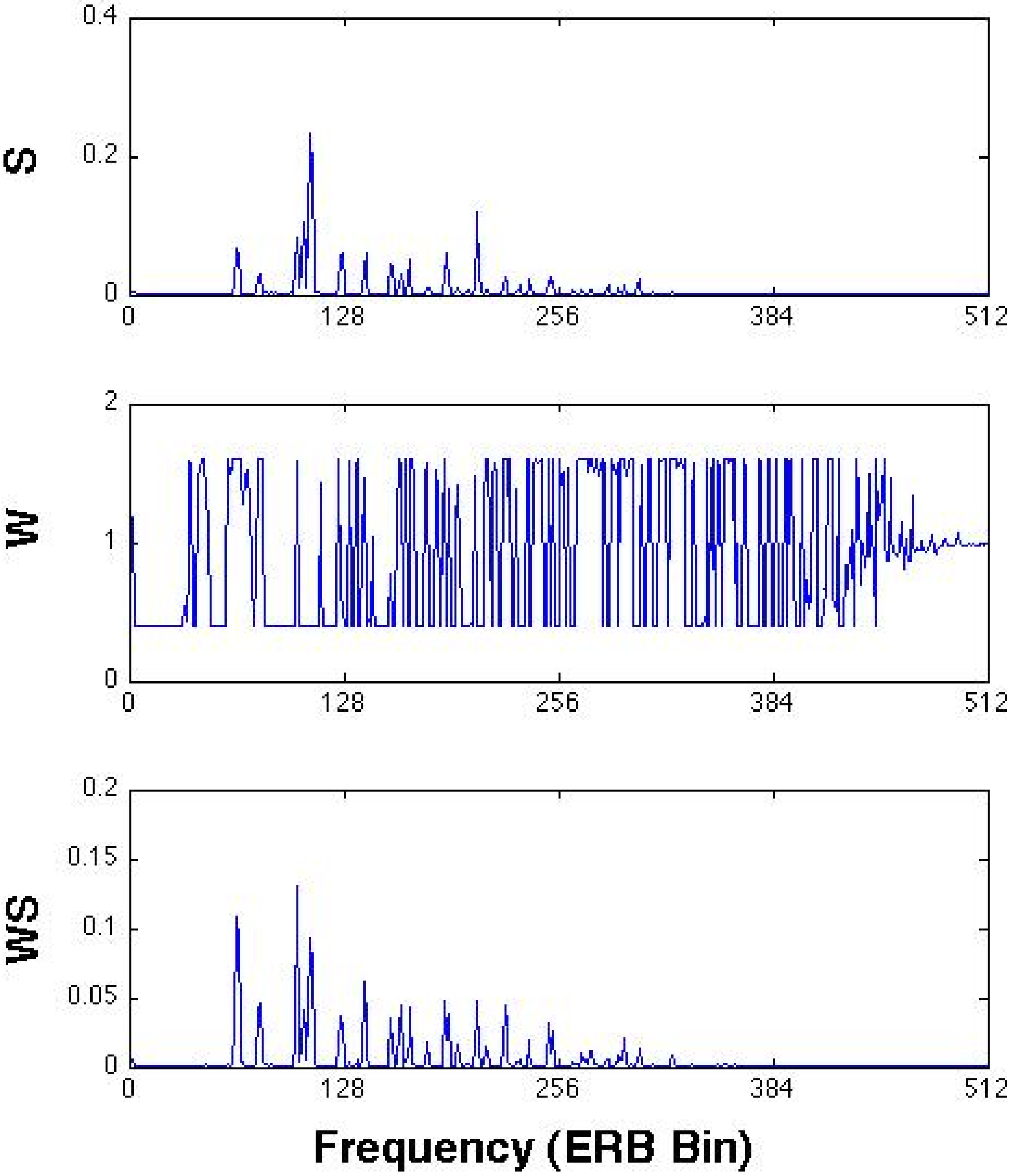

Figure 8.4: Example data point $\mathbf{s}_n$ (top), weighting vector $\mathbf{w}^n$ (centre) and weighted datapoint $\mathbf{W}^n\mathbf{s}_n$ (bottom).

Chapter 9

Sparse NMF

Non-negative Matrix Factorisation (NMF) is a popular tool in musical signal processing, which seeks a low rank approximation of a non-negative matrix. This factorisation is often referred to as a ‘parts-based’ representation as it is hoped that meaningful elements of the matrix, or signal, will be separated. Given a matrix, $\mathbf{S} \in \mathbb{R}^{M \times T}$, in which all entries are non-negative, NMF seeks to find a dictionary matrix, $\mathbf{D} \in \mathbb{R}^{M \times N}$, and an activation matrix, $\mathbf{X} \in \mathbb{R}^{N \times T}$, both of which are also completely non-negative, such that

$$\mathbf{S} \approx \mathbf{DX}. \tag{9.1}$$

NMF was originally proposed as Positive Matrix Factorisation (PMF) by Paatero and Tapper [97], who proposed to effect the approximation (9.1) by minimising a constrained Euclidean distance cost function:

$$\mathcal{C}_E = ||\mathbf{S} - \mathbf{DX}||_F^2 \quad s.t.\, \mathbf{D}, \mathbf{X} \geq 0 \tag{9.2}$$

where $\mathbf{D}$ and $\mathbf{X}$ are both unknown, thereby casting the NMF problem as

$$\mathbf{D}, \mathbf{X} = \arg\min_{D,X} ||\mathbf{S} - \mathbf{DX}||_F^2 \quad s.t.\, \mathbf{D}, \mathbf{X} > 0. \tag{9.3}$$

However the problem (9.3) is not convex in both variables simultaneously and is regarded to suffer from the presence of many local minima [61] [64]. Hence, a commonly used approach to NMF problem is to iterate through alternating projections, whereby one variable is fixed while the other is updated, then vice versa. In the case of PMF [97], this methodology is proposed through the use of Alternating Non-negative Least Squares (ANLS) projections

$$\mathbf{X} \longleftarrow \arg\min_{X} ||\mathbf{S}-\mathbf{DX}||_F^2 \quad s.t. \, \mathbf{X} \geq 0 \tag{9.4}$$

$$\mathbf{D} \longleftarrow \arg\min_{D} ||\mathbf{S}^T - \mathbf{X}^T\mathbf{D}^T||_F^2 \quad s.t. \, \mathbf{D} \geq 0 \tag{9.5}$$

whereby the NMF problem is solved through the iterative solution of convex subproblems.

While NMF was first introduced in the guise of PMF, it was popularised by Lee and Seung [74] who proposed using fast multiplicative gradient descent updates to perform the alternating projections as an alternative to the expensive NNLS (9.4) (9.5) calculations required for ANLS. Multiplicative updates were proposed in [74] for both the Euclidean distance (9.2) and Kullback-Leibler divergence cost funtions. The multiplicative updates were formed by selecting a fixed stepsize that enabled the normal additive update gradient descent to be rearranged into a multiplicative form. For instance, a stepsize of $\nu = \mathbf{X} \oslash \mathbf{D}^T\mathbf{DX}$ was inserted into the standard gradient descent update for the coefficient matrix

$$\mathbf{X} \longleftarrow \mathbf{X} + \nu \times [\mathbf{D}^T(\mathbf{S}-\mathbf{DX})] \tag{9.6}$$

leading to the multiplicative update, in the case of the Euclidean distance cost function

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes [\mathbf{D}^T\mathbf{S}] \oslash [\mathbf{D}^T\mathbf{DX}] \tag{9.7}$$

where $\otimes$ denotes elementwise multiplication and $\oslash$ denotes elementwise division. A similar formulation leads to the update for the dictionary :

$$\mathbf{D} \longleftarrow \mathbf{D} \otimes [\mathbf{SX}^T] \oslash [\mathbf{DXX}^T] \tag{9.8}$$

which can be seen as a transposed version of (9.7), similar to how (9.5) is equal to a transposed formulation of (9.4). It is shown in [74] that the updates (9.7) and (9.8) are both non-increasing in the Euclidean distance cost function (9.2). However, problems with convergence using multiplicative updates have been noted [4] and the ANLS framework is considered more robust [65].

While other cost functions, such as β-divergence [27], can be used for NMF [25] [27] [85], the Euclidean distance NMF is still popular for many applications [22]. The computational load of using NNLS projections for the subproblems in ANLS has been noted and many variants of the ANLS/PMF algorithm have been proposed using different strategies for the individual NNLS problems. For instance, projected gradient descent methods are proposed in [62]. Optimised active set methods, which take advantage of similarity between individual representation vectors coupled with the optimisations used in the Fast-NNLS algorithm [14] are used in [64]. The authors of [64] further propose the use of a block pivoting algorithm for ANLS [65]. Block pivoting is an active set algorithm that allows several atoms to be added and removed from the active set at each iteration, and was first proposed for NNLS in [105]. The authors of [65] propose to augment the block pivoting method of [105] by solving for multiple coefficient vectors simultaneously, in a similar fashion to that employed in [64]. A more recent development has seen coordinate descent based methods used for the individual NNLS problems. This approach, referred to as Hierarchical Alternating Least Squares (HALS), was first proposed in [26], with each alternating projection consisting of cyclic sequential coordinate descents.

Incorporating sparsity

A noted feature of NMF factorisations is that they tend to be sparse [74] [55]. This somewhat echoes the fact that NNLS decompositions tend to be sparse [114]. This sparsity is a desirable property for many applications and several attempts have been made in NMF research to control the sparsity level of a signal, or matrix, representation. Similar to the sparse representations literature, a typical approach to Sparse NMF is to introduce a sparsity penalty term to, for instance, the Euclidean distance :

$$\mathcal{C}_S = ||\mathbf{S} - \mathbf{D}\mathbf{X}||_F^2 + \lambda \sum_{t=1}^{T} ||\mathbf{x}_t||_p \quad s.t. \quad \mathbf{D},\mathbf{X} \geq 0 \tag{9.9}$$

where λ is a parameter controlling the sparsity and $||.||_p$ is an ℓ_p vector norm. Typically, in the NMF literature, as in the sparse representations literature, a ℓ_1 norm is used for the penalty term applied to the activation matrix, in which case the cost function (9.9) can be seen as a non-negative matrix variant of the LASSO [124], or Basis Pursuit Denoising (BPDN) [23].

The ℓ_1 penalty term was first introduced to NMF by Hoyer [56] using a multiplicative update

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes [\mathbf{D}^T\mathbf{S}] \oslash [\mathbf{D}^T\mathbf{D}\mathbf{X} + \lambda] \tag{9.10}$$

which simply augments the Euclidean coefficient update (9.7) with the sparsity parameter λ. In the same work an option to enforce the dictionary to be sparse, using an iteration of projected gradient descent, was proposed. A later work by the same author proposed a different strategy for enforcing sparsity in a coefficient vector. A sparsity measure incorporating the number of atoms in the dictionary, the ℓ_1 and ℓ_2 norms of an activation vector was proposed. Iterative projections of each coefficient vector were performed until the predefined sparsity measure was met. An alternative multiplicative update strategy for the ℓ_1 penalised coefficient matrix update using subtraction rather than addition:

$$\mathbf{X} \longleftarrow \mathbf{X} \otimes \left[\max\left(\mathbf{D}^T\mathbf{S} - \lambda, \varepsilon\right)\right] \oslash \left[\mathbf{D}^T\mathbf{D}\mathbf{X} + \varepsilon\right] \tag{9.11}$$

was proposed in [119] and [24] where ε is a small value added to prevent divide by zero errors

A non-negative variant of Iterative Soft Thresholding (IST) [138] was employed to perform the coefficient matrix update using the ℓ_1-penalised cost function (9.10) for a sparse version of NMF, called Generalised Morphological Component Analysis (GMCA) in [107]. The IST algorithm is known to converge for the LASSO or BPDN [138] and the authors also suggest a tempering approach, initialising the algorithm with a large value of λ that is slowly decreased at each iteration. In the ANLS framework, Kim and Park [64] [63] proposed to apply a squared ℓ_1-penalty term by performing the NNLS approximation:

$$\mathbf{X} \longleftarrow \arg\min_{X} \left\| \begin{bmatrix} \mathbf{M} \\ \mathbf{0}_{1\times N} \end{bmatrix} - \begin{bmatrix} \mathbf{D} \\ \sqrt{\lambda}\,\mathbf{1}_{1\times K} \end{bmatrix} \mathbf{X} \right\|_F^2 . \tag{9.12}$$

Non-negative dictionaries inherently have correlated atoms [16], and it is shown in [124] that ℓ_1 penalisation can perform worse than ℓ_2 penalisation, or ridge regression, in the case of correlated atoms in an undercomplete dictionary. In more specific terms of non-negative sparse approximations, it has been shown recently that Thresholded NNLS outperforms non-negative ℓ_1-minimisation approaches such as LASSO/BPDN, due to the innate regularisation of the non-negative constraint [114]. It would therefore seem questionable if ℓ_1 is an appropriate term for NMF. Indeed, in the context of NMF, and in particular the GMCA algorithm [107] an iterative strategy using hard thresholding was often seen to perform better than the IST approach, with the authors noting that the hard thresholding strategy tends towards an ℓ_0 penalty.

Algorithm 9.1 NN-K-SVD

```
 Input
    $\mathbf{S} \in \mathbb{R}^{M \times T}, \quad k$
 Initialise
    $\mathbf{X}^0 = 0^{N \times T}; \quad \mathbf{D} \in \mathbb{R}^{M \times N}; \quad \mathbf{D} > 0$
 repeat
    Approximate using sparse coding algorithm
       $\mathbf{X} \longleftarrow \arg\min_X \|\mathbf{S} - \mathbf{D}\mathbf{X}\|_F^2 \quad s.t. \, \|\mathbf{x}_t\|_0 = k$
    Update dictionary and coefficient matrix
       FOR $n \in \{1, ...., N\}$
          $\Gamma_n = \{t \mid [X]_{n,t}) > 0\}$
          $\mathbf{E}_n = [\mathbf{S} - \mathbf{D}\mathbf{X} + \mathbf{d}_n \mathbf{x}^n]_{\Gamma_n}$
          $\mathbf{E}_n = \mathbf{U}\Delta\mathbf{V}^T$
          IF $\sum_i \mathbf{v}_i < 0$
             $\mathbf{x}^n_{\Gamma_n} = -\mathbf{V}$
             $\mathbf{d}_n = -\mathbf{U}$
          ELSE
             $\mathbf{x}^n_{\Gamma_n} = \mathbf{V}$
             $\mathbf{d}_n = \mathbf{U}$
          ENDIF
       ENDFOR
 until stopping condition met
```

An alternative approach to the sparse NMF problem (9.9) is offered by Non-Negative K-SVD (NN-K-SVD) [3], outlined in Algorithm 9.1 where it is seen to iterate through two separate steps. The first step uses a non-negative sparse approximation algorithm to identify the sparse support. An algorithm referred to as Non-Negative Basis Pursuit (NN-BP), consisting of several iterations of the sparse multiplicative update (9.10) followed by a k-sparse thresholding for sparse approximation, is proposed in [3]. However, the authors note that any non-negative sparse approximation algorithm can be used. The second step uses the K-SVD [2] update. The standard K-SVD updates atoms in a sequential fashion, and also updates the corresponding coefficient vector while updating an atom. However, it is possible that the signs of the atoms and their corresponding coefficients may be reversed, and NN-K-SVD has to correct for this. The sign direction of the atom in which most energy is present is set as the positive side, while a corresponding reflection of the coefficient vector is performed if necessary. Remaining negative coefficients in the atom and the dictionary are then set to zero. It is noted [3] that the cost function (9.9) may actually increase during the update due to this setting of non-negative elements to zero, in which case it is proposed to use subsequent multiplicative updates in order to reduce the cost function. While this fix may lead to consideration that K-SVD may not be the most appropriate dictionary update step in a non-negative framework, it may conversely be queried if this fix is absolutely necessary

Algorithm 9.2 NMF-ℓ_0

```
Input
  $\mathbf{S} \in \mathbb{R}^{M \times T}, \quad k, \quad J$
Initialise
  $\mathbf{X}^0 = 0^{N \times T}; \quad \mathbf{D} \in \mathbb{R}^{M \times N}; \quad \mathbf{D} > 0; \quad ||\mathbf{d}_n||_2 = 1 \ \forall n$
repeat
  Normalise columns of D  $s.t. \quad ||\mathbf{d}_n||_2 = 1 \, \forall n$
  Approximate using sparse coding algorithm
    $\mathbf{X} \longleftarrow \arg\min_X ||\mathbf{S} - \mathbf{D}\mathbf{X}||_F^2 \quad s.t. \quad ||\mathbf{x}_t||_0 = k$
  FOR $j = 1 : J$
    Update D using (9.8)
    Update X using (9.7)
  ENDFOR
  Normalise
    $\mathbf{x}^n \leftarrow \mathbf{x}^n / ||\mathbf{d}_n||_2 \ \forall n$
    $||\mathbf{d}_n||_2 = 1 \ \forall n$
until stopping condition met
```

as monotonicity, while often considered desirable, may not be necessary when solving NMF [61] [62]. A stated advantage of the (NN-) K-SVD algorithm is that it may be used with overcomplete dictionaries [2] [3].

The NMF-ℓ_0 algorithm [101], also seeks to approximate a ℓ_0 norm penalty in the sparse NMF problem (9.9) through use of a sparse approximation algorithm. Once the sparse support is identified, several iterations of the Euclidean NMF multiplicative updates (9.7) (9.8) are performed, thereby updating both the coefficient matrix and the dictionary in a similar manner to NN-K-SVD. For sparse approximation the authors propose a gradient variant of NN-OMP [16] and also compare the use of NN-BP, described earlier in the context of NN-K-SVD. Experimental results in [101] describe NMF-ℓ_0 outperforming both NMF and NN-K-SVD in a range of experiments.

In the rest of this chapter two pieces of research are presented that were performed with a large separation in time. The following section describes some early experiments performed in the context of recovery of harmonic dictionaries that formed the first publications submitted during the course of this research project [92] [93]. It had previously been shown [8] that NMF significantly outperformed NN-K-SVD for the purpose of AMT. This seemed a curious result and some toy experiments with harmonic dictionaries and coefficient matrices of known sparsity are presented, which serve to highlight the usefulness of sparsity in NMF and the importance of using an apt sparse approximation algorithm. This led to research, as described in previous chapters of this thesis, exploring sparse approximation in the presence of the non-negative constraint.

A variant of Sparse NMF, referred to as ℓ_0-Sparse NMF, is proposed that uses BF-NNLS algorithm to update the coefficient matrix and NNLS to update the dictionary. As the proposed approach is couched in the ANLS framework for NMF, it is also capable of learning overcomplete dictionaries. Experiments are presented, showing improved recovery of overcomplete non-negative dictionaries relative to NMF-ANLS and some other Sparse NMF algorithms.

9.1 Structure-Aware Dictionary Learning

NMF is a very popular tool in the musical signal processing community and is used for other applications as well as AMT. As previously mentioned, NMF and NN-K-SVD were compared for the purpose of AMT in [8], and NMF was seen to perform better for this purpose. A large discrepancy between the results reported for the two algorithms is reported in [8]. This may seem strange, as NN-K-SVD is a specific, sparse, approach to the NMF problem. Indeed, it could be hypothesised that incorporation of the sparse constraint, seen in the case of NN-K-SVD, should not be detrimental to performance in the AMT problem, with its underlying sparse assumption.

This motivates further investigation of the aforementioned algorithms. In particular it is queried whether the performance gap may be an effect of the K-SVD update, or if errors are more likely to be a result of the sparse approximation algorithm. In particular, Matching Pursuit (MP) , constrained to select atoms with non-negative inner products with the residual signal, was used as the sparse coding algorithm for NN-K-SVD in [8]. A particular facet of MP is that a backprojection step is not taken. This may lead to erroneous atom coefficients, a problem that may be amplified in this coherent setting, and a similarly incorrect residual, which may effect the K-SVD step.

In this context, substitution of other sparse approximation algorithms affords simple isolation of the effect of using MP for NN-K-SVD. Similarly it is desired to isolate the effect of the K-SVD update. Hence, a further algorithm, referred to as Sparse Multiplicative Update Dictionary Learning (SMUDL), outlined in Algorithm 9.3, is proposed. SMUDL performs alternating projections, using a sparse coding algorithm to update the coefficient matrix, then updating the dictionary with one iteration of the multiplicative dictionary update (9.7). The difference in this approach relative to NMF-ℓ_0 is small; NMF-ℓ_0 uses a sparse approximation algorithm followed by several iterations of both multiplicative updates, thereby updating both $\mathbf{X}$ and $\mathbf{D}$ given a fixed sparse support, in a similar manner to NN-K-SVD. Indeed, it is possible that NMF-ℓ_0 may possibly outperform SMUDL in this context. However, the intention here is solely to isolate the

Algorithm 9.3 SMUDL

```
Input
  $\mathbf{S} \in \mathbb{R}^{M \times T}, \quad k$
Initialise
  $\mathbf{X}^0 = 0^{N \times T}; \quad \mathbf{D} \in \mathbb{R}^{M \times N}; \quad \mathbf{D} > 0$
repeat
  Normalise dictionary : $\quad \|\mathbf{d}_n\|_2 = 1 \, \forall n$
  Approximate using sparse coding algorithm
    $\mathbf{X} \longleftarrow \arg\min_X \|\mathbf{S} - \mathbf{DX}\|_F^2 \quad s.t. \quad \|\mathbf{x}_t\|_0 = k$
  Update $\mathbf{D}$ using (9.8)
until stopping condition met
```

different elements of the NN-K-SVD algorithm, in order to test their efficacy, or otherwise.

One problem with using NMF for AMT is that certain signal elements may not be separable. To give a simple example, it is easy to consider two notes that are always played coincidentally in a piece to be factorised, a scenario that may not be that unusual due to the chordal nature of western music. It is to be expected that a factorisation algorithm would learn one atom containing both sources, in the absence of prior knowledge. Indeed, this problem of learning dual source atoms may be extended to other less specific situations.

In the case of many musical pieces a harmonic structure is expected in signal elements, and harmonically constrained variants of NMF have been proposed in order to leverage this structure for the AMT problem. The earliest of these [106] proposes to use a dictionary initialised with harmonic structure. In particular, the dictionary was initialised with 88 atoms, each representing a different note of the piano scale. Each atom was initialised with a non-zero element at each frequency frame that represented a multiple of the fundamental frequency of the note it sought to represent, while all other frequency frames were set to zero. A feature of multiplicative updates is that elements that are equal to zero are unaffected. A problem with this method [106] noted by the authors is that the atoms representing notes that do not exist in the signal interact with the atoms of active notes. Hence, the authors proposs a penalisation strategy for correlated atoms.

An alternative formulation of Harmonic NMF (H-NMF) is proposed in [133], in which the adaptive harmonic dictionaries are proposed. In this approach each full spectrum pitched harmonic atom is formed from a variable superposition of several narrowband harmonic atoms labelled with the same pitch, and a semi-supervised NMF algorithm is used to learn the optimal superpositions to constitute each full spectrum atom. The results given for this approach are considered state-of-the-art for AMT using NMF. An alternative approach, proposed in [126], sought to extract harmonic atoms sequentially from a non-negative matrix. Experiments using this approach were described using synthetic harmonic dictionaries, with results showing improved dictionary recovery relative to unconstrained NN-K-SVD and NMF.

It is easy to consider that such harmonically constrained approaches may be limited when presented with a signal consisting of a mix of harmonic and inharmonic features. Nonetheless it seems worthwhile to experiment with harmonically constrained variants of the different NMF methods mentioned above. A harmonically constrained SMUDL (H-SMUDL) is easily effected using the same initialisation as H-NMF [106] as the dictionary update is similarly based on the multiplicative update, and zeros are maintained. A harmonic version of NN-K-SVD is also proposed (H-NN-KSVD). Unlike the multiplicative update, the NN-K-SVD update does not necessarily preserve structure, and requires filtering at each atom update

$$\mathbf{d}_n \longleftarrow \mathbf{d}_n \otimes \mathbf{v}_n \tag{9.13}$$

where $\mathbf{V}$ is a binary matrix, of similar dimensions to $\mathbf{D}$, that encodes the harmonic structure by placing ones at the harmonic partials and zeroes elsewhere, in a similar fashion to the H-NMF initialisation.

Experiments

Several questions have been raised by reference to the literature just described. Firstly can the use of different sparse coding algorithms effect the efficacy of sparse NMF methods such as NN-K-SVD when used in the context of dictionaries with harmonic structure? If so, how does this then compare with NMF? Finally, if the sparse structure of the dictionary is known beforehand, how much might this improve the recovery rate? Attempting to answer these questions suggests an experimental setup that is quite pointed in its intentions. Hence, toy dictionary recovery experiments were run with synthetic noiseless spectrograms. Another consideration in real-world factorisation problems is the selection of the learning order, or the rank of the factorisation. The effect of the learning order on NMF and NN-K-SVD in the context of AMT is noted in [8]. In order to constrain the number of variables in the experimental setup the learning algorithms are performed with the learning order set to the number of known atoms in the dictionary. Similarly, the sparse algorithms are also aware of the fixed number of atoms active at each frame.

The experiments are based upon those described in [126], with a harmonic dictionary consisting of ten atoms synthesised. From this dictionary a spectrogram is synthesised and the dictionary is initially constructed so that the spectrogram simulates a STFT with sampling frequency $44.1kHz$ and a signal window of dimension 4096. Two variations in the experiment are used, which differ only in how the harmonic dictionary $\mathbf{D} \in \mathbb{R}^{2048 \times 10}$ is created. In both cases a unique fundamental frequency from the set

$$f_0 \in \{200, 300, 400, 500, 600, 800, 900, 1000, 1200, 1500\}$$

is assigned to each atom, thereby producing a large harmonic overlap between different atoms. At each multiple of the fundamental frequency a harmonic peak is assigned. In the first set of experiments, referred to as *Experiment A*, the coefficients of each harmonic peak are assigned using a fixed spectral envelope

$$\phi_f = \mathbf{e}^{\left(\frac{-f}{512}\right)^2}$$

similar to that used in [126]. Sidelobes for each harmonic peak are assigned by filtering with a Gaussian window. For the second set of experiments, *Experiment B*, the same harmonic structure is used; however the coefficients of each peak and sidelobe is selected randomly from a equiprobable distribution in the interval [0, 1]. In both sets of experiments, all dictionary atoms are normalised to have unit ℓ_2 norm. A non-negative coefficient matrix, $\mathbf{X} \in \mathbb{R}^{10 \times 100}$, is also synthesised, with each column consisting of 5 randomly selected non-zero elements, with coefficients sampled from a equiprobable distribution in the interval [0.02, 1]. The synthetic spectrogram is formed from the product of the dictionary and coefficient matrix.

For both sets of experiments, 100 different spectrograms are synthesised. In the case of *Experiment B*, a unique dictionary is used for each individual experiment. Experiments are run using the NN-K-SVD, SMUDL and NMF algorithms, and their harmonic variants. For the sparse NMF algorithms, NN-K-SVD and SMUDL, three different sparse approximation algorithms are used; NN-MP as used in [8], NN-OMP [16] and T-NNLS [114]. It is noted that T-NNLS displays a similar performance as NN-BP, in which the multiplicative update step (9.10) is observed to tend towards a NNLS solution. NN-MP and NN-OMP stop iterating when $k = 5$ atoms are selected at each frame of the spectrogram, while T-NNLS uses the k-sparse thresholding strategy (§3.4), also with $k = 5$. The same spectrograms and initial dictionaries are used for all algorithms, in order to effect a fair comparison. Each algorithm is run for 100 iterations, after which the learnt dictionary is compared with the original dictionary.

To measure the similarity between the two matrices, two metrics are used. First, similar to

[126], a *hit* is registered when an atom in the learnt dictionary has a correlation of 0.9 or greater with an atom from the original dictionary, and the number of hits is recorded. An *Accuracy* measure is given as the percentage of atoms for which a hit is recorded:

$$Acc = \frac{number\, of\, hits}{number\, of\, atoms} \times 100\%.$$

The average maximum correlation between the dictionaries is also measured using the matrix $\bar{\mathbf{G}} = \bar{\mathbf{D}}^T \mathbf{D}$, where $\mathbf{D}$ is the original dictionary and $\bar{\mathbf{D}}$ is the learnt dictionary. Vectors of the row maximums and column maximums of $\bar{\mathbf{G}}$ are formed, and the mean of both of these vectors is taken. The average correlation, ρ is given as the minimum of these two averaged values:

$$\rho_{mean} = \frac{\min\{\sum_{k=1}^{K} \max \bar{\mathbf{g}}_k, \sum_{k=1}^{K} \max \bar{\mathbf{g}}^k\}}{K}. \tag{9.14}$$

The minimum of the column/row maximums is used to avoid extra hits where a learnt atom may contain energy of two signal elements and hence be highly correlated with two synthetic atoms.

Results

The results for the experiments are given in Table 9.1, from which several observations are made. First, it is seen that NMF performs better than NN-K-SVD (NN-MP) in both experiments, a result that resonates with those in [8]. However, this pattern is reversed when the other sparse approximation algorithms, NN-OMP and T-NNLS are used, in which the results are superior to those for NMF by a considerable margin.

In *Experiment A*, NMF is seen to recover 61.9% of atoms, at the correlation threshold specified, while NN-K-SVD recovers 95.4% and 98.8% of the atoms using NN-OMP and T-NNLS, respectively. In comparision, the SMUDL algorithm has a recovery rate of 84.2% with NN-OMP, and 90.6% with T-NNLS. In terms of sparse approximation algorithms the T-NNLS is seen to improve on the results given using the NN-OMP for both NN-K-SVD and SMUDL. The results using the SMUDL algorithm relative to NMF show that the sparse approximation step effects a large improvement in the dictionary recovery performance. However, the K-SVD update is seen to outperform the multiplicative update in terms of updating the dictionary. This is in contrast to the original set of experiments, described in [92] [93], in which SMUDL was seen to outperform NN-K-SVD. In these experiments [92] [93], code for NN-K-SVD published by the authors [3] was used, with minor modifications performed to incorporate the harmonic structure. It was

	Experiment A		*Experiment B*	
Algorithm	$Acc(\%)$	$\mathcal{C}_{mean}$	$Acc(\%)$	ρ_{mean}
NMF	61.9	0.91	89.7	0.96
NN-KSVD (NN-MP)	39.9	0.86	39.1	0.87
NN-KSVD (NN-OMP)	95.4	0.98	99.8	**1.00**
NN-KSVD (T-NNLS)	**98.8**	**1.00**	**100**	**1.00**
SMUDL (NN-MP)	25.6	0.84	26.2	0.83
SMUDL (NN-OMP)	84.2	0.96	98.9	**1.00**
SMUDL (NNLS)	90.6	0.97	99.9	**1.00**
H-NMF	98.8	0.99	**100**	**1.00**
H-NN-KSVD (NN-MP)	42.0	0.86	46.2	0.88
H-NN-KSVD (NN-OMP)	97.0	0.99	**100**	**1.00**
H-NN-KSVD (T-NNLS)	99.5	**1.00**	**100**	**1.00**
H-SMUDL (NN-MP)	65.9	0.92	62.2	0.91
H-SMUDL (NN-OMP)	99.3	**1.00**	**100**	**1.00**
H-SMUDL (NNLS)	**99.8**	**1.00**	**100**	**1.00**

Table 9.1: Results from experiments comparing NMF with NN-K-SVD using different sparse coding algorithms

noted in subsequent observations that atoms were occasionally omitted from the representation completely at the sparse approximation step. In the code provided, the replacement strategy for unsupported atoms was to form a new atom based on the signal residual. In the context of the experiments presented here, this was seen not to be effective as the newly formed atom tended not to be included at further sparse approximation steps. An alternative strategy, retaining unsupported atoms, was used and led to a large difference in the results. It is worth recalling that the experimental setup used here is similar to that used in [126], and the results seen here for NN-K-SVD are superior to those reported in [126].

Experiment B is seen to be easier for all methods, with both sparse algorithms achieving almost perfect recovery, while NMF is seen to recover 89.7% of atoms. It is recalled that the only difference between *Experiment A* and *Experiment B* is in assignment of dictionary coefficients, with a fixed spectral envelope used in the former, and random coefficients in the latter. An analysis in terms of a global dictionary coherence measure is offered. The correlation of each element in the dictionary is given as an element of the Gram matrix, $\mathbf{G} = \mathbf{D}^T\mathbf{D}$. A global correlation measure is given by

$$\mu_G = ||\mathbf{G} - \mathbf{I}||_F^2 \tag{9.15}$$

when the dictionary elements are normalised. A fixed dictionary is used in *Experiment A*, for which it is found that $\mu_G = 3.36$. In *Experiment B*, a different dictionary is used for each separate experiment. To compare, each individual dictionary used in *Experiment B* is measured using (9.15). It is found the mean of μ_G for these dictionaries is 2.79, with a standard deviation of 0.06, while the maximum and minimum values of μ_G found are equal to 2.94 and 2.63, respectively. In other words the fixed spectral envelope dictionary is much more correlated, or coherent. While sparsity is often considered advantageous in NMF [55], it may become important, or even essential, when correlated underlying factors exist.

The harmonically constrained versions of the dictionary learning algorithms are given in the bottom half of Table 9.1. The recovery rates are high for all algorithms, except for the sparse algorithms when NN-MP is employed as the sparse approximation algorithm. Indeed, for *Experiment B* perfect recovery is seen for all algorithms when NN-MP is not used. In *Experiment A*, some slightly different patterns to those for the unconstrained algorithms are seen.

H-SMUDL algorithm outperforms H-NN-K-SVD relative to the sparse approximation algorithm used, while H-NMF outperforms H-NN-K-SVD, except when T-NNLS is used as the sparse approximation algorithm. This could suggest that the harmonic structure favours the multiplicative update to the K-SVD update; however, the differences are relatively insignificant. The unrealistic experimental setup is noted, with the learning algorithm aware of both the correct number of atoms and their structure. In terms of AMT it is recalled that H-NMF [106] required a penalty term to counter the interactions of atoms representing notes not present in the signal. Nonetheless, a similar variant of K-SVD, referred to as Musical Structure K-SVD (MS-K-SVD) [48], which is very similar to H-NN-K-SVD has recently been proposed for the purpose of AMT, with promising results reported.

Some Further Experiments

Some results in the previous set of experiments demand some further exploration. In particular, poor performance was observed for NMF in *Experiment A*. It is worth investigating if this is just an effect of slower convergence when sparsity is not enforced, or whether sparsity may actually be required in order to find an appropriate factorisation. While NN-K-SVD outperformed other methods, the K-SVD update is relatively computationally expensive. Equivalent performance may be achievable using SMUDL, or other algorithms, with less computation required.

In light of this, *Experiment A* from the previous section is repeated using unconstrained NMF and sparse NMF algorithms. Multiplicative update NMF is run for 1000 iterations while S-MUDL is run for 500 iterations. Two other NMF approaches are also compared. NMF using

Algorithm	*#iterations*	*Acc*	ρ_{mean}
NN-KSVD (NN-OMP)	100	95.4	0.98
NN-KSVD (T-NNLS)	100	98.8	**1.00**
SMUDL (NN-OMP)	100	84.2	0.96
SMUDL (NN-OMP)	500	93.2	0.98
SMUDL (T-NNLS)	100	90.6	0.97
SMUDL (T-NNLS)	500	98.3	0.99
NMF	100	61.9	0.91
NMF	500	71.7	0.94
NMF	1000	73.3	0.95
ANLS-NMF	100	74.9	0.95
ANLS-NMF	500	76.3	0.95
ANLS-NMF (T-NNLS)	100	**99.2**	**1.00**
β-NMF	100	62.0	0.91
β-NMF	500	64.6	0.92

Table 9.2: Further results for Experiment A

ANLS [97] is run for 500 iterations. Further to this the β-NMF with $\beta = 0.5$, is run for 500 iterations. A sparse variant of ANLS-NMF is also proposed, with thresholding applied to the coefficient matrix. This is referred to as ANLS-NMF(T-NNLS) and is run for 100 iterations.

The results for the further experiments are shown in Table 9.2, where an increase in performance can be seen for both SMUDL and NMF after further iterations, with the performance of SMUDL coming close to that of NN-K-SVD. However, NMF is seen to perform relatively poorly. After 500 iterations, an improvement of almost 10% is seen relative to the performance after 100 iterations. However, a further 500 iterations yield only an extra 1.6% of atoms recovered, at which point the recovery rate is still more than 20% lower than for NN-K-SVD after 100 iterations. Considering the slow improvement observed between 500 and 1000 iterations, it is apparent that the multiplicative update version of NMF may have converged at this stage. Further unreported experiments with a larger amount of iterations yielded no significant improvement. The β-NMF is seen to perform similar to the Euclidean NMF, suggesting that the alternating multiplicative update approach is not suitable for unsupervised NMF in such scenarios. ANLS-NMF is also seen to fail to recover the dictionaries, with performance only slightly better than that of the Euclidean NMF. On the other hand, experiments run using the sparse variant of ANLS-NMF, with T-NNLS used to update the coefficient matrix are seen to slightly improve upon the performance of the NN-K-SVD, giving the best results of all. From these results, it is obvious that introduction of sparsity to the NMF problem may be essential in some cases.

9.2 ℓ_0-Sparse NMF

Experimental results in the previous section outline the importance of sparsity in NMF when underlying correlated signal elements are present in the matrix to be factorised. In the toy experiments presented, all algorithms that did not employ a sparse approximation step were seen to fail to recover the dictionaries. Meanwhile, all algorithms employing a sparse approximation step, other than MP, were relatively successful. The employment of NN-OMP was seen to improve upon standard NMF approaches, even though NN-OMP can be considered to be somewhat ill-suited in such a coherent problem. Further improvements were observed when T-NNLS was used instead of NN-OMP, motivating a further examination of sparse approximation for NMF.

It is worth recalling the NN-K-SVD [3] and NMF-ℓ_0 [101] NMF approaches that employ sparse approximation algorithms. More recent research with these two approaches has considered different sparse approximation steps. NN-OMP is used with NN-K-SVD in [48] for the purpose of AMT. Further research on the NMF-ℓ_0 approach was recently published [100], in which the use of different non-negative sparse approximation algorithms is considered. A backwards elimination algorithm called reserve sparse NNLS (rsNNLS) was proposed in [100], and was seen to outperform NN-OMP and non-negative ℓ_1-minimisation in sparse approximation tasks. rsNNLS, similar to BF-NNLS (§5.1) is a backwards elimination algorithm; however, some differences between the two approaches should be noted. Both approaches start from an initial NNLS solution, performing a downdate or elimination at each step. While BF-NNLS takes an optimal step at each iteration, rsNNLS eliminates the atom with the smallest coefficient in the downdated NNLS coefficient vector. rsNNLS is proposed for a k-sparse experimental setup, while BF-NNLS more easily accommodates a threshold on the energy loss in the step, and the optimality in the step suggests that a better ℓ_0 approximation should be formed.

A variant of sparse NMF, referred to as ℓ_0-Sparse NMF, (ℓ_0S-NMF) using the BF-NNLS sparse approximation step with the modified sparse cost function (§5.2) is proposed. ℓ_0S-NMF, outlined in Algorithm 9.2 employs the standard ANLS NMF [97] methodolgy, and differs only through incorporation of the backwards elimination step after the NNLS coefficient matrix estimation step. The ℓ_0S-NMF can also be employed with other dictionary update steps, such as the NN-K-SVD or NMF-ℓ_0 approaches. Alternatively stated, the BF-NNLS algorithm can be used as the sparse approximation step for the NN-K-SVD and NMF-ℓ_0 algorithms. However, using the ANLS framework for ℓ_0S-NMF, the backwards elimination approach can also be used to enforce

Algorithm 9.4 ℓ_0 S-NMF Algorithm

```
Input   $\mathbf{S} \in \mathbb{R}^{M \times T}$,  $N$,  $\lambda$
Initialise   $\mathbf{D} \in \mathbb{R}^{M \times N}$
repeat
  Perform BF-NNLS
  $\mathbf{X} = \arg\min_{\mathbf{X}} \|\mathbf{S} - \mathbf{D}\mathbf{X}\|_2^2 \quad s.t. \quad \mathbf{X} \geq 0$
  for t = 1:T do
    $\Gamma_t = \{n | [X]_{n,t} > 0\}$
    repeat
      $\Delta r_t = \mathbf{x}^{[2]} \oslash diag([\mathbf{D}_{\Gamma_t}^T \mathbf{D}_{\Gamma_t}]^{-1})$
      $\hat{n} = \arg\min \Delta^n \mathbf{r}_t$
      $\bar{\Delta}^{\hat{n}} \mathbf{r}_t = \sqrt{\|\mathbf{r}_t\|_2^2 + \Delta^{\hat{n}} \mathbf{r}_t} - \|\mathbf{r}_t\|_2$
      $\Gamma_t \leftarrow \Gamma_t \backslash \hat{n}_t$
    until $\bar{\Delta}^{\hat{n}} \mathbf{r}_t < \lambda$
    $\mathbf{x}_t = \arg\min_x \|\mathbf{s}_t - \mathbf{D}_{\Gamma_t} \mathbf{x}\|_2^2 \quad s.t. \quad \mathbf{x} \geq 0$
  end for
  Update Dictionary
  $\mathbf{D} = \arg\min_D \|\mathbf{S}^T - \mathbf{X}^T \mathbf{D}^T\|_2^2 \quad s.t. \quad \mathbf{D} \geq 0$
until stopping condition
```

sparsity on the dictionary as well, thereby considering a cost function

$$\mathcal{C}_S = \sum_n \{\|\mathbf{s} - \mathbf{D}\mathbf{x}_n\|_2 + \lambda \|\mathbf{x}_n\|_0\} + \eta \|\mathbf{D}\|_0 \tag{9.16}$$

where $\|\mathbf{D}\|_0$ is the number of non-zero elements in the dictionary and η is a parameter enforcing sparsity on the dictionary. In the transposed NNLS problem (9.5) the solution given is $\mathbf{D}^T$, the transpose of the dictionary. Application of BF-NNLS in this case sets to zero some elements in a given row of the dictionary.

Experiments

Some synthetic dictionary recovery experiments were designed to test the proposed ℓ_0-SNMF approach. Random, twice overcomplete non-negative dictionaries $\bar{\mathbf{D}}$ of dimension 200×400 were generated, with each element sampled from a flat equal probability distribution in the range [0,1], and all dictionary columns were normalised to unit ℓ_2 norm. A coefficient matrix $\bar{\mathbf{X}}$ of dimension 200×800 was synthesised using a equal distribution in [0.02, 1]. Between 5 and 10 entries of $\bar{\mathbf{X}}$ were randomly selected to be active in each column for all experiments, and all other entries of $\bar{\mathbf{X}}$ were set to zero. Experiments were performed using three different sparsity levels in the dictionary, with $\{10, 25, 50\}\%$ of entries set as non-zero. The matrix $\mathbf{S} = \bar{\mathbf{D}}\bar{\mathbf{X}}$ was synthesised. Subsequent factorisation was performed using different NMF approaches, each run for

50 iterations of alternating projections. All approaches use the transposed ANLS approach (9.5) to perform the dictionary update, while different algorithms are used to estimate the coefficient matrix $\mathbf{X}$ at each iteration.

The proposed ℓ_0S-NMF is used with $\lambda = 0.02$, the minimum value of an activation in the synthesised dictionary. NMF was performed using the ANLS approach. OMP was used as a sparse approximation step, with non-negative constraints applied. OMP stopped iterating when either 15 atoms were selected, or the relative error $\frac{\|\mathbf{r}_n\|_2^2}{\|\mathbf{s}_n\|_2^2} < 0.05$. Thresholded NNLS (T-NNLS) was performed using two different values of the threshold $\lambda = 0.02$ and $\lambda = \sqrt{0.02}$. An ℓ_1-SNMF approach was also performed, with $\lambda = 0.02$, and also with $\lambda = 0.04$ (ℓ_1-SNMF (2λ)).

Early efforts attempted to use accelerated active set methods [65] [64] that attempt to solve multiple right-hand sides of the NNLS problem simultaneously. However these were seen to be problematic, inducing scaling errors. It is suspected that this is an effect of the overcomplete dictionaries used. For all NNLS calcuations the active set Fast-NNLS [14] method was used, with each column of $\mathbf{S}$, or $\mathbf{S}^T$ in the transposed case, being decomposed independently.

Similar to the experiments in the previous section, the goal is to find a dictionary that is similar to the original dictionary, using the described NMF techniques. In order to measure the similarity between the original and estimated dictionaries, the measure ρ (9.14) , which measures the average maximum correlation of an atom from the original dictionary with a learnt atom, is again used. A value of $\rho = 0.95$ is considered success in this set of experiments, and an additional measure $\mathcal{I}$, relates the number of iterations taken to achieve $\rho = 0.95$, when averaged across all experiments.

Results

The results for the experiments are shown in Table 9.3, while Figure 9.1 plots the average value of ρ across all experiments at each iteration, for the three different dictionary sparsity levels. It is observed that NMF performs poorly, being unsuccessful for all dictionary sparsity levels with the average correlation falling relative to the initialised dictionary in all cases. A stated advantage of the ANLS approach to NMF is that overcomplete dictionaries can be used [26]. However, from these experiments it would appear that a sparse approach may actually be necessary when the dictionary is overcomplete. With a relatively low threshold, $\lambda = 0.02$, the T-NNLS approach was seen to perform almost exactly the same as ANLS-NMF, and the results are not recorded. When a higher value of $\lambda = \sqrt{0.02}$ was used, the T-NNLS approach was seen to perform well for

	50%		25%		10%	
	ρ_{max}	$\mathcal{I}$	ρ_{max}	$\mathcal{I}$	ρ_{max}	$\mathcal{I}$
ℓ_0-S-NMF	**0.992**	27	**0.992**	**12**	0.973	**10**
NMF	0.408	-	0.507	-	0.660	-
T-NNLS	0.990	**17**	0.970	14	0.897	-
ℓ_1-SNMF	0.432	-	0.555	-	0.865	-
ℓ_1-SNMF(2λ)	0.476	-	0.893	-	**0.995**	21
OMP	0.818	-	0.958	15	0.976	11

Table 9.3: Different NMF algorithms compared for different dictionary density levels with dictionaries synthesised from equiprobable distribution

the densest dictionary, with 50% of active dictionary elements. In this case the T-NNLS approach learnt quicker than all other approaches, reaching $\rho = 0.95$ in the smallest amount of iterations. However, the performance of this approach is seen to deteriorate when the dictionaries become sparser, and is unsuccessful in recovering the dictionary of 10% density.

The proposed ℓ_0-SNMF approach is seen to be the only algorithm successful in all experiments. However, a small drop-off in performance is observed in the case of the sparsest dictionary, where ρ_{max} is reached after around 15 iterations, and not subsequently improved. A sparsity constraint on the dictionary, such as in (9.16), may improve this performance. NN-OMP is also shown to perform well, with success in the case of the sparser dictionaries, and a relatively high value of ρ_{max} in the case of the densest dictionary.

Using an ℓ_1-SNMF approach was relatively unsuccessful, using the considered parameters. Success was demonstrated only when the higher threshold (2λ) was used for the experiments using the sparsest dictionaries. In this case, however, ρ_{max} was seen to be higher than for all other algorithms. While learning was slower than with other algorithms, the correlation for the ℓ_1-SNMF(2λ) was seen to continue increasing when other successful approaches had ceased to improve.

The effect of the use of a higher threshold is obvious in the results, as seen in the case of T-NNLS. Some initial experiments were run with the ℓ_1^2-penalty norm suggested by [65], using a value of $\lambda = 0.02$ for which good performance was observed. However, this is an effect of the scaling that is inherent in the ℓ_1^2-norm, and performance using this approach was seen to be poor in similar experiments that were scaled down. Other initial experiments performed using high values of λ with the ℓ_1-SNMF approach were seen to bring similar improvements in the dictionary recovery to the ℓ_1^2 approach. These observations would seem to validate the approach taken

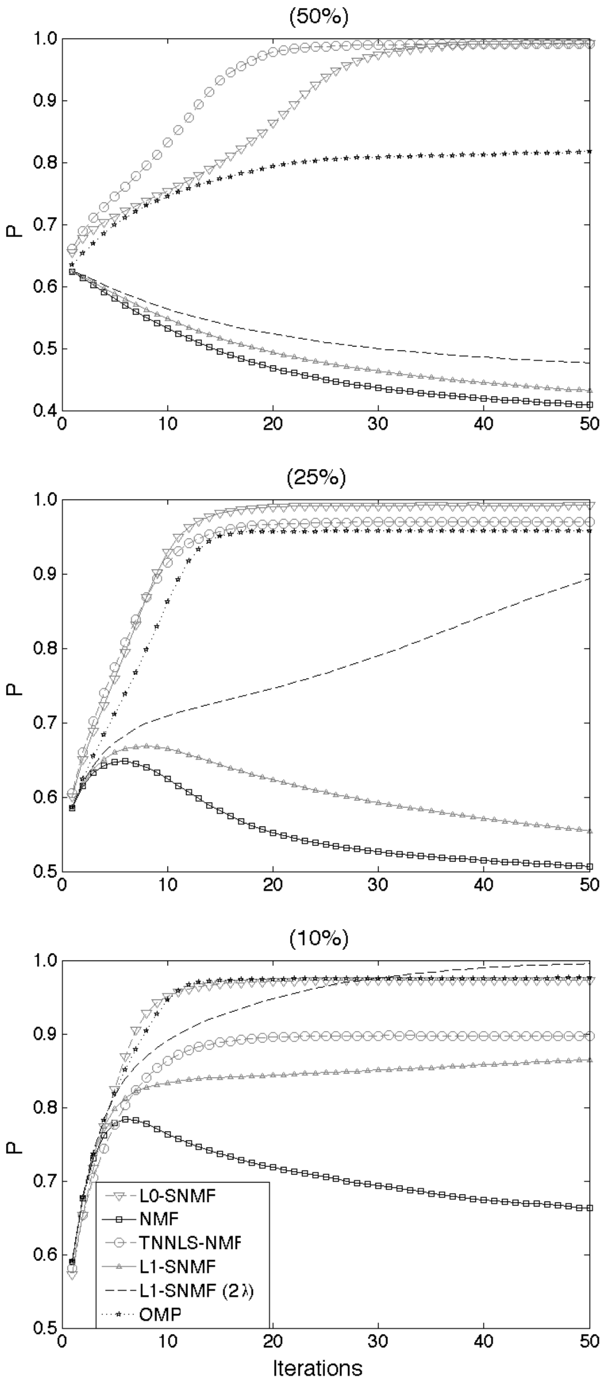

Figure 9.1: Comparison of NMF algorithms learning in terms of $\mathcal{P}$ when the dictionary is 50% dense (top) and 25% dense (middle) and 10% dense (bottom).

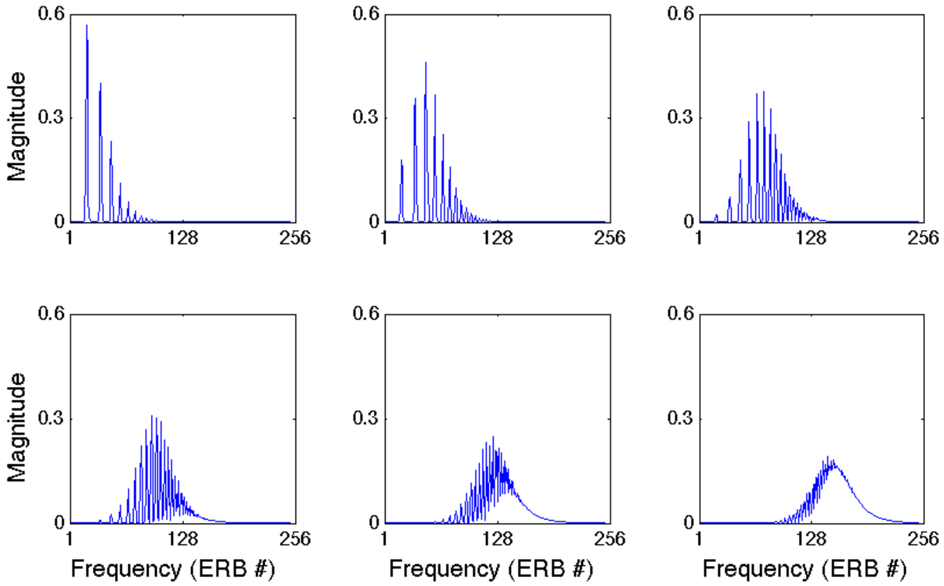

Figure 9.2: Group of atoms used to represent one note in adaptive harmonic dictionary

in [107] where a large value of λ, used at initial iterations, was gradually decreased. However, it is noted that in all cases the proposed ℓ_0 approach performs similarly, without requiring any scaling.

9.3 Is NMF necessary for AMT?

Spectrogram decompositions are considered to provide better results for the application of AMT than unsupervised learning methods such as NMF, particularly if the dictionary used represents the sources in the signal well. However, such a dictionary is not alway available. For this reason, NMF is popular tool for AMT and musical signal processing, affording a fast, data-driven approach. However, the problems with NMF in the context of AMT, such as learning order, separability of coincident notes and a tendency to learn meaningless atoms, are well noted in the literature [8] [133] [106].

An alternative NMF-based approach is the semi-supervised NMF of Vincent et al, proposed in [133]. In this method a hierarchical dictionary is used. At the top of the hierarchy is a dictionary consisting of pitched atoms, with one atom for each note on the piano scale, 88 in all. Each of these pitched atoms is actually a superposition of a group of narrowband harmonic atoms, each with the same fundamental frequency. An example of one of these groups is shown in Figure 9.2. Each of these narrowband atoms is fixed. In [133] a semi-supervised β-NMF (SSβ-

NMF) approach is used to learn harmonic atoms. Rather than learning a dictionary, SSβ-NMF seeks to learn the coefficients of a group of narrowband atoms, such as those in Figure 9.2, that constitute a good broadband harmonic atom. The learning is performed by using an alternating projections methodology, using β-NMF. In one projection the spectrogram is decomposed using the 88 broadband pitched atoms. In the alternative projection, the coefficients of the narrowband atoms are optimised. Each group of narrowband atoms comprising a given broadband atom are optimised with the coefficients of the broadband atom fixed. It is noteworthy that each alternating projection does not consist of one multiplicative update. In this approach, each alternating projection iterates until convergence of the given cost function.

Results in [133] place the performance for this method at 10% ahead of unsupervised NMF, in terms of $\mathcal{F}$-measure, and 7% ahead of the H-NMF [106]. Indeed, this method is generally considered state-of-the-art for NMF-based transcription systems. It is noted that, similar to the NMD experiments described in the same paper and from which the benchmark experiment described in (§3.5) is derived, that the authors of [133] performed the supervised NMF experiments using β-divergence NMF for a wide range of values of β.

It is possible to take an alternative view of the adaptive harmonic dictionary used in [133], and consider it as a subspace dictionary, and hence a candidate for the group sparse decompositions used throughout this thesis. While the learnt dictionaries were seen to perform well, they were recorded in a similar environment to the piano pieces that they were used to decompose. However, the dictionary used in [133] was not designed with decomposition methods in mind. For instance, the narrowband structure employed results in some atoms being present in areas of the spectrogram where little energy is present, and it is possible that a group sparse penalty might be unevenly distributed between the narrowband atoms constituting a pitch.

Experiments

Some experiments were run to compare the performance of group sparse NMD methods, from Chapter 7, against the SSβ-NMF approach of Vincent et al [133]. The software to construct these dictionaries and perform the related SSβ-NMF learning is available online. Dictionaries were created using the default settings in the software, and the SSβ-NMF was performed using the code supplied.

To perform group sparse NMD, all narrowband atoms were normalised to unit ℓ_2 norm and a group structure, $\mathcal{L}$ (§2.2.2.) was formed by applying common indexing of atoms of the same

	$\mathcal{F}$	δ_{opt}(dB)	$\mathcal{P}$	$\mathcal{R}$
G-KL-NMD (§7.2.1)	66.6	29	69.0	64.4
G-β-NMD (§7.2.1)	67.5	29	**71.9**	63.5
SSβ-NMF [133]	**67.7**	29	70.3	**65.3**

Table 9.4: Experimental results comparing semi-supervised β-NMF with G-NMD.

pitch (2.18). It is noted that the groups or subspaces created have varying size P, with $P = 6$ being the maximum, and used for low pitches, while $P = 3$ is the minimum, used only for the highest pitches. A total of 496 atoms were created. The spectrograms of Transform $\mathbb{E}1$, the default setting for the SSβ-NMF approach, were decomposed using Group-NMD (§7.2.1) decompositions using both the Kullback-Leibler (KL) and β-divergence, with $\beta = 0.5$, cost functions with a group sparse penalty. For the KL-divergence the $\ell_{\perp,1}$ mixed norm was used as the group sparse penalty, while the β-divergence was penalised with a $\ell_{\perp,0.5}^{0.5}$ norm. In both cases λ was set to 1.

The results are shown in Table 9.4. Here it is seen that there is little to separate the approaches. While SSβ-NMF method performs best, the improvement in $\mathcal{F}$-measure seen is 1.1% relative to the KL-divergence and only 0.2% relative to the G-β-NMD.. This is an interesting result, questioning the requirement of NMF methods for the purpose of decomposing such structured signals. It is possible that with a different dictionary design, the decomposition-based method may have performed as well. Alternatively, it may be possible to improve on the results of SSβ-NMF by enhancing the dictionary learning process, as the decomposition step is now superior to that used before.

9.4 Discussion

In this chapter, prior research on NMF and Sparse NMF was first described. A noted difficulty in NMF is the problem of the separability of overlapping factors [115], as described in the original paper proposing the use of NMF for AMT. In particular, the authors of [115] consider that each note in a signal to be factorised may have to be played in isolation at least once to be recognised. While this was later shown not to be the case [8], problems with the quality of atoms learnt using NMF are often noted [106] [133]. A curious result relating to the comparison of NMF and NN-K-SVD in [8] for the purpose of AMT was recalled. Some experiments were run to compare the use of some NMF and Sparse NMF algorithms, with different sparse approximation algorithms employed also. The experiments were synthesised from a small dictionary, designed to emulate musical spectrograms with a presence of harmonic overlap. It was found that several

NMF algorithms failed to recover the underlying dictionaries. However, all approaches that used a sparse approximation method, other than MP, were relatively successful. The implication is that sparse NMF methods may be desirable when factorising musical spectrograms.

A considerable difference was noted in the results for the expeirments above when T-NNLS and NN-OMP were used as sparse approximation algorithms. Previously it was observed that BF-NNLS, proposed in Chapter 5, is a good non-negative sparse approximation algorithm. Some further synthetic experiments, this time synthesised from overcomplete dictionaries were run to compare various NMF and Sparse NMF algorithms, all within the ANLS framework. The proposed ℓ_0-S-NMF was seen to perform most consistently amongst the approaches compared. The worst performance for ℓ_0-S-NMF was observed when the underlying dictionary was sparse. However, a strategy to enforce sparsity on the dictionary, again using BF-NNLS was proposed, albeit not implemented here.

Finally, some group sparse decompositions were performed using a generic harmonic dictionary, which was designed for use with a state-of-the-art NMF approach to AMT. The results were perhaps surprising. The group sparse constraint alone performed almost as well as the semi-supervised β-NMF approach. This opens a wide range of possibilities. For instance, group sparse NMF may be worth exploring, particularly in such a harmonically constrained case. An alternative perspective may also be taken. Structured sparse representations, with flexible, generic dictionaries may be sufficient for AMT, in which case dictionary design may become a pertinent area for future research.

Chapter 10

Conclusions

Throughout this thesis, a stated goal has been to inform the Automatic Music Transcription (AMT) problem through incorporation of concepts and methodology from the sparse representations repetoire. In particular the concepts of structured sparsity and dictionary coherence, and use of stepwise methods were introduced to AMT. To conclude this thesis, a summary of the research contained herein is given, with the main contributions then outlined. Finally some pointers to future research made more accessible by the findings of this thesis are offered.

10.1 Summary

In ***Chapter 4*** an exploration of greedy methods for AMT was undertaken. Some prior work in the literature was used as a reference point, particularly the use of OMP with large overcomplete dictionaries containing datapoint atoms. The alternative approach of subspace modelling was considered, and the use of group sparsity was deemed necessary when using dictionaries formed from a union of pitched subspaces. Magnitude spectrograms were used, as is typical in AMT spectrogram decompositions, leading to the proposal of novel non-negative variants of group OMP algorithms, many simply derived through suppression of non-negative inner products in the calculation of group selection coefficients. However, the best performing of the group non-negative pursuits, NN-NS-OMP, was not so simply adapted to the non-negative framework. Hence, a novel accelerated approach, F-NS-OMP, was derived through bounding the coefficients of the NNLS projection, bringing the computational load associated with this approach close to that of the other non-negative greedy group methods. The use of a gradient-based approximate backprojection step was also explored in order to lessen the computational load of backprojection incurred with the subspace dictionaries. However, limitations in terms of fractured temporal continuity, inability to correct early bad selections and the difficulty in choosing a good stopping condition are notable when using OMP for AMT. Nonetheless, some positive conclusions and motivation for further research were provided in this exploration of OMP. In particular it was found that subspace modelling is a apt approach, leading to development of other group sparse algorithms. The large variation in performance relative to the type of spectrogram transform used was noted, leading to a later analysis in terms of dictionary coherence.

Chapter 5 began with a demonstration, through a simple three-atom example, of how OMP can easily select an incorrect atom, in the context of musical signals. This provided the motivation for an exploration of stepwise methods that include a facility for backtracking. More specifically, the use of stepwise optimal methods was considered, prompted by the difficulty in selecting an good stopping condition when OMP is employed. This approach had not previously been considered in the context of AMT. Some methods from the sparse methodology incorporating backtracking were first described, with note taken of their tricky assimilation to a non-negative framework. Closer consideration of the problem led to the proposal of a backwards elimination approach, BF-NNLS, initialised from a non-negative least squares decomposition. In k-sparse experiments this was seen to be perform similarly to other approaches, while simpler in its execution. A modified sparse cost function was then proposed, due to the observation that the elimination cost for an atom was related to the square of its current NNLS coefficient. In this case improved results relative to NNLS were observed. A group sparse variant of the BF-NNLS approach was then proposed, using a modified group sparse cost function. This group sparse backwards elimination approach led to AMT results that were superior to the benchmark experiments, previously considered state-of-the-art for framewise decomposition-based AMT.

The research presented in ***Chapter 6*** was also motivated by the failings of OMP-based approaches; this time, with regard to the fractured time continuity observed in correctly detected signal elements. In this chapter, molecular sparse methods were considered. Molecular methods are most often considered in time-frequency representations, with tracking performed across time frames of pitch-similar atoms. However, this approach is not necessarily apt when the dictionary is coherent, as in the case of non-negative decomposition for AMT. A new approach was proposed to counter these problems. First, clustering of pitch-similar time-adjacent active atoms, in a NNLS-based piano roll, into molecules is performed. These molecules form an alternative type of dictionary, for which the new M-NS-OMP algorithm was proposed to perform molecular spectrogram decompositions. Promising results were observed using the proposed M-NS-OMP, both in frame- and event-based analyses. Comparison with a ground truth, or oracle, decomposition was proposed in order to analyse the decomposition-based AMT problem, and led to some interesting observations. Systematic problems in the onset detector employed were noted. While it was observed that very low thresholds are required to achieve a high value of Recall, the use of subspace modelling was validated as a higher potential Recall was recorded using an oracle decomposition. Finally a molecular norm was proposed, prompting new molecular algorithms such as Molecular Hard Thresholding (MHT) and Molecular Backwards Elimination. Experimental results indicate that the molecular approach is potentially robust, with performance observed to be equivalent regardless of the transform used, while the use of subspace modelling and backwards elimination were again validated.

Chapter 7 sees a departure from the use of stepwise methods, which result in subset NNLS decompositions. Previous AMT research has reported better results when cost functions other than the Euclidean distance are used, with the generalised β-divergence particularly popular in musical signal processing. Some exploration of the more recent generalised $\alpha\beta$-divergence, previously not considered in the context of AMT, was undertaken. A novel η divergence, itself a special case of the $\alpha\beta$-divergence, was proposed, leading to state-of-the-art AMT decompositions for the STFT. Monotonic descent is proved for the η-divergence, using an NMF update with a larger stepsize than previously considered in the NMF literature, for $\eta \in [0.5, 1]$ the range of enhanced performance. Sparse and group sparse penalised NMD approaches were then explored.

An alternative perspective is taken in ***Chapter 8***. Throughout many chapters of this thesis, varying performance was observable in AMT experiments relative to the transforms employed in the decomposition. This variation led to an analysis of the different dictionaries in terms of coherence, a perspective not previously undertaken in AMT. A relationship between the dictionary coherence in a given transform and AMT performance was shown. Considering then that coherence is somewhat related to AMT performance, a novel row-weighting conditioning approach for application in a non-negative framework was proposed. A new effective coherence measure, incorporating coherence and activity through the coefficients of an initial decomposition, was proposed. Projected gradient descent was then performed to find a row-weighting that reduced this effective coherence measure. This row weighting was then applied to both the signal and the dictionary before a further decomposition was performed, leading to improved AMT, while the observation was made that row weighting in a noiseless scenario has no effect on the final decomposition coefficients.

Fiinally, ***Chapter 9*** explores the field of Sparse Non-negative Matrix Factorisation (SNMF), with a focus on methods that consider a ℓ_0 sparse approximation algorithm. Some toy experiments showed that sparsity may necessary in some cases for NMF when underlying correlated factors exist in the matrix to be factorised. Further to this, a variant of sparse NMF, ℓ_0-SNMF was proposed, further indicating the usefulness of BF-NNLS as a non-negative sparse approximation algorithm. The chapter finishes with an observation of some experimental results, comparing group sparse decompositions using a generic harmonic dictionary with a state-of-the-art NMF-based method, bringing into question the necessity of NMF for AMT.

10.1.1 Main Contributions

- Coherence-based conditioning for AMT using newly proposed effective coherence measure.
- The η-divergence.
- BF-NNLS backwards elimination approach with modified sparse cost function; seen to be particularly useful for group sparse decompositions.
- ℓ_0-SNMF, a novel variant of Sparse NMF employing the BF-NNLS approach.
- Subspace modelling of musical notes and group sparse algorithms.
- Molecular sparse methods for AMT.
- F-NS-OMP; a fast non-negative variant of SMP.

10.2 Future Research

Several contributions to decomposition-based AMT have been outlined briefly in the previous section. While some of these, such as the molecular approach, can be considered application specific, many are general methods that may possibly find application elsewhere. However, possible avenues for further research in the context of AMT are suggested, some of which are outlined below.

10.2.1 Modelling Approaches

In the work presented subspace modelling of musical notes was seen to be beneficial, affording improved AMT relative to typical atomic pitch dictionaries in supervised decompositions. Further improvements may be possible by using a more considerate subspace learning approach. The individual subspaces were learnt in a simple fashion, using unconstrained Euclidean NMF, and the representation power of individual atoms was poor, with many harmonic partials missing from each atom. Some simple possibilities that might be considered include normalisation of the data, and the use of different cost functions to learn the atoms. More structured approaches are also possible, for instance using approaches that encourage co-activity in harmonic partials, or otherwise. The employment of temporally constrained learning could possibly lead to pitched-subspaces that themselves contain overlapping groups, which may better model the evolution of a note.

In a similar fashion to the subspace modelling, the molecular clustering approach was primitive with all time-adjacent active atoms clustered, occasionally leading to large molecules containing several notes, or molecules that substantially overrun the natural length of a single note. Ideally, a molecule should be more semantically meaningful. For instance in AMT each molecule could represent a single note, an approach that may lead to better thresholds or stopping parameters for stepwise methods. However, this might be a tricky problem, as it could be hypothesised that some type of onset detection would be necessary in order to derive such meaningful molecules. Further work could build on the observed deficiencies in the threshold-based onset detector used in Chapter 6.

While improved decompositions were implemented as part of this thesis, the post-processing of a decomposition was not developed, with adherence to simple global thresholding in order to derive a piano roll. It may be worthwhile considering more complex classification at each time-pitch point. For instance, in the AMT literature the use of Hidden Markov Models is seen to be popular. Other approaches to be considered include a simple mixed global and local thresholding.

10.2.2 Methodology

The backwards elimination methodology was seen to be useful when the modified sparse cost functions were introduced, affording good transcription with the use of only one stopping condition parameter, in the frame-based approach, and also performing well when a molecular approach was used. OMP-based approaches have previously been used in a multi-instrument setting [76], where gradient-based methods are seen to denote co-activity [133]. In this context, backwards elimination approaches may be able to provide the best elements of both approaches, affording the selectivity of greedy approaches with the better overall performance of other approaches. Simultaneous stepwise decompositions, in which different spectrograms are decomposed at once, in a similar manner to simultaneous sparsity [130] [54], may also be considered

In the exploration of the $\alpha\beta$-divergence it was seen that model-weighted cost functions performed well in the context of AMT, and consideration should be given to other divergences in the statistics literature that share this property, for which multiplicative update descent algorithms could be derived. An exploration of other approaches than the popular multiplicative updates might also be welcome. For instance, the use of proximal methods has been proposed for the Kullback-Leibler divergence [36], while coordinate descent algorithms have previously been proposed for the α- and β-divergences. Research should be performed into extension of these approaches, with a particular focus on group sparse penalisation.

An inherent weakness of decomposition based AMT is the requirement of a dictionary that represents the sources in the signal well. While sparse variants of NMF were proposed, experimentation with musical signals was not performed. One possibility is the use of group sparse NMF, which may be implemented by employing the GBF-NNLS algorithm in the ℓ_0 Sparse NMF. However, the final experiments in Chapter 9 suggest that spectrograms decomposition methods may be capable of similar performance to factorisation based approaches when a well-structured dictionary is used. Conversely, the performance of the decomposition methods in this case might indicate the potential for enhanced NMF based methods.

An alternative approach may be to use adaptive decompositions, whereby the signal or the dictionary or both are transformed. An example of this, using coherence-based row-weighting, was proposed in this thesis, and there are many future possibilities for such approaches. Initial work would consider incorporation with group sparsity. The genetic methods proposed in [109] provide an alternative example of an adaptive type of approach, whereby spectral templates are used and adapted. However, a more numerical based approach may be possible. While a decomposition adaptive approach was taken using the effective coherence measure in this thesis, consideration of more explicit adaptation to the signal may be worthwhile.

Bibliography

[1] S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by non-negative sparse coding of power spectra. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 318–325, Barcelona, 2004.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006.

[3] M. Aharon, M. Elad, and A. M. Bruckstein. K-SVD and its non-negative variant for dictionary design. In *Proceedings of the SPIE conference (Wavelets XI)*, pages 327–339, Baltimore, 2005.

[4] R. Badeau, N. Bertin, and E. Vincent. Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(12):1869–1881, December 2010.

[5] E. Benetos and S. Dixon. A shift-invariant latent variable model for automatic music transcription. *Computer Music Journal*, 36(4):81–94, Winter 2012.

[6] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchoff, and Anssi Klapuri. Automatic music transcription: Breaking the glass ceiling. In *Proceedings of the 13th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 379–384, Porto, 2012.

[7] D. Bernstein. *Matrix Mathematics*. Princeton University Press, 2005.

[8] N. Bertin, R. Badeau, and G. Richard. Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, pages 65–68, Honolulu, 2007.

[9] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio Speech and Language Processing*, 18(3):538–549, March 2010.

[10] T. Blumensath and M. Davies. On the difference between orthogonal matching pursuit and orthogonal least squares. Technical report, Department of Engineering, University of Edinburgh, 2007.

[11] T. Blumensath and M. E. Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, June 2008.

[12] S. Bock and M. Schedl. Polyphonic piano note transcription with recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–124, Kyoto, 2012.

[13] A. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.

[14] R. Bro and S. De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, 11(5):393–401, September 1997.

[15] G. J. Brown and D. L. Wang. Separation of speech by computational auditory scene analysis. In *Speech Enhancement*, pages 371–402. Springer, 2005.

[16] A. M. Bruckstein, M. Elad, and M. Zibulevsky. On the uniqueness of non-negative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54(11):4813–4820, November 2008.

[17] J. J. Burred. The Acoustics of the Piano, revised edition, translated by D. Ripplinger. Professional Conservatory of Music, Arturo Soria, Madrid, 2009.

[18] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

[19] E. J. Candes, M. B. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1-minimisation. *Journal of Fourier Analysis and Applications*, 14(5):877–905, December 2008.

[20] J. J. Carabias-Orti, P. Vera-Candeas, F. J. Canadas-Quesada, and N. Ruiz-Reyes. Music scene adaptive harmonic dictionary for unsupervised note-event detection. *IEEE Transactions Audio Speech and Language Processing*, 18(3):473–486, March 2010.

[21] C. Chafe and D. Jaffe. Source separation and note identification in polyphonic music. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1289 – 1292, Tokyo, 1986.

[22] D. Chen and R. J. Plemmons. Nonnegativity constraints in numerical analysis. In *A. Bultheel and R. Cools (Eds.), Symposium on the Birth of Numerical Analysis*, pages 109–140. World Scientific Press, 2009.

[23] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, December 1998.

[24] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He. Extended smart algorithms for non-negative matrix factorization. *Lecture notes in Artificial Intelligence, 8th International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, 4029:548–562, 2006.

[25] A. Cichocki, S. Cruces, and S. Amari. Generalized alpha-beta divergences and their application to robust non-negative matrix factorization. *Entropy*, 13(1):134–170, January 2011.

[26] A. Cichocki and A. H. Phan. Fast local algorithms for large scale non-negative matrix and tensor factorizations. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E92-A(3):708–721, March 2009.

[27] A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 32–39, Florida, 2006.

[28] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Elsevier, 2010.

[29] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, May 2009.

[30] L. Daudet. Sparse and structured decompositions of signals with the molecular matching pursuit. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1808–1816, September 2006.

[31] M. Davy, S. J. Godsill, and J. Idier. Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498–2517, April 2006.

[32] A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 489–494, Utrecht, 2010.

[33] S. Dixon. On the computer recognition of solo piano music. In *Proceedings of the Australasian Computer Music Conference*, pages 31–37, Brisbane, 2000.

[34] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, November 2001.

[35] D. L. Donoho, Y. Tsaig, I. Drori, and J. L. Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. Technical report, Department of Statistics, Stanford University, 2006.

[36] M. El Gheche, J.-C. Pesquet, and J. Farah. A proximal approach for optimization problems involving kullback divergences. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5984–5988, Vancouver, May 2013.

[37] Y. C. Eldar, P. Kuppinger, and H. Bolsckei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, June 2010.

[38] E. Elhamifar and R. Vidal. Block sparse recovery via convex optimisation. *IEEE Transactions on Signal Processing*, 60(8):4094–4107, August 2012.

[39] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio Speech and Language Processing*, 18(6):1643–1654, August 2010.

[40] C. Fevotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[41] C. Fevotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, September 2011.

[42] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, December 2007.

[43] M. A. T. Figueirido, J. M. Bioucas-Dias, and R. D. Nowak. Majorization-minimization algorithms for wavelet-based restoration. *IEEE Transactions on Image Processing*, 16(12):2980 – 2991, December 2007.

[44] N. H. Fletcher and T. D. Rossing. *The physics of musical instruments, 2nd edition*. Springer, 1998.

[45] N. Fonseca and A. Ferreira. Measuring music transcription results based on a hybrid decay/sustain evaluation. In *Proceedings of the 7th Triennial Conference of the European Society for the Cognitive Science of Music (ESCOM '09)*, Jyvaskyla, 2010.

[46] V. Franc, V. Hlavc, and M. Navara. Sequential coordinate-wise algorithm for the non-negative least squares problem. In A. Gagalowicz and W. Philips, editors, *Computer Analysis of Images and Patterns*, volume 3691 of *Lecture Notes in Computer Science*, pages 407–414. Springer Berlin Heidelberg, 2005.

[47] A. Ganesh, Z. Zhou, and Y. Ma. Separation of a subspace sparse signal: Algorithms and conditions. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3141–3144, Taipei, 2009.

[48] M. Genussov and I. Cohen. Multiple fundamental frequency estimation based on sparse representations in a structured dictionary. *Digital Signal Processing*, 23(1):390–400, January 2013.

[49] D. Godsmark and G. J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366, 1999.

[50] G. H. Golub and C. F. van Loan. *Matrix Computations*. North Oxford Academic, 1983.

[51] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, pages 287–288, Baltimore, 2002.

[52] A. Gretsistas. *Sparse Representations and Compressed Sensing for Direction of Arrival Estimation*. PhD thesis, Queen Mary University of London, 2013.

[53] R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51(1):101–111, January 2003.

[54] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! *Journal of Fourier Analysis and Applications*, 14(5):655–687, December 2008.

[55] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, November 2004.

[56] P.O. Hoyer. Non-negative sparse coding. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pages 557–565, Martigny, 2002.

[57] C. J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1064–1072, San Diego, 2011.

[58] A. Huang, G. Guan, Q. Wan, and A. Mehbodniya. A block orthogonal matching pursuit algorithm based on sensing dictionary. *International Journal of the Physical Sciences*, 65(5):992–999, March 2011.

[59] A. Huang, G. Guan, Q. Wan, and A. Mehbodniya. A re-weighted algorithm for designing data dependent sensing dictionary. *International Journal of the Physical Sciences*, 6(3):386–390, February 2011.

[60] H. Huang and A. Makur. Backtracking-based matching pursuit method for sparse signal reconstruction. *IEEE Signal Processing Letters*, 18(7):391–394, July 2011.

[61] D. Kim, S. Sra, and I. S. Dhillon. Fast newton-type methods for the least squares nonnegative matrix approximation problem. In *Proceedings of the 6th SIAM International Conference on Data Mining (SDM)*, pages 343–354, Minneapolis, 2007.

[62] D. Kim, S. Sra, and I. S. Dhillon. Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Statistical Analysis and Data Mining*, 1(1):38–51, February 2008.

[63] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating nonnegativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[64] J. Kim and H. Park. Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, July 2008.

[65] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing (SISC)*, 33(6):3261–3281, November 2011.

[66] H. Kirchoff, S. Dixon, and A. Klapuri. Shift-variant non-negative matrix deconvolution for music transcription. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 125–128, Kyoto, 2012.

[67] A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, 20004.

[68] A. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Audio, Speech and Language Processing*, 11(6):804–816, November 2003.

[69] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):255–266, February 2008.

[70] M . Kowalski, K. Seidenberg, and M. Dorfler. Social Sparsity: Neighbourhood systems enrich structured shrinkage operators. *IEEE Transactions on Signal Processing*, 61(10):2498–2511, May 2013.

[71] M. Kowalski and B. Torresani. Sparsity and Persistence: Mixed norms provide simple signal models with dependent coefficients. *Signal Image and Video Processing*, 3(3):251–264, September 2009.

[72] S. Krstulovic, R. Gribonval, P. Leveau, and L. Daudet. A comparison of two extensions of the matching pursuit algorithm for the harmonic decomposition of sounds. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 259–262, New Paltz, 2005.

[73] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice Hall, 1974.

[74] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS 14)*, pages 556–562, Denver, 2000.

[75] A. Lefevre, F. Bach, and C. Fevotte. Itakura-Saito nonnegative matrix factorization With Group Sparsity. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–24, 2011.

[76] P. Leveau, E. Vincent, G. Richard, and L. Daudet. Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1):116–128, January 2008.

[77] B. Mailhe, D. Barchiesi, and M. D. Plumbley. INK-SVD: Learning incoherent dictionaries for sparse representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3573–3576, Kyoto, 2012.

[78] B. Mailhe, R. Gribonval, F. Bimbot, and P. Vandergheynst. A low complexity orthogonal matching pursuit for sparse signal approximation with shift-invariant dictionaries. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3445– 3448, Taipei, 2009.

[79] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way 3rd edition*. Elsevier, 2009.

[80] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(3):3397–3415, December 1993.

[81] I. Markovsky and S. Van Huffel. Overview of total least-squares methods. *Signal processing*, 87(10):2283–2302, October 2007.

[82] K. D. Martin. A blackboard system for automatic transcription of simple polyphonic music. Technical Report 385, Massachusetts Institute of Technology Media Laboratory Perceptual Computing, 1996.

[83] B. Moghaddam, A. Gruber, Y. Weiss, and S. Avidan. Sparse regression as a sparse eigenvalue problem. In *Information Theory and Applications Workshop (ITA)*, pages 219 –225, San Diego, 2008.

[84] J. A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, pages 32–38, November 1977.

[85] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative update algorithms for nonnegative matrix factorization with the β-divergence. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 283–288, Kittila, 2010.

[86] B. Natajaran. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, April 1995.

[87] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3), May 2009.

[88] K. O'Hanlon, N. Keriven, and M. D. Plumbley. Structured sparsity using backwards elimination for automatic music transcription. In *Proceedings of IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, 2013.

[89] K. O'Hanlon, H. Nagano, and M. D. Plumbley. Group non-negative basis pursuit for automatic music transcription. In *Proceedings of the Workshop on Music and Machine Learning (MML) at ICML 2012*, Edinburgh, 2012.

[90] K O'Hanlon, H. Nagano, and M. D. Plumbley. Oracle analysis for automatic music transcription. In *Proceedings of 9th International Symposium on Computer Music Modelling and Retreval (CMMR)*, London, 2012.

[91] K. O'Hanlon, H. Nagano, and M. D. Plumbley. Structured sparsity for automatic music transcription. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 441–444, Kyoto, 2012.

[92] K. O'Hanlon and M. D. Plumbley. Structure-aware dictionary learning with harmonic atoms. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1761–1765, Barcelona, 2011.

[93] K. O'Hanlon and M. D. Plumbley. Structure-aware non-negative dictionary learning. In *Proceedings of the 4th Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, page 68, Edinburgh, 2011.

[94] K. O'Hanlon and M. D. Plumbley. Non-negative group sparsity. In *Proceedings of the IMA Conference on Numerical Linear Algebra and Optimisation*, Birmingham, 2012.

[95] K. O'Hanlon and M. D. Plumbley. Row-weighted decompositions for automatic music transcription. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, 2013.

[96] K. O'Hanlon and M. D. Plumbley. Learning overcomplete dictionaries with ℓ_0 sparse non-negative matrix factorisation. In *Proceedings of the IEEE Global Conference on Signal and Information Processing*, to appear.

[97] P. Paatero and U. Tapper. Positive matrix factorisation: A non-negative factor model with optimal utilization of error. *Environmetrics*, 5:111–126, 1994.

[98] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 40–44, Pacific Grove, CA, 1993.

[99] P. H. Peeling, A. T. Cemgil, and S. J. Godsill. Generative spectrogram factorization models for polyphonic piano transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):519–527, March 2010.

[100] R. Peharz and F. Pernkopf. Sparse nonnegative matrix factorisation with ℓ_0 constraints. *Neurocomputing*, 80:38–46, March 2012.

[101] R. Peharz, M. Stark, and F. Pernkopf. Sparse nonnegative matrix factorisation using ℓ_0 constraints. In *IEEE Interational Workshop on Machine Learning for Signal Processing*, pages 83–88, Kittila, 2010.

[102] M. D. Plumbley. Recovery of sparse representations by polytope faces pursuit. In *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pages 206–213, Florida, 2006.

[103] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. Sparse representations in audio and music: From coding to source separation. *Proceedings of the IEEE*, 98(6):995–1005, June 2010.

[104] G. Poliner and D. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal of Advances in Signal Processing*, 8(1):154–162, January 2007.

[105] L. F. Portugal, J. Judice, and L. Vicente. Comparision of block pivoting and interior point algorithms for linear least squares problems with nonnegative variables. *Mathematics of Computation*, 63(208):625–643, October 1994.

[106] S. Raczynski, N. Ono, and S. Sagayama. Extending non-negative matrix factorisation - a discussion in the context of multiple frequency estimation of musical signals. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 934–938, Glasgow, 2009.

[107] J. Rapin, J. Bobin, A. Larue, and J. Starck. Robust non-negative matrix factorization for multispectral data with sparse prior. In *Proceedings of the 7th Conference on Astronomical Data Analysis (ADA 7)*, Cargese, 2012.

[108] L. Rebollo-Neira and D. Lowe. Optimized orthogonal matching pursuit approach. *IEEE Signal Processing Letters*, 9(4):137 –140, April 2002.

[109] G. Reis, F. Fernadez de Vega, and A. Ferreira. Automatic transcription of polyphonic piano music using genetic algorithms, adaptive spectral envelope modelling, and dynamic noise level estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8):2313–2328, October 2012.

[110] R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit. Technical report, Technion, Haifa, April 2008.

[111] M. P. Ryynanen and A. Klapuri. Polyphonic music transcription using note event modeling. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 319–322, New Paltz, 2005.

[112] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Transactions in Signal Processing*, 56(5):1994–2002, May 2008.

[113] C. Schoerkhuber and A. Klapuri. Constant-Q transform toolbox for music processing. In *Proceedings of the 7th Sound and Music Computing Conference (SMC)*, Barcelona, 2010.

[114] M. Slawski and M. Hein. Sparse recovery by thresholded non-negative least squares. In *Advances in Neural Information Processing Systems (NIPS 24)*, pages 1926–1934, Granada, 2011.

[115] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180, New Paltz, 2003.

[116] P. Smaragdis, B. Raj, and M. Shashanka. A probabilistic latent variable model for acoustic modeling. In *Workshop on Advances in Models for Acoustic Processing at NIPS*, Vancouver, 2006.

[117] C. Soussen, R. Gribonval, J. Idier, and C. Herzet. Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares. *IEEE Transactions on Information Theory*, 59(5):3158 – 3174, May 2013.

[118] P. Sprechmann, I. Ramirez, P. Cancela, and G. Sapiro. Collaborative sources identification in mixed signals via hierarchical sparse coding. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5816–5819, 2011.

[119] S. Sra and I. Dhillon. Non-negative matrix approximation : Algorithms and applications. Technical report, University of Texas at Austin, June 2006.

[120] B. L. Sturm and M. G. Christensen. Cyclic matching pursuits with multiscale time-frequency dictionaries. In *Conference Record of the 44th Asilomar Conference on Signals, Systems and Computers*, pages 581–585, Pacific Grove, CA, 2010.

[121] B. L. Sturm and M. G. Christensen. Comparison of orthogonal matching pursuit implementations. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 220 –224, Bucharest, 2012.

[122] B. L. Sturm, J. J. Shynk, and S. Gauglitz. Agglomerative clustering in sparse atomic decompositions of audio signals. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 97–100, Las Vegas, 2008.

[123] M. E. Davies T. Blumensath. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, November 2009.

[124] R. Tibrishani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, pages 267–288, 1994.

[125] S. K. Tjoa and K. J. Ray Liu. Factorization of overlapping harmonic sounds using approximate matching pursuit. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 257–262, Miami, 2011.

[126] S. K. Tjoa, M. C. Stamm, W. S. Lin, and K. J. Ray Liu. Harmonic variable-size dictionary learning for music source separation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 413–416, Dallas, 2010.

[127] B. E. Trevor, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2002.

[128] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions in Information Theory*, 50(10):2231–2242, October 2004.

[129] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86:589–602, April 2006.

[130] J. A. Tropp, A. C. Gilbert, and J. M. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit . *Signal Processing*, 86(3):572–588, April 2006.

[131] B. Varadarajan, S. Khudanpur, and T. D. Tran. Stepwise optimal subspace pursuit for improving sparse recovery. *IEEE Signal Processing Letters*, 18(1):27–30, January 2011.

[132] E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic non-negative matrix factorisation for polyphonic music transcription. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 109–112, Las Vegas, 2008.

[133] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):528–537, March 2010.

[134] T. Virtanen. Monaural sound source separation by non-negative matrix factorisation with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, March 2007.

[135] Z. Yang, H. Zhang, Z. Yuan, and E. Oja. Kullback-leibler divergence for nonnegative matrix factorization. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *Artificial Neural Networks and Machine Learning ICANN 2011*, volume 6791 of *Lecture Notes in Computer Science*, pages 250–257. Springer Berlin Heidelberg, 2011.

[136] C. Yeh, A. Roebel, and X. Rodet. Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *IEEE Transaction on Audio, Speech and Language Processing*, 18(6):1116 – 1126, August 2010.

[137] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1):49–67, February 2006.

[138] M. Zibulevsky and M. Elad. L1-L2 Optimisation in signal and image processing: Iterative shrinkage and beyond . *IEEE Signal Processing Magazine*, 27(3):76–88, May 2010.