Improving the perceptual quality of single-channel blind audio source separation

Tobias Stokes

Submitted for the Degree of Doctor of Philosophy
Declaration of originality

This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images, or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography, or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification. I agree that the University has the right to submit my work to the plagiarism detection service TurnitinUK for originality checks. Whether or not drafts have been so-assessed, the University reserves the right to require an electronic version of the final document (as submitted) for assessment as above.
Abstract

Given a mixture of audio sources, a blind audio source separation (BASS) tool is required to extract audio relating to one specific source whilst attenuating that related to all others. This thesis answers the question “How can the perceptual quality of BASS be improved for broadcasting applications?”

The most common source separation scenario, particularly in the field of broadcasting, is single channel, and this is particularly challenging as a limited set of cues are available. Broadcasting also requires that a source separator is automated, capable of handling non-stationary, reverberant mixtures and able to separate an unknown number of sources. In the single-channel case, the time-frequency mask is common as a method of separation. However, this process produces artefacts in the separated audio.

The perceptual evaluation for audio source separation (PEASS) toolkit represents an efficient way to generate a multi-dimensional measure of perceptual quality. Initial experimental work, using ideal target and interferer estimates, uses PEASS to test variations on the ideal binary mask and shows continuous masks are perceptually better than binary while identifying a trade-off between artefacts and interferer suppression.

To explore the optimisation of this trade-off, a series of sigmoidal functions are used to map target-to-mixture ratios to mask coefficients. This leads to a mask, with less target-to-mixture based discrimination than those typically found in literature, being identified as the optimum. Further experiments applying offsets, hysteresis, smoothing and frequency-dependency to the mask do not show any benefit in audio quality.

The optimal sigmoidal mask is demonstrated to also be superior under non-ideal conditions using a non-negative matrix factorisation algorithm to produce the estimates. A final listening test compares the outputs of binary, ratio and optimal sigmoidal masks concluding that listeners prefer the ratio mask to the sigmoidal mask and both continuous masks to the binary mask.
Acknowledgements

My foremost thanks belong with my supervisors: Tim Brookes and Chris Hummersone. They have both played a vital role throughout the completion of this project and the preparation of my thesis. I am immensely grateful for their time, wisdom and guidance.

The input of the BBC into this work has made it so much more than an academic research project. I’m particularly grateful to Andrew Mason for his supervision of the industrial side of this work. Early on, conversations with Chris Baume and Dave Marston played an important part in shaping the direction of this research. Further thanks belong with the many BBC sound supervisors and studio managers who either allowed me to observe their work or took time out of their busy schedules to discuss my work and listen to some of the audio produced.

Before commencing my PhD, I was warned that the existence of a PhD student can be incredibly lonely. However, I have been fortunate to be a part of an incredible cohort of PhD students during my time in the Institute of Sound Recording. My thanks to: Andy Pearce, Cleo Pike, Daisuke Koya, Jon Francombe, Khan Baykaner, Kirsten Hermes, Marek Olik, Phil Coleman and Tommy Ashby.

Finally, I’d like to thank my wife, Rhiannon, for her love, encouragement, patience and belief in me throughout this process.

The work in this thesis was supported by funding from the Engineering & Physical Sciences Research Council (EPSRC) and the British Broadcasting Corporation Research & Development Department (BBC R&D) by way of an Industrial Cooperative Award in Science & Technology (iCASE).
# Table of contents

List of figures vii  
List of tables ix  
List of equations x  
Publications arising from this thesis xiii  

Chapter 1: Introduction 1  
1.1 What is blind audio source separation? 2  
1.2 What is time-frequency masking? 2  
1.3 Applications of blind audio source separation 3  
1.4 Thesis aim 3  
1.5 Thesis structure 4  

Chapter 2: Blind audio source separation 7  
2.1 Classification of BASS problems 7  
2.1.1 Classification of the mixtures to be separated 8  
2.1.2 Classification of un-mixing tasks 12  
2.1.3 Summary 15  
2.2 Methods for BASS 16  
2.2.1 Computational auditory scene analysis 16  
2.2.2 Independent component analysis 31  
2.2.3 Non-negative matrix factorisation 36  
2.2.4 Discussion 39  
2.2.5 Summary 42  
2.3 Chapter summary 44  

Chapter 3: Evaluating audio source separation 45  
3.1 Objective evaluation 46  
3.1.1 Schobben et al.’s method 46  
3.1.2 The ideal binary mask ratio 47  

Chapter 4: Conclusions 49
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.3</td>
<td>Vincent et al.’s method</td>
<td>47</td>
</tr>
<tr>
<td>3.2</td>
<td>Subjective assessment of separation quality</td>
<td>51</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Stubbs &amp; Summerfield’s test</td>
<td>51</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Kornycky et al.’s test</td>
<td>51</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Emiya et al.’s test</td>
<td>52</td>
</tr>
<tr>
<td>3.3</td>
<td>Objective modelling of subjective metrics</td>
<td>52</td>
</tr>
<tr>
<td>3.3.1</td>
<td>PEMO-Q</td>
<td>53</td>
</tr>
<tr>
<td>3.3.2</td>
<td>The PEASS system</td>
<td>56</td>
</tr>
<tr>
<td>3.4</td>
<td>Discussion</td>
<td>58</td>
</tr>
<tr>
<td>3.5</td>
<td>Chapter summary</td>
<td>59</td>
</tr>
</tbody>
</table>

Chapter 4: The requirements of the broadcasting industry | 61
---
| 4.1     | The observations | 62   |
| 4.1.1   | Live broadcasting | 62   |
| 4.1.2   | Event recording | 63   |
| 4.1.3   | Programme editing | 65   |
| 4.2     | Noise control | 65   |
| 4.2.1   | Intellimix | 66   |
| 4.2.2   | Dialogue noise suppressor | 66   |
| 4.3     | Implications for an un-mixing tool | 67   |
| 4.3.1   | The interface with the engineer | 67   |
| 4.3.2   | Inputs and outputs | 67   |
| 4.4     | Opinions of the sound supervisors | 68   |
| 4.4.1   | Data compression | 68   |
| 4.4.2   | The art of mixing | 69   |
| 4.5     | Chapter summary | 69   |

Chapter 5: Developing source separation for the broadcasting industry | 71
---
| 5.1     | The TF mask for broadcasting | 72   |
| 5.1.1   | The case for the TF mask | 72   |
| 5.1.2   | The binary mask | 73   |
| 5.2     | The performance of binary masking | 74   |
# Table of contents

5.2.1 Representation ........................................ 76
5.2.2 Feature extraction .................................... 76
5.2.3 Masking ................................................. 77
5.3 How can the binary mask be improved? ......................... 78
  5.3.1 Alternatives to the binary switching function .......... 78
  5.3.2 Smoothing of the binary mask ........................ 80
5.4 Chapter summary ........................................ 81

Chapter 6: Perceptual quality improvement of time-frequency masks 82
6.1 Method ...................................................... 83
  6.1.1 Audio material ...................................... 83
  6.1.2 PEASS ............................................... 84
6.2 The experimental masks ...................................... 85
  6.2.1 TF representation ................................... 85
  6.2.2 The ideal binary mask ................................ 87
  6.2.3 The dithered binary mask ............................. 87
  6.2.4 The noisy binary mask ................................ 88
  6.2.5 The cepstrally-smoothed binary mask ................. 89
  6.2.6 The segmented binary mask .......................... 95
6.3 Comparison .................................................. 95
  6.3.1 Discussion ............................................ 95
6.4 Chapter summary ........................................ 97

Chapter 7: Ideal sigmoidal masking 99
7.1 Background .................................................. 100
7.2 Method ...................................................... 101
  7.2.1 Audio mixtures and overlap calculation ............... 102
  7.2.2 Estimates ............................................. 103
7.3 Results ...................................................... 104
  7.3.1 PEASS results ....................................... 105
  7.3.2 BSS Eval results ..................................... 107
  7.3.3 The relationship between TF overlap and optimal sigmoid 107
Table of contents

7.4 Resolution .................................................. 107
7.5 Chapter summary .............................................. 108

Chapter 8: Non-ideal sigmoidal masking .......................... 110
8.1 Gao & Woo’s algorithm ........................................ 111
8.2 Experimental procedure ....................................... 112
8.3 Results ....................................................... 113
  8.3.1 PEASS results .............................................. 113
  8.3.2 BSS Eval results .......................................... 115
8.4 Chapter summary ............................................... 115

Chapter 9: Optimised sigmoidal masking ...................... 116
9.1 Offset sigmoidal masking .................................... 117
  9.1.1 Method .................................................. 117
  9.1.2 Results .................................................. 118
  9.1.3 Discussion ............................................... 119
9.2 Hysteresis .................................................... 119
  9.2.1 Results .................................................. 121
  9.2.2 Discussion ............................................... 121
9.3 Smoothing .................................................... 121
  9.3.1 Method .................................................. 123
  9.3.2 Results .................................................. 123
  9.3.3 Discussion ............................................... 123
9.4 Frequency dependent masking .............................. 125
  9.4.1 Results .................................................. 125
  9.4.2 Discussion ............................................... 127
9.5 Chapter summary ............................................... 127

Chapter 10: Subjective assessment of separated audio quality .. 129
10.1 Method ..................................................... 130
  10.1.1 Assessors ................................................. 130
  10.1.2 Test environment and set up .......................... 130
  10.1.3 Stimuli .................................................. 133
10.2 Analysis .................................................... 134
# Table of contents

## 10.2.1 Post-screening of assessors

## 10.2.2 Data pre-processing

## 10.2.3 Initial analysis

## 10.2.4 Model creation

## 10.2.5 Post hoc tests

## 10.3 Paired comparison test

### 10.3.1 Method

### 10.3.2 Results

## 10.4 Discussion

## 10.5 Chapter summary

## Chapter 11: Conclusions and further work

### 11.1 Thesis summary

#### 11.1.1 Main aim

#### 11.1.2 Chapter 1

#### 11.1.3 Chapter 2

#### 11.1.4 Chapter 3

#### 11.1.5 Chapter 4

#### 11.1.6 Chapter 5

#### 11.1.7 Chapter 6

#### 11.1.8 Chapter 7

#### 11.1.9 Chapter 8

#### 11.1.10 Chapter 9

#### 11.1.11 Chapter 10

### 11.2 Contributions to knowledge

#### 11.2.1 The sigmoidal continuum between the flat and binary masks

#### 11.2.2 Sigmoidal optimisation of artefact-interferer trade-off

#### 11.2.3 Perceptual preference for ratio mask over a binary mask

#### 11.2.4 Negative findings

#### 11.2.5 Data to question PEASS

### 11.3 Further work

#### 11.3.1 Understanding the relationship between the ideal and non-ideal cases
### Table of contents

11.3.2 An improved perceptual model ........................................ 156
11.3.3 The optimal TF basis ...................................................... 157
11.3.4 Sample level mask smoothing ......................................... 157

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of acronyms</td>
<td>158</td>
</tr>
<tr>
<td>List of symbols</td>
<td>161</td>
</tr>
<tr>
<td>References</td>
<td>164</td>
</tr>
</tbody>
</table>
List of figures

1.1 Thesis question structure ............................................. 6
2.1 A mixture of two sound sources being recorded at one microphone . 10
2.2 A mixture of two sound sources being recorded at one microphone in a reverberant environment . .......................... 11
2.3 The spectrogram of a viola playing the note A3 .................. 18
2.4 The time domain waveform of the viola playing A3 .............. 19
2.5 The spectrum of a viola playing the note A3 .................... 20
2.6 The histogram produced by a pattern matching algorithm ..... 21
2.7 The ACF of the viola playing A3 .................................. 22
2.8 The SACF of a bank of 128 gammatone filters .................. 24
2.9 The absolute analytic signal of an amplitude modulated sine wave . 25
2.10 Gaussian smoothing for onset and offset detection in drum music . 26
2.11 Cross correlation to detect the ITD ............................... 28
2.12 Plots of the Laplacian and Gaussian distributions both with zero mean and unity standard deviation. ....................... 32
2.13 Plot of entropy approximations ................................. 35
2.14 The bases and coding matrices produced by an NMF algorithm . 37
2.15 Overview of an NMF algorithm that separates spatial information . 39
3.1 A block diagram of the PEMO-Q system .......................... 54
3.2 A block diagram of the PEASS system ............................. 57
5.1 Overview of a typical under-determined BASS system. A TF representation is calculated before the target and interferer are estimated. Based on these estimates, the mixture is masked allowing the target estimate to be resynthesised. ............................. 75
5.2 A box plot comparison of APS for the ideal and non-ideal masks .. 77
5.3 Mask value against TMR for binary, ratio and sigmoidal masks . 79
6.1 The experimental TF masks calculated for a segment of audio . . . 86
List of figures

6.2 Optimisation of the dithered binary mask . . . . . . . . . . . . . . . 88
6.3 Optimisation of the noisy binary mask . . . . . . . . . . . . . . . . 90
6.4 CBM optimisations . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
6.5 CBM optimisations . . . . . . . . . . . . . . . . . . . . . . . . . . . 93
6.6 CBM optimisations . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
6.7 Results of the initial experiment . . . . . . . . . . . . . . . . . . . . 96

7.1 Sigmoidal switching functions spaced on powers of two scale . . . . . . . 101
7.2 Perceptual results for ideal sigmoidal masking . . . . . . . . . . . . . 104
7.3 Physical results for ideal sigmoidal masking . . . . . . . . . . . . . . 105
7.4 Analysis of the relationship between peak OPS and TF overlap . . . . 106
7.5 OPS scores for sigmoids using different TF resolutions . . . . . . . . 109

8.1 PEASS results for sigmoidal masking using Gao & Woo’s algorithm 113
8.2 Physical metrics for non-ideal sigmoidal masking . . . . . . . . . . . 114

9.1 Offset sigmoidal curves . . . . . . . . . . . . . . . . . . . . . . . . . 118
9.2 The effect of adding hysteresis to a sigmoidal mask . . . . . . . . . . 122
9.3 The results of the smoothing experiment . . . . . . . . . . . . . . . . 124
9.4 OPS results calculated using frequency-dependent sigmoidal masking 126

10.1 The MUSHRA testing interface . . . . . . . . . . . . . . . . . . . . . 132
10.2 eGauge analysis of assessor reliability . . . . . . . . . . . . . . . . . 136
10.3 eGauge analysis of assessor discrimination . . . . . . . . . . . . . . 137
10.4 Box plots for each programme items . . . . . . . . . . . . . . . . . . 139
10.5 Histograms of normalised assessor responses . . . . . . . . . . . . . 141
10.6 Bootstrapped confidence intervals for ratings of each TF mask . . . 142
10.7 The UI from the AB test . . . . . . . . . . . . . . . . . . . . . . . . . 146
10.8 Results of the paired comparison test . . . . . . . . . . . . . . . . . 147
List of tables

2.1 A summary of the mixing matrix properties that classify the nature of a mixture. ................................................. 9
2.2 The various classes of BASS tasks and their end goals ............ 13
2.3 Summary of CASA’s extraction of spectral, spatial and temporal features. ............................................................. 29
2.4 A comparison of the processing of spectral, temporal, spatial and statistical data by the ICA, NMF and CASA approaches. ...... 41
2.5 The number of mixtures required to separate audio using different BASS methods. ................................................ 43
3.1 A comparison of BASS evaluation techniques. ......................... 60
5.1 Perceptual quality results for the TF mask from Araki et al. (2012) 75
9.1 The results of the offset sigmoidal experiment ....................... 119
10.1 The programme items to be used for the listening test, their measured overlap and the optimal OPS scores observed when separating each mixture in Chapter 7. ................................. 133
10.2 Mauchly’s test output table from the R console. The output shows significant violations of sphericity for the mask and programme factors. 143
10.3 The output of the analysis of variance (ANOVA) process showing significant effects for the mask, programme and the interaction between them. ....................................................... 143
10.4 Results of the Bonferroni post hoc tests conducted between each pairing of mask and programme item ............................ 144
10.5 Showing significant differences between the ratings given to audio from each of the masks. ........................................ 145
10.6 Rank order comparisons of the three masks using four metrics: interference-related perceptual score (IPS), APS,OPS and the listener data from this chapter. .............................................. 148
List of equations

1.1 The general form of an un-mixing problem ................. 2
2.1 Jutten & Hérault’s (1988) model of the un-mixing problem .... 8
2.2 Delayed mixing .................................................. 11
2.3 Convolutive mixing model ..................................... 11
2.4 The auto-correlation function ................................. 22
2.5 The first derivative of the Gaussian Function ............... 25
2.6 The cross-correlation for ITD detection ...................... 27
2.7 ITD at different frequencies. ................................. 28
2.8 Calculation of the IID .......................................... 28
2.9 Calculation of kurtosis using moments ...................... 32
2.10 Calculation of entropy from PDF ............................ 33
2.11 Negentropy in terms of entropy ............................. 33
2.12 Generalised negentropy approximation .................... 34
2.13 Hyvärinen’s first negentropy approximation ............... 34
2.14 Hyvärinen’s second negentropy approximation .......... 34
2.17 Maximum a posteriori expression of ICA in TF domain. ... 36
2.18 The first NMF update rule for the bases matrix $W$ ........ 38
2.19 The second NMF update rule for the bases matrix $W$ .... 38
2.20 The NMF update rule for the bases matrix $H$ ............. 38
3.1 Schobben et al.’s distortion calculation .................... 46
3.2 Schobben et al.’s separation quality calculation .......... 46
3.3 The ideal binary mask ratio (IBMR) ........................ 47
3.6 Vincent et al.’s decomposition of a separation into error terms ... 47
3.7 The general form of an orthogonal projector ................ 48
3.8 Vincent et al.’s calculation of the target source in an estimate ... 48
3.9 Vincent et al.’s calculation of interference in an estimate ... 48
3.10 Vincent et al.’s calculation of error due to noise in an estimate ... 48
3.11 Vincent et al.’s calculation of error due to artefacts in an estimate . 48
List of equations

3.12 Calculation of error due to interference ........................................... 49
3.13 Calculation of error due to noise ...................................................... 49
3.14 Vincent et al.’s simplified calculation of error due to noise in an estimate .............................................................. 49
3.15 Calculation of source to distortion ratio ................................................ 50
3.16 Calculation of source to interferer ratio .............................................. 50
3.17 Calculation of sources to noise ratio .................................................... 50
3.18 Calculation of sources to artefact ratio ................................................ 50
3.19 Calculation of the ISR .................................................................. 50
3.20 The PEMO-Q assimilation ................................................................. 55
3.21 The cross-correlation coefficient .......................................................... 55
3.22 PEMO-Q’s measure of the overall error ............................................... 56
3.23 PEMO-Q’s measure of the target error ............................................... 56
3.24 PEMO-Q’s measure of the interference error ....................................... 56
3.25 PEMO-Q’s measure of the artefacts error .......................................... 56
3.26 The non-linear processing of the perceptual saliences ........................... 56

5.1 The ideal binary mask .................................................................. 73

6.1 The ideal binary mask .................................................................. 87
6.2 The dithered binary mask .............................................................. 87
6.3 The noisy binary mask ................................................................. 89
6.4 The cepstral transform applied to the ideal binary mask (IBM) .......... 90
6.5 The cepstrally-smoothed binary mask .............................................. 90
6.6 The inverse cepstral transform of a TF mask .................................. 91

8.1 The conventional non-negative matrix factorisation model ............ 111
8.2 The 2D non-negative matrix factorisation model .............................. 111

9.1 The modified optimal sigmoidal mask for offsetting .................... 117
9.2 The constraints applied to the offset sigmoidal masking functions .... 117
9.3 The behaviour of a Preisach hysteron ............................................ 120
9.4 The Preisach model summation ..................................................... 120
9.5 The \( \alpha \) curve used for hysteresis .............................................. 120
9.5 The $\beta$ curve used for hysteresis ........................................ 120

10.1 The eGauge reliability calculation for the $j^{\text{th}}$ subject ........... 135
10.2 The eGauge discrimination calculation for the $j^{\text{th}}$ subject ......... 136
10.3 The multimodality test specified in ITU-R BS.2300-0 (2014) ........... 140
Publications arising from this thesis

Conference papers


Book chapter


Data Archive

The data underlying the findings presented in this thesis are available from http://iosr.uk/unmixing.
Chapter 1

Introduction

A key property of sound waves is that they are linearly superposable; two or more sounds can be united, either in the air or electronically, to become a sum of their parts (Howard & Angus 1996). This property has been exploited for centuries by composers scoring great chorales or symphonies combining the sounds of hundreds of sources. More recently, it has been used in the production of audio for electronic media, mixing together the signals from microphones on different parts of a band, or capturing ambient sound along with a presenter’s speech to give a location recording a sense of realism.

Not every combination of sounds is desirable and a recording may require elements to be removed. Currently, if a desired sound and an undesired sound noticeably coincide on single channel, that channel is no longer of use to a mixing engineer. While creating a sound mixture from individual sources is a straightforward process, extracting one or more sources for individual playback from a mixture is a much greater challenge, particularly when nothing is known about the original sources. This is because while a forward problem, in this case mixing, allows analysis of a system’s inputs and parameters to generate a specific output, unmixing involves using observations of a system’s output to predict a system’s input and parameters, and is an inverse problem with no unique solution (Tarantola 2005).

Academia has named this challenge blind audio source separation (BASS) (Vincent et al. 2006). Alternative names include: audio un-mixing, audio separation (Srinivasan & Kankanhalli 2003) and blind source separation (BSS) (Chan
et al. 1995). This thesis details research undertaken to improve BASS for the broadcasting industry. This opening chapter will introduce the themes of BASS and time-frequency (TF) masking and outline the research questions and structure of the thesis.

1.1 What is blind audio source separation?

BASS describes the problem of separating acoustic energy related to one sound source from a multi-source recording. The mixed sources can be assumed to overlap in frequency as well as time meaning that they can not be separated by a single filter. The problem is often formulated mathematically as

\[ x(t) = As(t) \]  \hspace{1cm} (1.1)

where \( x(t) \) represents the signal that is a mixture of the sources, \( s(t) \), mixed according to the process \( A \) (Jutten & Hérault 1988). The aim is to uncover the inverse of \( A \) allowing the sources to be separated. In real world applications \( A \) can not be calculated directly.

It is often convenient to classify a specific signal being separated as the target, \( s_j \), and all other signals present, \( s_j' \), as the interferer. Even in the case where separation of all the signals is required this designation enables the separation of each target to be described. This is also referred to as the “one versus all” scenario (Vincent et al. 2003).

1.2 What is time-frequency masking?

In the context of BASS, TF masking refers to applying a weighting to each part of a spectro-temporal representation of an audio waveform (Yilmaz & Rickard 2004). This representation can be calculated by a variety of methods provided that they are reversible. A TF cell that is identified as being part of the target signal is given a greater weighting than one corresponding to energy that is mostly part of
Chapter 1. Introduction

the interferer signal. Having applied the weighting to the TF representation, the TF transform is inverted returning the separated signal.

A specific case of a TF mask is the binary mask where elements are weighted either one or zero, meaning they are either entirely included or excluded from the separated signal (Wang 2005). While binary masking allows the separation of audio it can introduce unwanted artefacts to the audio. These artefacts are sometimes referred to as musical noise as it results from “isolated noise energies” at specific frequencies (Wang 2008).

1.3 Applications of blind audio source separation

BASS has many potential applications beyond academic research. The field of broadcasting will be considered in this work as the British Broadcasting Corporation (BBC) is a stakeholder in this project. Within broadcasting there are many potential applications. These include: advanced denoising tools for recordings; separation of content, mixed in legacy formats, so that it can be remixed for a modern surround format; removal of music, for which no licence is available, from a programme; and, providing the audience with a speech-background balance tool.

Away from broadcasting there are further applications available for semantic audio tools such as automated music transcription (O’Hanlon & Plumbley 2014) and speech recognition systems (Raj et al. 2010). Hearing prosthesis could also benefit from being able to identify sounds and decide how useful they are to the impaired listener (Wang 2008).

1.4 Thesis aim

This research in this thesis is framed in the context of the broadcasting industry, where the quality of audio must be good enough that it does not distract from the content of the broadcast being listened to. As future chapters will show, BASS
algorithms do not produce separated audio that is of a high enough perceptual quality for consumption by an audience. The main aim of this thesis is to answer the question: “How can the perceptual quality of BASS be improved for broadcasting applications?”

1.5 Thesis structure

To address the main aim of this thesis nine further questions are answered. These questions are enumerated in Figure 1.1. Some of the enumerated questions are further sub-divided within the chapters in which they are answered. The remainder of this thesis is structured as follows:

- The goal and practice of BASS need to be established in order to give this project a starting point and direction. Chapter 2 therefore establishes what BASS is and how it can be achieved (Question 1).

- In order to assess how successfully an algorithm has separated audio it is necessary to have a way of measuring the quality of the separation. Chapter 3 therefore determines how BASS techniques can be evaluated (Question 2).

- The BBC is a major stakeholder in this project and presents an opportunity for any resulting technologies to be applied. To allow the BBC’s needs to steer the project Chapter 4 establishes the BASS needs of the broadcasting industry (Question 3).

- Studies of literature and industrial practice presented in earlier chapters need to inform further investigation. Chapter 5 determines the subject of experimental investigation (Question 4).

- The quality of audio separated by binary masking needs to be improved before it can be applied in a practical system. Chapter 6 investigates whether binary masking performance can be improved (Question 5).

- A range of sigmoidal masks presents a way of comparing multiple TF masks including the widely-used binary and ratio masks. Chapter 7 identifies the
sigmoidal mask which gives optimum quality separated audio (Question 6).

- As real-world unmixing systems do not have access to known target and interferer signals, Chapter 8 identifies the sigmoidal mask which provides optimum quality under non-ideal conditions (Question 7).

- Further processing of the optimal sigmoidal mask may provide even greater quality improvements. Chapter 9 determines if further improvements to the sigmoidal mask are possible (Question 8).

- While PEASS provides a good model of perceptual audio quality, separated audio may ultimately be consumed by real listeners; their opinions are important. Chapter 10 establishes which TF mask real listeners prefer (Question 9).

- The work in this thesis gives a number of insights into how the perceptual quality of separated audio may be improved and also prompts further questions. Chapter 11 describes how the perceptual quality of BASS can be improved for broadcasting applications, summarises the thesis and proposes further work.
How can the perceptual quality of BASS be improved for broadcasting applications?

1. What is BASS and how can it be achieved?
   - How are BASS problems classified?
   - How can sounds from different sources be separated?

2. How can BASS techniques be evaluated?

3. What are the BASS requirements of the broadcasting industry?
   - What is the role of the sound supervisor?
   - How is noise controlled?
   - What are the implications for an un-mixing tool?
   - What are the opinions of the sound supervisors?

4. What should be the subject of further investigation?
   - For broadcasting applications, which area of investigation looks most promising?
   - What are the current problems in this area?
   - How might these problems be addressed?

5. Can binary masking performance be improved?

6. Which sigmoidal TF mask provides optimal separated audio quality?

7. Which sigmoidal mask provides optimum quality under non-ideal conditions?

8. How might sigmoidal masking be further optimised?

9. Which TF mask do real listeners prefer?

**Figure 1.1:** The question structure of this thesis.
Chapter 2

Blind audio source separation

The opening chapter introduced this thesis’ goal: answering the question, “How can the perceptual quality of BASS be improved for broadcasting applications?” Before discussions about the possibility of improving the quality of audio separated by BASS can begin, an understanding of the current state of the art is required. Specifically, the goal and practice of BASS need to be established to give the project a starting point and a direction. The aim of this chapter is to answer this thesis’ first question: “1. What is BASS and how can it be achieved?” This aim will be achieved by a literature survey answering two sub-questions:

1.1 How are BASS problems classified?

1.2 How can sounds from different sources be separated?

These questions are answered in sections 2.1 and 2.2 respectively. Section 2.1 will provide a black-box understanding by describing BASS systems in terms of their inputs and outputs. Section 2.2 will explore a select group of systems in more depth by detailing how an audio mixture can be processed to reach a desired output.

2.1 Classification of BASS problems

Before attempting BASS it is necessary to define the exact nature of the problem that is to be solved. Authors define the BASS problem in different ways and with differing end goals. This section will answer the first sub-question of this chapter:
“1.1 How are BASS problems classified?”. This will be divided into two areas asking:

- “How are mixtures classified?” in Subsection 2.1.1; and,
- “How are the desired outcomes classified?” in Subsection 2.1.2.

The nature of the mixture determines how it might be separated. A mixture will contain signals produced by multiple sound sources. The assumption is made here that signals are overlapping both in time and frequency; otherwise they may be separated by splicing or filtering. Beyond this assumption the classification of the mixture is considered in three ways: the number of observations available, whether it is time variant and whether it is instantaneous or convolutive. The end result of BASS determines how successful a technique might be considered. At the minimum, a BASS technique must extract at least some information from a source which is contained in a mixture. At full scope, BASS is required to extract multiple sources from a single-channel mixture with complete attenuation of all other components.

2.1.1 Classification of the mixtures to be separated

The BASS problem is modelled in terms of sources, mixtures and a mixing process. Jutten & Hérault (1988) model the problem as

\[ x(t) = A s(t) \]  \hspace{1cm} (2.1)

where \( x(t) \) is one or more mixtures of the signals \( s(t) \) according to the mixing process \( A \). With only \( x(t) \) known, the system is required to determine information about \( A \) and \( s(t) \). As this section classifies different un-mixing problems it defines forms for these three variables. The general form of the BASS problem given in Equation 2.1 is a basis for specific classification. This section will focus particularly on the nature of the mixing matrix \( A \) and the audio implications of its various possible configurations. This section will specifically discuss whether the mixture is: over- or under-determined, stationary or non-stationary, and instantaneous or
Chapter 2. Blind audio source separation

<table>
<thead>
<tr>
<th>Property</th>
<th>Expression</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td>Number of rows, $m$, equal to number of columns, $n$</td>
<td>Problem is determined; $A^{-1}$ can be calculated</td>
</tr>
<tr>
<td>Stationarity</td>
<td>$A(t) = A(t + 1)$</td>
<td>$A$ is immutable. All data can be used in its calculation.</td>
</tr>
<tr>
<td>Instantaneous or Convolutive</td>
<td>2 or 3 dimensional</td>
<td>Each mixture contains multiple copies of each source at multiple delays and amplitudes.</td>
</tr>
</tbody>
</table>

Table 2.1: A summary of the mixing matrix properties that classify the nature of a mixture.

convolutive. A summary of this section is provided in Table 2.1.

The shape of the mixing matrix

BASS is an inverse problem. Given the outcome of a mixing process the input of that process must be sought. The shape of the mixing matrix $A$ determines how a solution to this problem may be sought. If the matrix, $A$, is square and non-singular then it has an inverse, $A^{-1}$, which, if it can be calculated, can then be used to extract $s$ from $x$. When the number of sources exceeds the number of mixtures the mixing matrix is not square and the problem is said to be over-complete or under-determined and no inverse of the mixing matrix exists. The implication of a square matrix for the BASS problem is that there must be exactly as many mixtures as sources. For techniques which rely on being able to find an inverse of the mixing matrix this is a significant constraint. Audio processing often takes many sources and mixes them to give a mono or stereo mixture.

Mixture stationarity

The model for BASS given by Equation 2.1 defines the mixing process, $A$, as time invariant. Assuming mixture stationarity assumes that throughout the recording there is no variation in the level, direction or positioning of any of the sources
Chapter 2. Blind audio source separation

Figure 2.1: A mixture of two sound sources being recorded at one microphone. $t_1$ and $t_2$ are the propagation times. Assuming $t_1 = t_2 \approx 0$ the mixture is instantaneous.

(Hyvärinen et al. 2001). Under circumstances where the audio is not constrained in this way the audio data only remain relevant to the un-mixing problem for a short window. Stationary mixtures have an advantage; all the data (often millions of samples in the case of audio) can be used to calculate the mixing matrix. In the case of a non-stationary mixture, the mixture signal must be windowed and recalculations made regularly to minimise errors due to non-stationarity. The mixture stationarity assumption is rarely valid for anything but the most synthetic of problems.

Instantaneous or convolutive

An important distinction is made between instantaneous and convolutive mixing. Under an instantaneous mixing model, sounds are recorded electronically at generation. In convolutive situations, delays are introduced between the creation of the sound and its recording at a microphone. In a reverberant environment, reflections then arrive at further delays. Pedersen et al. (2008) highlight the differences between instantaneous and convolutive mixtures. An instantaneous mixture happens as close to sound generation as possible. The mixture shown in
Figure 2.2: A mixture of two sound sources being recorded at one microphone in a reverberant environment. This mixture is convolutive.

Figure 2.1 will be instantaneous if the propagation times, $t_1$ and $t_2$, are equal to each other and close to zero. In the case where $t_1 \neq t_2$ the mixture is no longer instantaneous as the mixing of the signals now contains a lag. Instead of the mixing model given in Equation 2.1, each mixture, $x_j$, is now a combination of time shifted source signals

$$x_j(t) = \sum_i a_{ij} s_i(t - \tau_{ij})$$  \hspace{1cm} (2.2)

where $\tau_{ij}$ represents the lag on each signal, $i$ is the source index and $j$ is the mixture index. Allowing each signal to arrive at a delay introduces an additional variable to each mixture: the delay on each signal, $\tau$. This becomes further complicated by the introduction of a reverberant environment. As shown in Figure 2.2, rather than each signal arriving at one delay and amplitude it arrives at many delays and amplitudes.

$$x_j(t) = \sum_i \sum_{\tau} a_{ij}(\tau) s_i(t - \tau)$$  \hspace{1cm} (2.3)
2.1.2 Classification of un-mixing tasks

As well as classifying the nature of the input, it is of interest to specify different required outputs. Different authors pursue differing goals for BASS and clarification of the end goal will be useful as a descriptor of each BASS problem and in choosing a metric with which to assess experimental work. This section classifies un-mixing tasks using a similar structure to that of Vincent et al. (2003). When comparing tasks an important initial distinction is whether or not the end result is to be listened to (Vincent et al. 2003). In this section two scenarios are listed which have listenability as a requirement: extraction and scene modification. Databasing tasks are also described; these are not required to return audio which is to be listened to. A summary is provided in Table 2.2.

Extraction of individual components

The most prevalent goal of BASS is the extraction of individual components (Vincent et al. 2003). In this context, a successful algorithm will return one audio stream for each source in the mixture. In each stream, the ratio of the target source to other signals will be maximised. Using extraction as an aim does cause problems as any imperfections in the separation may be obvious when the separated signal is played in isolation.

Noise Removal

Noise removal can be an important task in the production of high quality audio. While many tools exist to remove spectrally stationary sources, some of which are detailed in Chapter 4 of this thesis, there may be interfering background sounds which do not fit this spectral or temporal profile. Noise removal in this context can be seen as a subset of extraction as the extracted noise is generally discarded and does not need to retain listenable quality but this process must not damage the remaining audio signal. If the mixture is convolutive, a decision must be made about how this affects extraction. Either the reverberation should be maintained
Chapter 2. Blind audio source separation

<table>
<thead>
<tr>
<th>Task</th>
<th>Separation quality required</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>The audio separated must be of the highest possible quality as it could be reused in any circumstance.</td>
<td>Complete remixing of a recording.</td>
</tr>
<tr>
<td>Scene Rebalancing</td>
<td>The process should not damage the quality of the overall audio but light distortion of individual sources are likely to be masked in the mix.</td>
<td>Automated or audience controlled intelligent balancing of a mixture.</td>
</tr>
<tr>
<td>Databasing</td>
<td>Sources must be recognisable by the databasing process but do not need to be listened to.</td>
<td>Archive management, Source Classification, Music Information Retrieval</td>
</tr>
<tr>
<td>De-noising</td>
<td>Target audio must not be damaged but sounds being removed do not need to be retained.</td>
<td>Next generation noise control tools capable of dealing with more than white noise sources</td>
</tr>
</tbody>
</table>

Table 2.2: The various classes of BASS tasks and their end goals

for each signal, extracting the source to sound as if it was on its own in the room, or the algorithm may be required to also perform deconvolution and leave only direct sound in the resulting separation. Whichever decision is made, it should be made explicitly as evaluation techniques described later in this thesis will need to assess for either criterion.

Modification of an audio scene

There are many situations where BASS does not need to fully extract sources but to simply allow the scene which they are in to be modified. This could involve changing the spectral characteristics, spatial positioning or relative levels of individual sources. This would be useful in a number of cases, for example, in a mix where the speech intelligibility is decreased by background sound an increase in level and spectral equalisation will be able to improve the signal. This scene
modification or rebalancing has been identified as a separate task by Vincent et al. (2003).

The point at which a scene modification becomes an extraction needs to be quantified for these two terms to provide a distinction. While extraction will allow complete scene rebuilding, there are a number of potential advantages to pursuing scene modification instead of the extraction approach. If an iterative extraction is attempted then it may be that iterations increasingly damage the remaining signals meaning increased distortion of signals extracted later in the process. The rebalancing approach allows artefacts of the process to be masked in the audio scene. This masking means perceived degradation due to artefacts is reduced. As well as the distortion advantages the rebalancing approach may also be a better approach for non-expert end users. While total extraction provides a tool for skilled restructuring of a soundtrack a rebalancing tool may provide an audience member with a way of altering the ratio of speech to background music in the mix they listen to at home. This can be controlled so the mix quality can not be impaired beyond what is necessary to provide improved speech intelligibility. The amount of flexibility provided by a BASS tool must be appropriately matched to the intended user. Rebalancing is advantageous over extraction when full control of the sources is not required but control over the balance of the mix is. If this tool is well designed it may be usable by audience members within defined constraints.

Databasing tasks

Some audio tasks are focused on extracting information from the audio rather than presenting it to an end listener. These tasks can be collectively referred to as databasing tasks and cover a number of possible scenarios.

Systems that aim to produce a score for a piece of music (Goto & Hayamizu 1999) are an example of a range of music information retrieval (MIR) tasks as described by Typke et al. (2005). Speaker identification as described by Sailaja et al. (2010) is also a databasing task. Within the broadcasting industry, speaker and program identification could be useful tools for management of an archive. The advantage of a databasing approach to BASS is that while the information in the audio...
being separated needs to be preserved the listenability does not. This approach does not seek to change the audio mixture in any way; the only aim is to extract information.

2.1.3 Summary

This section has answered the question “How are BASS problems classified?”. This has been done by classifying different mixtures and outcomes for BASS systems.

The simplest BASS systems are instantaneous, determined and stationary. There is no delay between sound generation and recording, for every source in the mixture there is an observation of the mixture to aid separation and the proportions of each sound in the mixture are time invariant. The reality is that systems are rarely this simple. Real world mixtures are convolutive. Reflections from the recording environment cause additional multiple delayed portions to become part of the signal. Mixing systems are also generally under-determined. Many sources are mixed to one or two channels. This stops a simple matrix inversion being sought to provide the separation. Stationarity is not normally a realistic constraint. The mix is altered by the movement of sources and changes made to electronic mixing equipment.

Classifying outcomes used four high level problems: source extraction, scene rebalancing, databasing and denoising. Source extraction is the most common goal of BASS but is also the most demanding as the resulting signal must be of good enough quality to listen to in isolation. Rebalancing still aims for audio that is listenable but the separation can be less perfect as some artefacts will be masked in the mixture. Databasing tasks focus on the significance of audio features and while not requiring perfect extraction may require more in depth identification of audio elements.
2.2 Methods for BASS

Separating audio from a mixture has been a research interest since Cherry (1953) described the “cocktail party problem”, the human listener’s ability to “recognise what one person is saying when others are speaking at the same time”. This section will answer this chapter’s second question: “1.2 How can sounds from different sources be separated?”. This section will focus on distinguishing sounds from different sources within a mixture. Four key criteria for distinguishing sources are presented: spectral similarity, temporal continuity, spatial distinction and statistical independence. These are described in the context of computational auditory scene analysis (CASA), independent component analysis (ICA) and non-negative matrix factorisation (NMF). At the end of the chapter a comparison of methods is provided in terms of the cues used for separation.

2.2.1 Computational auditory scene analysis

CASA is a technique inspired by Bregman’s (1990) work on source separation by the human auditory system. Bregman’s studies of auditory scene analysis (ASA) detail observations of the human auditory system’s ability to separate sounds from an audio scene. Wang & Brown (2006) bring together a number of key ideas in producing a computational model of ASA.

While the goal of early CASA research was to model the ASA process in humans (Brown 1992; Ellis 1996), researchers have used CASA principles for source separation. When trying to perform CASA inspired source separation, the computational goal is widely believed to be the ideal binary mask (IBM) (Wang 2005). From a TF representation of the mixture audio—which is often obtained using a gammatone filterbank (Patterson et al. 1987)—the IBM designates each TF cell with either a one or a zero depending on whether that element is primarily part of the target source’s energy or not. The processes described in this section aim to aid the calculation of the IBM. Knowledge of the IBM allows the TF cells attributed to a specific source to be re-synthesised into an audio waveform. This re-synthesis is performed by inversion of the masked TF representation. The IBM
is calculated from a signal using three stages:

1. the audio is processed to form a perceptually-relevant TF representation;

2. spectral, temporal and spatial features are extracted and used to segment
   the auditory scene; and

3. the segments are grouped by source.

This section will focus on features that can be used for separation, how they
are extracted by a separation algorithm and how sounds can be subsequently
separated. Spectral cues are discussed first, followed by temporal cues (page 23)
and finally spatial cues (page 26). A summary is given at the end of the section
on page 30.

Fundamental frequency detection

In harmonic sounds, for example pitched musical instruments or voiced speech, the
frequency-spacing of the resonances is related by an integer-multiple relationship.
The lowest frequency is the fundamental, $f_0$, and all its harmonics are related by
integer multiples. Figure 2.3 shows the harmonics of a note played by a viola.
The detection of fundamental frequencies provides a method of grouping parts
of a TF audio representation. Having established which peaks in the frequency
spectrum are fundamentals, peaks at integer multiples of the fundamentals can
be grouped. There are both spectral and temporal methods for $f_0$ detection.

This section details $f_0$ detection using three approaches: spectral, temporal and
spectro-temporal. Each approach is demonstrated using a monophonic example to
establish its principles. Polyphonic problems are then described as an extension
of the monophonic techniques. Throughout this section the example of a viola
playing the note A3 ($f_0 \approx 220$ Hz) is used. The recording is taken from the
University of Iowa’s sample library of anechoic instrument recordings\(^1\). A section
of the waveform from this signal is shown in Figure 2.4, in which the periodicity
of the waveform is clear.

\(^1\)http://theremin.music.uiowa.edu/MIS.html
Chapter 2. Blind audio source separation

Figure 2.3: The spectrogram of a viola playing the note A3. The harmonics are visible as lines running across the image. $f_0$ is the lowest of the harmonics.

Spectral $f_0$ detection There are a number of approaches to extracting the fundamental frequency from a spectrum; some are more generally applicable than others. The example spectrum shown in Figure 2.5 shows the fundamental is both the lowest and the largest peak. However this is not always the case and neither of these factors definitely indicate that a peak represents the fundamental. The fundamental can still be calculated in cases where it is entirely missing. The prevailing technique for calculation of $f_0$ from the spectrum was first applied to speech by Schroeder (1968) and is called pattern matching. This technique involves dividing the frequency of each peak in the spectrum by an incremental series of integers and plotting the results on a histogram. Figure 2.6 shows the final histogram that has been used to recover a fundamental of 220 Hz.

To search for multiple fundamental frequencies in a mixture Parsons (1976)
Figure 2.4: A short segment of the time domain waveform of a viola playing the note A3.

detailed an algorithm that found the first $f_0$ and then removed it along with all peaks harmonically related to it. The residual spectrum was then searched for a further $f_0$ with the detect-and-remove process being iterated as many times as necessary. This technique is flawed in cases where the multiple voices are harmonically related as they will share harmonics at certain peaks. Removal of peaks belonging to more than one source will deprive the algorithm of information to find fundamentals on later iterations.

Temporal $f_0$ detection Calculation of the fundamental frequency is also possible working in the time-domain. While techniques such as measuring the distance between peaks and counting zero-crossings work for a subset of periodic waveforms, a better estimate is obtained by the use of the auto-correlation function (ACF) (Licklider 1951). The ACF measures the correlation of a waveform, $x(n)$, to time-
Figure 2.5: The spectrum of a viola playing the note A3. The harmonic structure is shown by the peaks.


Figure 2.6: Pattern matching performed to find a missing fundamental of 220 Hz. To produce the histogram, locations of the peaks in the spectogram: 440, 660, 880 and 1760 Hz, were each divided by the integers 1 to 10. The most frequently occurring value in the histogram is 220 Hz.
shifted versions of itself. The discrete form of the ACF is given by

\[
\text{ACF}(\tau) = \frac{1}{W} \sum_{n=t+1}^{n=t+W} x(n)x(n + \tau)
\]  \hspace{1cm} (2.4)

where \( t \) is the time index, \( \tau \) the time lag and \( W \) is the window of summation. The ACF produces peaks spaced by \( \frac{1}{f_0} \), the fundamental time period. The ACF of the example viola note is shown in Figure 2.7. The time-domain approach has

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{acf.png}
\caption{The ACF of the viola playing the note A3. The peaks are clearly spaced by 200 samples. This is the time period of the fundamental frequency in samples. Division of 44.1 kHz by 200 samples recovers the frequency.}
\end{figure}

been used to solve multiple \( f_0 \) problems. Like the spectral technique the aim is to remove the first \( f_0 \) detected and then repeat the algorithm to detect the second \( f_0 \). The time-domain approach uses comb filtering with the notches spaced at the same width as the harmonics to remove all components of the initial detected pitch (Frazier et al. 1976). From this point an iterative process similar to the one
described previously for spectral techniques can be implemented.

Spectro-temporal $f_0$ detection  Inspired by Licklider’s (1951) duplex theory of pitch perception, modern CASA systems tend to use spectro-temporal methods to detect $f_0$. Licklider states, “That frequency and period are reciprocally related is not sufficient reason for throwing one away and examining only the other” and then goes on to demonstrate analysis of both within the human auditory system. To realise this in a system the original signal is passed through a filterbank and then an ACF performed on each frequency band (Wang & Brown 2006). The ACFs are then summed across all frequency bands at each lag. The resulting measure is often called the SACF. As with the single ACF approach discussed previously the distance between the peaks gives the period of the fundamental. The SACF is shown in Figure 2.8.

Envelope, onset and offset detection

There are a number of interesting temporal features of audio which can aid BASS. Detection of onset and offset and then full envelope calculation will be detailed here. Onset and offset information is useful to derive as it gives end points which should have common frequency content between them. Detection of the overall envelope of a sound may aid separation as changes in energy can easily be observed. Amplitude information helps identify temporal features of each source. The start and end point of sound events aid segmentation of the audio scene. Onsets and offsets allow the temporal characteristics of segments to be detected even when they occur in the presence of other sounds.

Envelope Calculation  The envelope or amplitude modulation of the signal is a measure of how the amount of energy in a signal changes over time. A common representation of a signal’s envelope is the absolute analytic signal. This can be calculated in four fast Fourier transform (FFT) based steps as demonstrated by Hartmann (1998):
Figure 2.8: Top: the ACF of each output of a bank of 128 gammatone filters (using code from Jin (2007)). Bottom: The SACF taken by summing across the top plot at each lag time. The audio is the same viola sample as used previously.

1. take the FFT of the signal;
2. set the negative frequency components to zero;
3. take the inverse FFT; and
4. calculate the absolute value and multiply it by two.

This technique will provide the absolute analytic signal as shown by the example in Figure 2.9.

Onsets and offsets A number of commonly occurring sounds are marked by a burst of energy as they begin. This is particularly true of spoken plosives and percussive instruments. While the audio signal contains a number of rises and falls these can
Figure 2.9: The absolute analytic signal of an amplitude modulated sine wave. The dashed line shows the modulated sine wave.

not all be interpreted as onsets. It is common to apply smoothing by means of convolution of the signal envelope with the first derivative of a Gaussian function (Wang & Brown 2006),

$$G'(t, \sigma) = \frac{-t}{\sigma^3 \sqrt{2\pi}} \exp \left( -\frac{t^2}{2\sigma^2} \right)$$

(2.5)

where $\sigma$ is the Gaussian width. This convolution with the first derivative provides a smoothed and differentiated signal. From this point, the processed signal’s peaks and troughs above a certain threshold are identified. Peaks are taken as onsets and troughs as offsets (Wang & Brown 2006). The effect of applying this process to some djembe music shown in Figure 2.10.
Analysis of spatial information

The human auditory system is binaural; we each perceive our surroundings using two ears. Binaural hearing provides the ability to determine spatial information about sound sources. Commercial audio is often produced in stereo, allowing spatial information to be conveyed and the listener to detect sound sources’ positions in recordings. Two fundamental binaural cues form the basis for analysis of spatial information: the inter-aural intensity difference (IID) and the inter-aural time difference (ITD). The human auditory system is capable of localisation under reverberant conditions aided by the suppression of delayed signal portions as evidenced by the precedence effect (Litovsky et al. 1999). When a two-channel stereo mixture is available it can be separated spatially with sounds from similar locations assumed to be from the same source. While methods such as
beamforming and multiple signal classification (MUSIC) (Schmidt 1986) rely on large microphone arrays to perform location based separation. In contrast, the techniques focused on here will aim to locate sources from a two-channel recording. The methods described here relate specifically to a recording made with a binaural head but similar techniques can be applied to a two-channel stereo recording. For two-channel stereo recordings created with a spaced pair of microphones there is no model of the human head between the microphones; time and level differences between the channels will be a function of the distance between the microphones and the speed and attenuation of sound in air. As a result, the time and level differences are likely to be smaller with a spaced pair of microphones than a binaural recording head.

The inter-aural time difference The ITD is the difference in time of arrival of a sound at each ear. The speed of sound in air is approximately 340 m/s and sound will arrive at each ear at different times when the source is not anywhere in the median plane of the listener. This difference allows azimuth to be calculated. When working with sine tones, it is common to refer to the inter-aural phase difference (IPD) as a time difference is a phase shift for a pure tone (Wang & Brown 2006). The ITD can be calculated using a cross-correlation of the two signals. Each filter bank channel can be cross correlated with its equivalent in the opposite ear. The cross correlation function (CCF) is calculated as

$$\text{ccf}(n, c, \tau) = \sum_{k=0}^{M-1} a_L(n-k, c)a_R(n-k-\tau, c)h(k) \tag{2.6}$$

using $M$ samples and a windowing function $h$, $n$ and $c$ index the time steps and filter bank channels respectively. The use of the CCF is shown by the plots in Figure 2.11. The lag which gives the maximum cross correlation, $\tau$, can be equated to $\theta$, the azimuth using

$$\tau = \begin{cases} 
(r/c)2\sin\theta & f \leq 500\text{Hz} \\
(r/c)(\theta + \sin\theta) & f > 500\text{Hz}
\end{cases} \tag{2.7}$$
where $r$ is the radius of the head (which is assumed to be spherical) and $c$ is the speed of sound (Wang & Brown 2006).

The inter-aural intensity difference  The IID, sometimes also referred to as the inter-aural level difference (ILD), is caused by the head attenuating the sound reaching the far ear. The IID can be expressed in dB as

$$
\text{IID} = 20 \log(L/R) \tag{2.8}
$$

where $L$ and $R$ represent the sound intensities and the left and right ears. The IID can not be as accurately predicted as the ITD but it is known to be dependent on angle of arrival and frequency, reaching differences of up to 25 dB (Wang & Brown 2006). Computational models of the IID are not as developed as those for the ITD.
Chapter 2. Blind audio source separation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cues</th>
<th>CASA extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral</td>
<td>Each fundamental frequency is detected and grouped with its harmonics</td>
<td>Fundamental frequency detection performed by the summary auto-correlation function. Multiple fundamental frequencies can be extracted by multiple algorithms including the double difference function (DDF).</td>
</tr>
<tr>
<td>Temporal</td>
<td>Onsets and offsets</td>
<td>The signal is convolved with a differentiated Gaussian function. The peaks and troughs of the resulting signal represent onsets and offsets respectively.</td>
</tr>
<tr>
<td>Spatial</td>
<td>ITD and IID</td>
<td>ITD extracted by summary cross-correlation of the left and right signals. IID is measured by comparing the signal intensities in the two channels.</td>
</tr>
</tbody>
</table>

Table 2.3: Summary of CASA’s extraction of spectral, spatial and temporal features.

with the work of Birchfield & Gangishetty (2005) being one of few preliminary studies.

Segmentation and grouping

Having identified key auditory features, a CASA algorithm must use the features extracted to segment the TF representation of the audio. Wang & Brown (2006) define a segment as a TF region where the “underlying acoustic energy originates primarily from the same sound source”. The segments form part of the CASA goal, the IBM, each segment is a part of the IBM for a source. To construct the
IBM for each source its segments must be grouped. Grouping is performed in two stages: simultaneous grouping and sequential grouping. Simultaneous grouping takes segments that occur simultaneously and groups them if they are part of the same source. This can be achieved using pitch tracking, for harmonic sounds, and onset and offset detection for inharmonic parts of the sound (Wang & Brown 2006).

Sequential grouping aims to group sounds from the same source across time. This is more challenging than simultaneous grouping as the characteristics of a source can change over time. Simpler algorithms perform sequential grouping using measurements of spectral similarity or pitch. Model-based grouping algorithms, such as that developed for speech by Barker et al. (2005), compare the segments with models being developed for each source and perform a maximum-likelihood estimation.

Summary

This subsection has answered the question “How does CASA separate audio?”. CASA’s separation of audio is performed by estimating the IBM. The main focus has been on feature extraction, which is necessary to perform this estimation. The key spectral feature is the fundamental frequency, $f_0$, which can be estimated using spectral, temporal or spectro-temporal techniques. Temporal features that can be extracted are the onset and offsets, or, for a more complete picture of temporal activity, the absolute analytic signal can be calculated to represent the temporal envelope. Spatial cues are extracted through calculation of time differences and intensity differences. These features are summarised in Table 2.3. Extracted features are used to group TF elements together to create a binary mask for each source in the mixture. Simultaneous grouping brings together harmonic and inharmonic segments belonging to the same source. Sequential grouping is used to group segments through time, based on the similarity of each section.
2.2.2 Independent component analysis

ICA is a technique being researched for source separation in a number of fields including BASS. Rather than looking to separate the audio in time or frequency, statistical groupings are used. To statistically separate sources, an initial assumption is required: sounds from physically independent sources will produce statistically independent signals (Stone 2004). In a mixture situation where this assumption is valid, an algorithm can then seek to separate the signals in a way that maximises their statistical independence.

ICA is a technique that seeks to separate components based on statistical independence (Jutten & Hérault 1988; Bell & Sejnowski 1995; Hyvärinen 1999). The technique aims to find the inverse mixing matrix which provides the most independent separated source signals. ICA techniques do not calculate independence directly but instead rely on simpler metrics which indicate independence.

The central limit theorem dictates that summing together independent components will lead to a variable which has a distribution that is more Gaussian than the distribution of any of the variables used to create it. From this it can be inferred that the un-mixing matrix providing the least Gaussian separated signals would most likely be the correct matrix. The rest of this section describes different measures of Gaussianity and their application in performing ICA.

There are limitations to the independence assumption: it is a poor assumption for music where multiple sources have been mixed to complement each other’s spectro-temporal content. Puigt et al. (2009) studied the independence of music and speech mixtures when measured over different excerpt sizes. Musical mixtures were shown to be highly dependent over shorter time windows. Shorter time windows are necessary when separating non-stationary mixtures.
Chapter 2. Blind audio source separation

Figure 2.12: Plots of the Laplacian and Gaussian distributions both with zero mean and unity standard deviation.

Kurtosis

Kurtosis is a measure of how raised a distribution is at its central point in comparison with a Gaussian distribution. Figure 2.12 shows the Laplacian and Gaussian distributions and the higher kurtosis of the Laplacian is clear. The classical statistical calculation of kurtosis uses moments

\[ k(x) = \mathbb{E}\{x^4\} - 3|\mathbb{E}\{x^2\}|^2 \]  \hspace{1cm} (2.9)

where \( \mathbb{E}\{\cdot\} \) is the expectation of a variable.
Entropy

Entropy is a measure of the uniformity of a distribution or alternatively can be viewed as the amount of information present in a signal (Shannon 1948). A uniform distribution displays maximum entropy. The entropy of a discrete variable can be evaluated using

\[
H(x) = -\frac{1}{N} \sum_{t=1}^{N} \ln P_x(x_t),
\]

where \( x \) is a variable with \( N \) possible values and \( P_x(x_t) \) is the probability that \( x = x_t \). Entropy is a useful feature of the signal as it can be compared with the entropy of a Gaussian signal to determine how much more information is being given. The concept of comparing signal entropy with the entropy of a Gaussian signal is the basis for negentropy. Negentropy is defined as the difference between the entropy of a dataset and the entropy of a Gaussian distribution. This measure has two desirable characteristics: firstly, it is zero for Gaussian distributions putting the point of reference at the origin; secondly, negentropy is always non-negative. Negentropy can therefore be viewed as a robust measure of how non-Gaussian the signal is. The negentropy, \( J(x) \), of a signal is defined as

\[
J(x) = H(\mu) - H(x),
\]

where \( H(\mu) \) is the entropy of a Gaussian signal with equal mean and variance to \( x \). Negentropy is considered a more robust measure of non-Gaussianity than kurtosis. The measure provided by kurtosis is sensitive to outlying data (due to the rapid growth of the quartic term in Equation 2.9).

Approximating negentropy

Due to a lack of knowledge about the PDFs of the signals to be extracted and for computational efficiency it is necessary to approximate the entropies in Equation 2.11 to give an estimated negentropy. While Equation 2.11 gives a precise calculation of negentropy a more general difference of two functions can be taken.
With $G$ as any non-quadratic function negentropy can be approximated to

$$J(x) \propto [E\{G(x)\} - E\{G(v)\}]^2$$  \hspace{1cm} (2.12)

While this approximation is not always accurate, it will remain consistent with negentropy as a robust non-negative measure of how non-Gaussian a signal is. Hyvärinen (1999) suggests two functions as approximations for entropy. They are selected for their similarity in shape and less than fourth order growth characteristics. These two functions are:

$$G_1(y) = \frac{1}{a_1} \log(cosh(a_1y))$$  \hspace{1cm} (2.13)

$$G_2(y) = -\exp(-y^2/2)$$  \hspace{1cm} (2.14)

with $1 \leq a_1 \leq 2$ but often set as one. These functions are favoured as approximations due to their similarity in shape to kurtosis. Further advantages are given by the ease with which the functions may be differentiated to obtain their gradients. A graphical representation of the negentropy approximations is shown in Figure 2.13.

**Iteration**

With a model for independence an un-mixing matrix must then be initialised and optimised to produce maximally independent separated signals. This optimisation is performed by use of a gradient algorithm, the most popular of which is Hyvärinen & Oja’s (1997) FastICA algorithm. The optimisation algorithm updates the un-mixing matrix so that the independence of the audio it separates is optimised towards a maximum. Assuming the independence assumption was valid, the algorithm will converge to provide an un-mixing matrix will that produce separated sources.
Figure 2.13: The functions in equations 2.13 and 2.14 with the quartic term as used in kurtosis for comparison.
Extension to the TF case

Whilst it is common to explain ICA in the time-domain it is not purely a time-domain technique. One case of adaptation of the mixture model and cost functions to further dimensions is detailed in Naik & Kumar (2011). Redefining the sources as

\[ s = C\Phi \]  \hspace{1cm} (2.15)

for coefficients \( C \) of the TF basis \( \Phi \). The new mixture model

\[ x = AC\Phi \]  \hspace{1cm} (2.16)

allows an estimate of \( AC \) to be calculated giving the weightings of the TF regions for the separated sources. These matrices are calculated using a maximum a posteriori approach expressed as

\[
\max_{A,C} P(A, C|x) \propto \max_{A,C} P(x|A, C)P(C) 
\]  \hspace{1cm} (2.17)

2.2.3 Non-negative matrix factorisation

NMF was introduced by Lee & Seung (1999) for the learning of images by parts. The algorithm has also proved useful for BASS. The NMF approach is similar to ICA: it aims to factorise a mixture into two matrices. The constraints of NMF are different to those of ICA. The requirement for a mixture to be determined is removed and the NMF approach works with single channel audio. Extra information is gained by using a TF representation of the input signal.

NMF aims to factorise the TF representation of a mixture into two matrices, the bases, \( W \), and the coding \( H \). The bases matrix is formed from a set of unique spectral structures; each basis does not represent a source in the mixture but rather temporally-coincident energy that is part of a single source. For example, the signal from a piano would be divided into each individually occurring note or
speech into individual formants. The number of bases, usually termed $r$, is taken in by the algorithm as prior information. Example bases and coding matrices are shown in Figure 2.14.

![Figure 2.14: The bases and coding matrices produced by an NMF algorithm. Two flute notes played separately and then together are shown. Their spectral information is contained in $W$, the left-hand matrix, while their activations are shown in $H$, the right-hand matrix. Produced using the NMF library available made available by Grindlay (2010).](image)

After giving the algorithm a TF representation, $V$, and the number of bases, $r$, each matrix is initialised with random non-zero values to prevent divide-by-zero errors. The algorithm then iterates update rules on $W$ and $H$ with the improvements in one allowing the other to be further improved. The bases or dictionary matrix $W$ is iterated by two update rules. The first works through time ensuring the amount of energy assigned to each frequency in each component is consistent with the total amount of energy in that frequency and the latest value estimate for the
activation of that component at that point in time. This is achieved by the first update equation,

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$$  \hspace{1cm} (2.18)

where $i$ indexes frequency, $\mu$ indexes time steps and $a$ indexes bases. A ratio greater than one causes the algorithm to add energy from that frequency and less than one takes away energy. The ratio is multiplied by $H$ for that component at that point in time. This means energy is apportioned to the component according to its level of activation at that point in time.

The second update rule for $W$ updates the matrix to constrain each column to sum to unity. This prevents more energy being apportioned at a given frequency than was present in the original signal. This is achieved by dividing each element by the sum of its column

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$  \hspace{1cm} (2.19)

The coding matrix $H$ is updated by only one equation that is similar to Equation 2.18. This works through the frequency spectrum and increases the activations of components that contain more spectral energy and reduces the activation of the signals with less spectral energy. The update equation is given as,

$$H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$  \hspace{1cm} (2.20)

where $i$ is used as a frequency index. For stopping conditions the algorithm either runs to a fixed number of iterations or measures the change on each update and stops when this change falls below a minimum value.

Smaragdis & Brown (2003) highlighted the issue of prior information in NMF algorithms. As this algorithm separates audio “based on the system’s accumulated experience from the presented input and not on predefined knowledge” this means that “all unique events are understood to be a new component”; simultaneous components that have not occurred already separately are recorded as one component. Therefore, were the example in Figure 2.14 to have occurred in reverse order, i.e. the two notes together and then separately, the algorithm would have
identified the combined note as one component and the first individual note as the second. This problem can be overcome by providing the algorithm with training data. This process initialises the basis matrix with expected components rather than random values. Providing the algorithm with prior information has been shown to improve results compared to using randomly initialised matrices (Wang & Plumbley 2006).

Spatial information

The basic algorithm presented above is designed for single-channel audio and hence does not offer a method of calculation of the position of sources. An extension to the NMF algorithm presented by Parry & Essa (2006) does allow a further matrix to be added to the factorisation which details spatial positioning for a two channel mixture. This technique is built assuming spatial stationarity and may prove difficult to adapt to signals which are moved in space. The initial factorisation is demonstrated in Figure 2.15.

\[
\begin{align*}
V_{\text{TF Representation}} & \quad = \quad W_{\text{Spectral Information}} \cdot Q_{\text{Spatial Information}} \cdot H_{\text{Temporal Information}} \\
\end{align*}
\]

**Figure 2.15:** The factorisation performed by Parry & Essa (2006). Showing the dimensions and arrangement of the spectral, spatial and temporal matrices. The algorithm’s inputs are \( V \), the TF representation, and \( r \) the number of sources.

2.2.4 Discussion

This chapter has so far detailed different features which can be used for BASS and how they are utilised by the different techniques. This section will now briefly compare the techniques. The comparisons made in this section relate to cues and
problem classification. The comparisons of cues is summarised in Table 2.4 and the comparison of problem classification in Table 2.5.

Spectral cues

The CASA approach to extraction of spectral cues is detection of the fundamental frequencies and grouping with harmonically-related energy. This provides a convenient way of grouping together spectral data but does not provide the entire spectral envelope of each component; it relies on grouping harmonic and inharmonic parts of a sound at a later stage. NMF approaches can separate spectral information by component. This process has been shown to be reliant on the understanding that unique events are individual components (Smaragdis & Brown 2003). NMF has been shown to be more effective when it is given prior information about the components in the mixture (Wang & Plumbley 2006). While CASA will spectrally group events based on harmonic relationship it uses temporal data to help further group the data. NMF bases its entire spectral grouping on temporal coincidence.

Temporal cues

CASA applies onset and offset detection methods to segment acoustical energy into temporal events. It is able to segment the auditory scene into as many events as it detects. In the NMF approach the number of acoustical events is defined by the prior information put into the algorithm. The algorithm will then identify as many unique events as specified starting from the beginning of the signal.

Spatial cues

CASA uses the summary cross-correlation function to detect the ITD and from this angle of arrival can be calculated. NMF is able to iteratively detect spatial information as part of its algorithm although this is limited to objects that are spatially stationary.
<table>
<thead>
<tr>
<th>Method</th>
<th>ICA</th>
<th>NMF</th>
<th>CASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral</td>
<td>-</td>
<td>The spectral bases matrix is initialised with random values to contain ( r ) spectral components for the mixture. These values are then iteratively updated, alternating with the temporal data, to calculate the spectral components.</td>
<td>Spectro-temporal ( f_0 ) detection by SACF or DDF for multiple ( f_0 ) detection. Segmentation performed by common ( f_0 ).</td>
</tr>
<tr>
<td>Temporal</td>
<td>-</td>
<td>The temporal activations matrix is initialised with random values and then iterated alternately with the spectral information to produce the activation data.</td>
<td>Onset and offset detection used to segment the audio temporally.</td>
</tr>
<tr>
<td>Spatial</td>
<td>-</td>
<td>NMF modifications allow spatial information detected in multi-channel audio. The spatial mixing matrix is initialised randomly and updated as part of the NMF cycle. The sources are assumed spatially stationary.</td>
<td>ITD estimated by taking a SCCF of each filterbank channel.</td>
</tr>
<tr>
<td>Statistical</td>
<td>Gradient algorithms used to maximise independence of separated signals.</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 2.4:** A comparison of the processing of spectral, temporal, spatial and statistical data by the ICA, NMF and CASA approaches.
Chapter 2. Blind audio source separation

Statistical cues

ICA makes use of statistical cues to separate audio. The algorithm seeks to separate mixtures in the way which provides the most statistically independent audio streams. Typically the negentropy of separation is calculated and maximised using a gradient function.

Problem classification

In Section 2.1, it was observed that BASS problems can be classified by the ratio of the number of sources to the number of mixtures. This concept is revisited in the discussion here so it can be related to the techniques discussed in this section. Table 2.5 shows the required number of mixtures for each technique discussed in this section.

2.2.5 Summary

This section has answered the question “How can sounds from different sources be separated?” Taking CASA, ICA and NMF in turn, particular focus was placed on their handling of spectral, temporal, spatial and statistical information. CASA’s modelling of the human ASA process means it handles spectral information using fundamental frequency detection and grouping together harmonically related information. Temporal onsets and offsets are used to segment the TF representation of the audio into as many individual events as are detected. These events are then grouped into streams believed to have originated from the same source. CASA models the human auditory system’s binaural processing of spatial information by performing a SCCF to detect the ITD. ICA relies on the assumption that physically independent sources will produce statistically independent signals. An ICA algorithm aims to factorise the mixture into its sources and mixing matrix. Sounds that satisfy the independence assumption can be separated by calculation of the un-mixing matrix and taking statistical measures of the separated signals it outputs. The proposed separated signals are measured.
### Table 2.5: The number of mixtures required to separate audio using different BASS methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of mixtures</th>
<th>Reason</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>$m = 1$</td>
<td>A single channel is used and then analysed in the TF domain by use of the STFT or similar transform.</td>
<td>Lee &amp; Seung (1999) Virtanen (2007)</td>
</tr>
<tr>
<td>CASA</td>
<td>$1 \leq m \leq 2$</td>
<td>Many techniques possible with one channel. Two channels used to model the binaural processing of spatial information.</td>
<td>Wang &amp; Brown (2006)</td>
</tr>
<tr>
<td>ICA</td>
<td>$n \leq m$</td>
<td>To recover $n$ sources. ICA requires at least $m$ mixtures. Performing ICA on a sparse audio representation can remove this constraint.</td>
<td>Jutten &amp; Hérault (1988) Naik &amp; Kumar (2011)</td>
</tr>
</tbody>
</table>
for either kurtosis or approximations of negentropy. These provide metrics which model independence without requiring the knowledge of a signal’s PDF. NMF is a technique which developed from ICA and also seeks to factorise the mixed audio into two parts. These take the form of the spectral bases and temporal activations. The algorithm takes a TF representation of the mixed audio and the number of bases contained within it as its input. This information is used to initialise the bases and activations matrices which are then optimised iteratively up to a fixed number of iterations or until the amount of change decreases below a minimum. Advancements to NMF allow it to process spatial information from multi-channel audio.

2.3 Chapter summary

In this chapter, a literature review was conducted to answer the question, “1. What is BASS and how can it be achieved?”. This chapter’s question was answered as two sub-questions:

1.1 How are BASS problems classified?

1.2 How can sounds from different sources be separated?

Answering the first sub-question demonstrated that BASS problems are categorised in terms of the type of mixture which is to be separated and the desired result at the end of the process. State-of-the-art BASS techniques must separate mixtures which are convolutive, non-stationary and under-determined. The desired end result varies between extraction of information about a source and extraction of the entire source from a mixture.

The second sub-question was answered by detailing the leading methods for audio un-mixing: CASA, ICA and NMF. These techniques come from different areas of academic research but use a common group of cues to separate audio. CASA and NMF use spectral and temporal cues and in multi-channel scenarios can use spatial information as well. ICA focuses on statistical cues.
Chapter 3

Evaluating audio source separation

The BASS techniques discussed in this thesis so far aim to produce audio separated from input mixtures. In order to assess how successfully each algorithm has separated the audio it is necessary to have a way of measuring the amount of separation. The aim of this chapter is to answer the question: “2. How can BASS techniques be evaluated?”.

This chapter details methods of comparing an extracted signal with the true pre-mixture signal either using the final audio output of the algorithm or the unmixing coefficients. The literature review details both objective and subjective evaluation of separated audio. Objective metrics allow calculations to be made on a separated signal to assess how well it has been separated whereas subjective evaluation allows the experience of the listener to make this judgement. Recent attempts to provide perceptually valid objective measures are also discussed.

When the aim of source separation is to separate one source from a mixture, the success of an algorithm is determined by how similar the separated source is to the true source before it was mixed. Any difference between the signals must be accounted for. The metrics discussed in this chapter are important to this project as a metric will be required for use in experimental work.

The remainder of this chapter is structured as follows: Section 3.1 details objective measurements of separation. In Section 3.2, subjective methods of separation assessment are listed. Section 3.3 describes a system for objectively modelling subjective metrics. A discussion of the relative merits of the methods detailed is included in Section 3.4.
3.1 Objective evaluation

Objective evaluation of source separation involves measuring the differences between the target signal and the estimate from the algorithm. Three different approaches are described here. They are all based on taking different ratios with parts of the target and the estimate.

3.1.1 Schobben et al.’s method

An evaluation system for BSS techniques was introduced by Schobben et al. (1999). The technique describes the distortion of the source and the amount of separation. Both metrics are SNR inspired and make use of expected values.

The distortion is expressed, in dB, by taking the ratio of the root mean square (RMS) difference between the amount of $s_j$ in the mixture $x_i$, and the source signal $s_j$ to the amount of energy in $x_i$ due to $s_j$. Using $x_{i,s_j}$ to represent the energy in $x_i$ due to $s_j$ gives the equation

$$D_j = 10 \log \left( \frac{E\{(x_{i,s_j} - \alpha_j \hat{s}_j)^2\}}{E\{(x_{i,s_j})^2\}} \right)$$

(3.1)

where $\alpha_j = E\{x_{j,s_j}^2\}/E\{\hat{s}_j^2\}$, the ratio of the amount of $s_j$ in $x_i$ to the estimate $\hat{s}_j$. This is included to scale $s_j$ so that, under zero-distortion circumstances, it equals $x_{i,s_j}$. This allows the equation to tend to $-\infty$ for perfect separation.

The separation is measured by the ratio of the squared expected value of the target signal to the squared expected value of all other signals. This is expressed as,

$$S_j = 10 \log \left( \frac{E\{(\hat{s}_{j,s_j})^2\}}{E\{\sum_{j' \neq j} (\hat{s}_{j'})^2\}} \right)$$

(3.2)

where $s_{j'}$ is used to represent all the signals except the target.
3.1.2 The ideal binary mask ratio

Hummersone et al. (2011) define the IBMR as a metric for assessing binary mask based BASS algorithms. The IBMR is defined in terms of $\lambda$, the number of cells correctly identified as containing source energy and $\rho$, the number of cells incorrectly identified. Correctly identified zeros are not included in the calculation. The IBMR is formulated as

$$\text{IBMR} = \frac{\lambda}{\lambda + \rho}$$

(3.3)

where for an estimated mask, $m$, and the ideal mask $m_{ibm}$

$$\lambda = \sum_{ij} m(i, j) \wedge m_{ibm}(i, j)$$

(3.4)

$$\rho = \sum_{ij} m(i, j) \oplus m_{ibm}(i, j)$$

(3.5)

with $\oplus$ representing a logical XOR and $\wedge$ representing a logical AND.

This metric is considered advantageous by Hummersone et al. (2011) in environments where robustness to convolutional distortion is important. Under such circumstances, the IBMR is shown to be more robust than simple SNR based metrics. Whilst this makes the IBMR a useful metric, it is limited in its application by the fact that it can only assess binary-mask-based separated algorithms.

3.1.3 Vincent et al.’s method

Vincent et al. (2006) describe a method contained in their BSS_eval toolbox. Their method is based on the idea of decomposing the error into three parts:

$$\hat{s}_j = s_j^{\text{target}} + e_j^{\text{interf}} + e_j^{\text{noise}} + e_j^{\text{artif}}$$

(3.6)

where the $e$ terms represent error due to interference, noise and artefacts respectively. This technique provides a fuller description of the error than the techniques already discussed and this definition ensures they are superposable.
Each error term is calculated differently depending on the “allowed distortions” (Vincent et al. 2006). The allowed distortion depends on the BASS task being attempted. In convolutive mixtures the allowed distortion may be a time-invariant filter whereas in instantaneous mixtures a time-invariant gain is the most likely permitted distortion. This section will focus on the time-invariant gain; further distortions are detailed in the original work.

The $e$ terms are separated out by means of a series of projections. Orthogonal projections are made using the estimate and the source, the interferers and the estimate, and the interferers and noise.

An orthogonal projection maps a vector onto a subspace. Finding the angle between two vectors and the norm of the vector being projected allows the projection to be calculated (Smith III 2007). The angle between two vectors is given by the inner product (denoted by $\langle \cdot, \cdot \rangle$) and the norm (denoted by $\| \cdot \|$) is calculated as the inner product of a vector with itself. This gives the orthogonal projector of $v_1$ onto $v_2$ as

$$P_{v_2}(v_1) = \langle v_1, v_2 \rangle \frac{v_1}{\|v_1\|^2} \quad (3.7)$$

To perform relevant projections, Vincent et al. (2006) define three projectors: $P_{s_j}$ the projector on to $s_j$, the true signal; $P_s$ the projection onto $s$, the interfering sources; and $P_{s,n}$ the projection on to the interfering sources and noise sources. The projectors are defined here for a time-invariant gain, to use the system with a different allowed distortion the projectors must be redefined but the remainder of the process is the same. Using these projections the terms in the decomposition shown in Equation 3.6 can be calculated as:

$$s_j^{\text{target}} = P_{s_j}(\hat{s}_j) \quad (3.8)$$

$$e_j^{\text{interf}} = P_s(\hat{s}_j) - P_{s_j}(\hat{s}_j) \quad (3.9)$$

$$e_j^{\text{noise}} = P_{s,n}(\hat{s}_j) - P_s(\hat{s}_j) \quad (3.10)$$

$$e_j^{\text{artif}} = \hat{s}_j - P_{s,n}(\hat{s}_j) \quad (3.11)$$
Chapter 3. Evaluating audio source separation

**Error due to interference**  The presence of interfering sources in a separated source is represented by $e_{j}^{\text{interf}}$. Calculation of the interference is achieved by projecting the estimate onto each source. The equation,

$$e_{j}^{\text{interf}} = \sum_{j' \neq j} \langle \hat{s}_j, s_{j'} \rangle \frac{s_{j'}}{||s_{j'}||^2}$$  \hspace{1cm} (3.12)

provides a decomposition for $e_{j}^{\text{interf}}$ for orthogonal sources. The inner product of the estimated source and the true interferers will evaluate to zero in the case of a perfect separation.

When the sources are not orthogonal a more complex decomposition is required. This involves creating a matrix containing the inner product of each interferer with the target source. This is then transposed and multiplied by each interfering signal in turn.

**Error due to noise**  The noise sources, $\mathbf{n}$, can be assumed to be orthogonal to the target signals $\mathbf{s}$. This allows the projection of $\hat{s}_j$ on to the noise sources and interferers to be calculated as a superposition of $P_s(\hat{s}_j)$ and $P_n(\hat{s}_j)$, the sum of the projector of $\hat{s}_j$ on to each noise source $n_i$.

$$P_{s,n}\hat{s}_j \approx P_s\hat{s}_j + \sum_{i=1}^{m} \langle \hat{s}_j, n_i \rangle \frac{n_i}{||n_i||^2}$$  \hspace{1cm} (3.13)

This assumption can be used to simplify Equation 3.10 to

$$e_{j}^{\text{noise}} = P_n\hat{s}_j.$$  \hspace{1cm} (3.14)

**Error due to artefacts**  The error due to artefacts is defined as distortions not due to interfering sources or noise. This is designed to include any distortion introduced by the processing of the BASS method. Artefacts may be introduced by inverse filtering processes on convolutive mixtures or loss of phase relationships between frequency bands when using a TF method.
**Performance measures** Vincent et al. use their decomposition to define the following performance metrics: the source to distortion ratio

\[
\text{SDR}_j = 10 \log_{10} \frac{\| s_j^{\text{target}} \|^2}{\| e_j^{\text{interf}} + e_j^{\text{noise}} + e_j^{\text{artif}} \|^2} \tag{3.15}
\]

the source to interferer ratio

\[
\text{SIR}_j = 10 \log_{10} \frac{\| s_j^{\text{target}} \|^2}{\| e_j^{\text{interf}} \|^2} \tag{3.16}
\]

the source to noise ratio

\[
\text{SNR}_j = 10 \log_{10} \frac{\| s_j^{\text{target}} + e_j^{\text{interf}} \|^2}{\| e_j^{\text{noise}} \|^2} \tag{3.17}
\]

and the source to artefact ratio

\[
\text{SAR}_j = 10 \log_{10} \frac{\| s_j^{\text{target}} + e_j^{\text{interf}} + e_j^{\text{noise}} \|^2}{\| e_j^{\text{artif}} \|^2} \tag{3.18}
\]

The above metrics are all expressed in dB and are similar to the SNR familiar in engineering. Using these power ratios the resulting signal and hence the algorithms they came from can be evaluated in terms of these four areas. Vincent et al. also go on to suggest time-localised calculations of the ratios for windowed portions of the signals to account for variations in each metric across the signal.

**Spatial Distortion** In further work (Vincent et al. 2007), this technique is expanded for use with stereo mixtures by adding the image to spatial distortion ratio (ISR). The error due to spatial distortion, \( e_j^{\text{spat}} \), is calculated by use of orthogonal projection of the estimated source on to the true source and then subtraction of the target signal. The ISR is then calculated by taking the ratio

\[
\text{ISR}_j = 10 \log_{10} \frac{\| s_j^{\text{target}} \|^2}{\| e_j^{\text{spat}} \|^2} \tag{3.19}
\]
3.2 Subjective assessment of separation quality

The metrics presented in Section 3.1 can be used to assess the audio separation of a given algorithm. Clearly less noise, artefact and interference in each source is desirable but the effect that any improvement will have on a listener’s perception of the audio is not necessarily clear.

Listening tests can be performed using the separated audio and asking listeners to rate its quality or identify words. The listening test can be argued to be the most relevant test for applications where the end goal is listenable audio.

3.2.1 Stubbs & Summerfield’s test

Stubbs & Summerfield (1988) devised one of the earliest tests for speech separation algorithms. Listeners were asked either to identify words from a mixture or to rate the quality of the audio. The experiment demonstrated that the algorithms on test improved speech intelligibility for impaired and unimpaired listeners.

3.2.2 Kornycky et al.’s test

Kornycky et al. (2008) created a source separation modelling system that allowed synthesised separated signals to be created from a weighted sum of a target signal, a randomly filtered interferer signal and Gaussian noise. The test implements the ITU-T P.835 (2003) recommendation for testing of a noise suppression algorithm. Kornycky et al. modified the questions from the ITU-T P.835 to ask assessors about the distortion of the background, how intrusive the background was and the quality of the separation. The mean opinion score (MOS) data generated from these tests is then compared to Vincent et al. (2006)’s metrics, which were previously discussed in Subsection 3.1.3. The results show that MOS data for intrusiveness and separation are highly correlated with SDR, while the results for SIR are highly correlated with the perceived background distortion.
3.2.3 Emiya et al.’s test

As part of their research into perceptually inspired objective metrics Emiya et al. (2011) performed listening tests asking subjects to rate the audio using four categories of degradation:

1. global quality;
2. preservation of the target source;
3. suppression of other sources; and
4. absence of additional artificial noise.

These categories are inspired by the objective model presented in Subsection 3.1.3. They do not provide a superposable decomposition as global quality is an opinion score taken from the listener not the sum of the other distortions.

3.3 Objective modelling of subjective metrics

This chapter has so far described assessment of the quality of a separation in terms of objective ratios and subjective scoring. The ratios described in Section 3.1 provide an easy-to-compute description of the quality of a separation thus making comparisons of different techniques on the same problem easy. The problem with these metrics is they do not give any information about the way a listener perceives the quality of a separation. Section 3.2 described how listening tests can be used to provide subjective assessments of the quality of a separation. These metrics are time consuming to generate and process in comparison to objective measures. This section will explore methods of using objective techniques to generate metrics which bear closer resemblance to subjective metrics generated by listening tests.

Furthering Vincent et al. (2006)’s work, Emiya et al. (2011), and later Vincent (2012), propose a system of metrics which provide an evaluation similar to that produced by listening tests. The system is named PEASS. Emiya et al.’s technique
Chapter 3. Evaluating audio source separation

obtains a perceptual similarity measure (PSM) for audio and then applies a non-linear weighting to bring the results closer to those that would be provided by a listening test. This section will describe the PEMO-Q test for producing PSMs for audio with a reference and the PEASS system which incorporates PEMO-Q.

3.3.1 PEMO-Q

With low bit-rate codecs being used for an increasing number of audio applications since the 1990s, research has been conducted and methods standardised for assessing the perceptual quality of audio in a number of situations. Initially, efforts focused on speech intelligibility over narrow-band codecs (ITU-R P.861 1996; ITU-R P.862 2001) but soon the evaluation of wide-band signals was also required.

PEMO-Q is an algorithm for the measurement of the perceptual similarity of a piece of audio to a reference (Huber & Kollmeier 2006). Building on the perceptual evaluation of audio quality (PEAQ) model (ITU-R BS.1387 1998), Huber & Kollmeier aimed to generate a “psychoacoustically validated model of auditory processing” which predicts perceived audio quality for any type of distortion to any type of signal. The end goal of PEMO-Q is the PSM, which is a cross-correlation of the reference and test audio, each having been passed through a model of the human auditory system.

Figure 3.1 represents the PEMO-Q system. The system produces two measures: an overall measure of perceptual similarity, PSM, and a time-localised measure PSM_t. The system uses a model of the human auditory system similar to that described in Wang & Brown (2006).

The output of these auditory models are assimilated and cross-correlated to calculate the PSM. The PSM_t is calculated using a weighted sum of the moving average and instantaneous audio quality.
Figure 3.1: A block diagram of the PEMO-Q system. Adapted from Huber & Kollmeier (2006)
Chapter 3. Evaluating audio source separation

Assimilation

The output of the auditory model is assimilated by halving the differences between each TF cell in the true signal, \( s_{tf} \), and the value in the equivalent cell in the estimate \( \hat{s}_{tf} \). When the absolute value of the estimate is smaller than that of the true signal the estimate value is replaced with the mean of the estimate and the true value. This gives the assimilated estimate,

\[
\hat{s}_{tf} = \begin{cases} 
\frac{\hat{s}_{tf} + s_{tf}}{2}, & |\hat{s}_{tf}| < |s_{tf}| \\
\hat{s}_{tf}, & |\hat{s}_{tf}| \geq |s_{tf}|
\end{cases}
\]  

This process is designed to mitigate the effect of missing components on the quality of the signal and emphasise the effect of additional components. The method is attributed to Berger (1998). Huber & Kollmeier (2006) state: “This approach follows the hypothesis that ‘missing’ components in a distorted signal are less disturbing than ‘additive’ components.”

Cross correlation

Having been assimilated the signals are cross-correlated to provide the PSM. Cross-correlation involves centring the reference and test signals before multiplying each TF value in the reference signal by the matching value in the test signal. Dividing by the magnitude gives \( r \), the cross-correlation coefficient of the two signals as

\[
r = \frac{\sum_{i,j=1}^{N,M} (s_{tf} - \bar{s})(\hat{s}_{tf} - \bar{s})}{\sqrt{\sum_{i,j} (s_{tf} - \bar{s})^2 \sum_{i,j} (\hat{s}_{tf} - \bar{s})^2}}
\]  

(3.21)

where \( N \) and \( M \) are the dimensions of the TF representation of the audio produced by the auditory model. For simplicity the signal index, \( j \), has been omitted.
3.3.2 The PEASS system

PEASS is a system of objective measures that aims to provide a better prediction of the perceptual quality than those in Equations 3.15–3.17 (page 50) and 3.19 (page 50). The process uses two stages to generate its metric: firstly, a series of PSMs are calculated and then a non-linear mapping is applied to give a better fit to the expected responses of listeners. The PEASS system is shown in the block diagram in Figure 3.2.

As Figure 3.2 shows the system is divided into two sections. Firstly, PSM measurements are taken and then a non-linear mapping is applied to enhance the similarity of the scores to those produced by a listening test. The perceptual saliences are assessed using the following PSM calculations:

\[
q_{\text{overall}}^{j} = \text{PSM}(\hat{s}_j, s_j)
\]
\[
q_{\text{target}}^{j} = \text{PSM}(\hat{s}_j, \hat{s}_j - e_{\text{target}}^{j})
\]
\[
q_{\text{interf}}^{j} = \text{PSM}(\hat{s}_j, \hat{s}_j - e_{\text{interf}}^{j})
\]
\[
q_{\text{artif}}^{j} = \text{PSM}(\hat{s}_j, \hat{s}_j - e_{\text{artif}}^{j})
\]

In these definitions \(e_{\text{interf}}^{j}\) and \(e_{\text{artif}}^{j}\) are as defined previously in Subsection 3.1.3. \(e_{\text{target}}^{j}\) represents distortion of the target source.

Having calculated these measures of perceptual salience each metric is passed through a non-linearity built around a network of sigmoid functions, based on \(g(x) = 1/(1 + e^{-x})\). Each sigmoid is weighted and biased giving the final output the form

\[
f_r(q) = \sum_{k=1}^{K} v_{rk} g(w_{rk}^T q + b_{rk})
\]

where \(w_{rk}\) represents the input weighting, \(b_{rk}\) represents the sigmoid biasing and \(v_{rk}\) is an output weighting. \(q\) is a vector of the four PSMs defined as \(q = [q_{\text{overall}}^{j}, q_{\text{target}}^{j}, q_{\text{interf}}^{j}, q_{\text{artif}}^{j}]\). The system is indexed by the size of the sigmoid network, \(K\), which ranges from 1 to 8; and \(r\), which indexes the \(q\) metrics by
Figure 3.2: A block diagram of the PEASS system. Adapted from Emiya et al. (2011)
Chapter 3. Evaluating audio source separation

ranging from 1 to 4. The values of the weight and bias are varied for each \( q \) function. Emiya et al. trained the neural-network against a database of 6400 subjective scores in order to create the predictions of perceptual quality.

This network delivers the final results of the PEASS system. The vector \( q \) has been weighted in four different ways to provide the four PEASS metrics: the target-related perceptual score (TPS), the IPS, the APS and the OPS. Each measure is scaled from 0 to 100 with higher values always being more desirable.

3.4 Discussion

The evaluation metrics discussed in this chapter aim to quantify the success of a separation from a mixture. Selecting a metric to compare techniques can depend on a number of factors, a few of which will briefly be considered here. This discussion is summarised in Table 3.1.

An important factor in the choice of an evaluation algorithm is its relevance to the task. As previously demonstrated, there are many different types of BASS mixtures and end goals. The PEASS system allows users to specify which distortions are allowed and has also been used to evaluate a number of different BASS tasks (Araki et al. 2012). In comparison the IBMR is built for situations where the preservation of convolutional distortion is important in separation.

A second factor to be considered is the amount of detail that is provided about the distortion. While single metrics provide a comparison point between two systems, a decomposition of the distortion will allow researchers to see where algorithms have strengths and weaknesses. This extra detail will help the improvement of existing algorithms and set design goals for future techniques. Schobben et al. (1999) provides two metrics but they are not decompositions of the total distortion. The metrics provided by the PEASS system are more detailed and provide the advantage of perceptual relevance.

A final consideration is how widespread the use of a given metric is. The purpose of metrics is to provide a reference point between researchers. If a metric is already
in widespread use it provides more opportunities for comparison of results. Araki et al. (2012) details the use of PEASS to evaluate the work of 18 different research institutions whose algorithms are aimed at a range of BASS problems.

Based on the above factors the use of the PEASS system as an evaluation tool for BASS research is recommended. This research will eventually focus on a specific subset of BASS tasks and the PEASS will provide flexibility to evaluate whichever area of work is chosen. A number of recent algorithms have already been tested with this software providing a number of comparison points for future developments.

3.5 Chapter summary

This chapter has answered the question, “2. How can BASS techniques be evaluated?” by surveying literature about BASS evaluation. Evaluation techniques can be split into objective and subjective metrics. Both techniques use known premixture signals to compare with the extracted signal. In an objective metric this is performed by computational measurement of the differences between the extracted and true signals. A subjective measurement is calculated by asking users to rate different parts of the signal. Due to its widespread use and multi-dimensional description of quality, the PEASS computational model of listener ratings was deemed most useful for further work in this project.
<table>
<thead>
<tr>
<th>Metric</th>
<th>Relevance</th>
<th>Detail</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSS Eval</td>
<td>Designed to give choice of one of four “allowed” distortions: time-invariant gain, time-varying gain, time-invariant filtering and time-variant filtering. All other distortion contributes to objective metrics.</td>
<td>The original work prescribes four metrics: SDR, SIR, SNR and SAR. A follow up paper introduces the ISR.</td>
<td>Used in two cross-institutional campaigns. (Vincent et al. (2009) and Araki et al. (2012))</td>
</tr>
<tr>
<td>IBMR</td>
<td>Designed for situations where convolutional distortion must be accounted for.</td>
<td>Provides a single metric.</td>
<td>Used by some authors working on speech problems (Yu et al. 2014).</td>
</tr>
<tr>
<td>Listening Test</td>
<td>The listening test can be modified to suit different tasks. Results accurately reflect listener preferences.</td>
<td>Listeners can be asked to rate audio on a number of scales. Emiya et al. (2011) used four categories similar to those above.</td>
<td>Not a widely used method of assessment for BASS.</td>
</tr>
<tr>
<td>PEASS</td>
<td>Built on BSS Eval and able to allow the same distortions. Measurements are also perceptually relevant allowing reflection of a listener experience.</td>
<td>Rates the separation using four metrics: OPS, TPS, IPS and APS.</td>
<td>Used in two cross-institutional campaigns. (Vincent et al. (2009) and Araki et al. (2012))</td>
</tr>
</tbody>
</table>

Table 3.1: A comparison of BASS evaluation techniques.
Establishing the needs of the broadcasting industry is an important early aim of this research. The BBC is a major stakeholder in this project and presents an opportunity for any resulting technologies to be applied. This chapter will answer the third question of this thesis: “3. What are the BASS requirements of the broadcasting industry?”.

The work presented in this chapter is the result of a number of observations and conversations with BBC sound engineers during December 2011. In describing a number of situations where the BBC’s audio operations have been observed, insight will be provided into the needs of an un-mixing algorithm and the constraints which will need to be applied in its design.

In this chapter, three distinct operations—live broadcasting, event recording and audio editing—will be discussed before implications for an un-mixing tool are suggested. Focus is also given to noise control methods and the opinions of the sound supervisors.

The descriptions of the observations can be found in Section 4.1, specific details of noise control solutions are detailed in Section 4.2, a discussion of the implications of the observations for BASS in broadcasting are included in Section 4.3 and Section 4.4 details the opinions of the BBC sound supervisors on BASS.
Chapter 4. The requirements of the broadcasting industry

4.1 The observations

This subsection gives details of the observations of BBC procedures. In all cases, the information is drawn from time sitting behind a mixing desk watching members of BBC staff with responsibility for sound while they worked. A range of different tasks were observed; they can not be thought to be entirely representative of BBC working procedures as it is such a large organisation but they do offer an insight into at least some of the work the BBC does with audio and how it is carried out. A number of similarities were observed across the different settings and these similarities are the focus here.

4.1.1 Live broadcasting

Live broadcasting was observed in an outside broadcast van during a Radio 1 Xtra show. The outside broadcast van was parked behind a students’ union bar that was being used for the show. Audio was transmitted to the van in the multi-channel audio digital interface (MADI) format over optical fibre. Inside the van the audio was routed digitally and mixed using a control surface. Audio was sent to Broadcasting House, over an integrated services digital network (ISDN) connection, for transmission to listeners.

The programme was formed of a number of inputs for music to be played from as well as two hand-held radio microphones for the presenters and two pairs of audience microphones positioned at either side of the room. The mix operated very much in two modes: when music was playing all microphones were closed and when the presenter was talking the hand-held microphones were used in conjunction with audience microphones where appropriate. The mix was structured into three groups: presenters, music and audience. Heavy dynamic range compression was applied to the presenting microphones and to the output mix. This is the style of BBC Radio 1 and is appropriate for typical listening environments which is in a vehicle or at home listening while doing other tasks. While the broadcast was made in stereo, only the audience channels were panned off centre, again due to the station’s preferred style (music tracks retained their full stereo width).
The role of the sound supervisor in this setting is to mix the audio and also act as a point of connection between the producer in the venue and staff at Broadcasting House. When not mixing the show, they spent time making phone calls and using the internal communication system to relay information related to the broadcast.

Deploying an un-mixing tool in the live broadcasting context would have a number of requirements. As the sound supervisor has more responsibilities than creating the mix, an un-mixing tool must be quick to use so they are able to set it up and then carry on with other tasks. The audio to be un-mixed will contain speech and music, spatially indistinct sources and the reverberance of the room.

### 4.1.2 Event recording

Two event recordings were observed as part of this research, one for radio and the other for television. The radio recording involved a performance by the BBC Concert Orchestra while the television recording involved a quiz show. Both tasks display a range of differences, despite both being audio capturing exercises.

The television quiz show was captured in BBC Television Centre’s studio 4. The recording involved seven lapel radio microphones, two hand-held radio microphones and ten microphones over the audience. In addition to the microphone inputs there were audio inputs from the company running the quiz software and a studio sound effects desk. The engineer mixed the recording live and the output was sent to the recorder in a number of configurations that may be useful to the editor. At the editing stage there is still access to pre-fade recordings of each of the contestants’ microphones as well as a mix down of audience microphones and a mixed down effects channel. These are only for use in the case that there is a problem with the live mix, which ideally should be used in its entirety.

The engineer has use of a number of techniques to help prevent unwanted sounds being recorded to the master track. An IntelliMix system (described in Subsection 4.2.1) is used across all lapel microphones. A Cedar dialogue noise suppressor (DNS) system (described in Subsection 4.2.2) is also available for the
removal of wideband noise. In a television studio, the noise produced by the lighting equipment is audible over the microphones and these tools help control the noise. The Intellimix system is routed back to the desk on its own channel allowing the engineer to control one fader knowing it is the current speaker.

The BBC Concert Orchestra performance involved a whole orchestra covered by a Decca tree with additional spot microphones of various types on each desk (1–2 musicians). In addition to this a number of audience microphones were hung from the ceiling. This recording was made in London’s Mermaid Theatre. The audio was mixed in an outside broadcast van with the audio being mixed down to a stereo file on a flash card recorder.

Between the two recordings observed there were a number of similarities and differences noted. In radio the value placed on audio is higher, the producer was constantly checking that the sound supervisor in the van was happy throughout the rehearsal, whereas in television the sound crew are consulted less while extensive work is put into sorting out cameras and lighting.

There are also a number of similarities. In both situations the audio is mixed straight to a stereo file for later editing. While multi-tracking facilities are available they are not considered advantageous. There is clear evidence of the sound recordist’s maxim “don’t fix it in the mix” at work. The aim is to get the sound recording right in the live setting leaving minimal work to do during editing.

As with the outside broadcast, sound staff were found to be responsible for other technical equipment not directly related to the broadcast audio, this included the internal communication system.

Un-mixing the orchestral recording would present a number of technical challenges. The independence assumption of ICA would not be satisfied by instruments playing in harmony and the final stereo mix is under-determined. Whilst the musicians considered the Mermaid Theatre “pretty dead” there is still an amount of reverb present. The quiz show’s studio presented a lot of noise problems to the sound supervisor. These were mainly controlled by use of an IntelliMix system. The audio was made up of voices and quiz sound effects. The mix was non-stationary.
due to the regular change of speaker that is necessitated by the format of the programme.

4.1.3 Programme editing

Observing programme editing took place while a producer and sound engineer were putting together a documentary for BBC Radio 4. In advance of the session the producer had captured large amounts of source material with the presenter in Pakistan and then recorded links in a London studio. The producer had also collected a range of sound effects and positioned everything in the timeline of multi-track digital audio software. The team were editing to produce two versions of a programme: a half an hour episode for Radio 4 and a lengthened two-part series for BBC World Service.

The task for the editing session was to reduce the large pool of sound source files prepared by the producer to a stereo file of the correct length. This process involved constant reference to the documentary’s script. The main purpose is to work through the programme and ensure every sound starts and finishes correctly with appropriate automated crossfades. Certain pieces of audio require other improvements such as equalisation. The editing work is completed rapidly to ensure completion of the task during the allocated studio time.

For an un-mixing tool to be used in this situation it would need to be quick and powerful. A number of edits were observed that removed unwanted speech from recordings. The tool would need to be able to detect voices against background sounds.

4.2 Noise control

The way unwanted noise and background sounds were controlled was a matter of particular interest during the observations as an un-mixing solution will provide a noise and background sound control tool. This refers to both common white noise sources and incidental sounds which distract from the target audio.
4.2.1 Intellimix

During the television recording microphone signals were passed through Shure’s patented Intellimix system (Gilbert & Canfield 2007). The Intellimix unit makes use of automated audio gating to ensure only the microphone being addressed is open. Having several microphones open increases the build up of noise in the mix. The unit keeps at least one microphone open at all times to ensure the ambient sound of the room is never lost. While this tool is not useful in every circumstance it works well over the lapel microphones used on the quiz show presenter and contestants. By controlling which microphones are open, unwanted noise from the lighting system and other parts of the busy studio are less likely to be recorded.

While the Intellimix system can close microphones that are not in use it can not prevent microphone cross talk. In situations such as the quiz show where the contestants are spaced closely, sources are already spread across several microphones and the Intellimix system will simply select one of those microphones but not do anything to separate its contents.

4.2.2 Dialogue noise suppressor

This tool produced by Cedar Audio was used in the television studio and in the documentary editing process. Cedar’s DNS is one of a range of broadband noise reduction tools available in the audio industry and is used by the BBC for removing noise sources that are particularly white, for example air conditioning.

While the DNS performs a task that an un-mixing solution could potentially be used for, a fully functional un-mixer would provide far greater flexibility as it would be able to remove non-white sources.

The use of the DNS may have implications for an un-mixing tool. Audio that has passed through the DNS may have its spectral characteristics altered by the process which may then affect the performance of an un-mixing technology.
4.3 Implications for an un-mixing tool

The observations detailed in this section of the thesis have given a number of insights into the nature of an un-mixing tool for use in the broadcasting industry. These have come about from watching the engineers work and also studying the infrastructure of the audio systems.

4.3.1 The interface with the engineer

In live situations, the engineers are working under time pressure to ensure that they make their next cue. In the editing situation, the time pressure is instead to finish the mixing progress within the allotted time. The sound processing observed all took place very quickly with very little time being spent using a particular tool. This principle will need to apply to an un-mixing tool. The engineer will not have time to set and monitor parameters but will need the tool to provide a limited number of powerful controls allowing the engineer to continue with their other work.

4.3.2 Inputs and outputs

An implementation of an audio un-mixing solution will need to be designed with a model for inputs and outputs in mind. The situations observed with the BBC have given a number of insights into the likely configuration of an un-mixing problem. Generally foreground speech is the most important piece of programme content, conversely unwanted speech is a high priority for removal as it is particularly distracting to the listener. Any algorithm created for use in the broadcasting industry should expect speech to be part of an input signal and its removal or enhancement to be a requirement by the output stage.

In Chapter 2, it was seen that convolutive effects are a problem for sound un-mixing. Reverberation was noticeable in all content creation but least apparent during the outside broadcast due to the use of dynamic microphones at close proximity and the compressed nature of the signal. The mixtures taken in by an
un-mixing tool for broadcasting are likely to be convolutive. The nature of the recording technique also renders it non-stationary, due to the nature of the sound sources and technologies processing them.

Noise is not noticeable on any of the recordings but will be present in some quantity and should be accounted for in a model. A variety of microphone-to-source ratios were observed and if signals were to be un-mixed some problems would be under-determined.

4.4 Opinions of the sound supervisors

During the observations discussions were had with the engineers about the application of an audio un-mixing within their work. The engineers’ in depth knowledge of the technology and procedures provides insights into the way an un-mixing technology can be applied.

4.4.1 Data compression

A number of sound supervisors were swift to highlight the use of lossy data compression on their audio. In both the digital video broadcast (DVB) and digital audio broadcast (DAB) standards, which are used for digital television and radio respectively, MPEG encoding is used for the audio. Internal transmission of audio is also often bit rate reduced, the outside broadcast that was observed sent audio data over ISDN using the proprietary enhanced aptX (eaptX) codec with a redundant back-up channel transmitting AAC.

Lossy compression reduces the amount of data by discarding perceptually redundant data. Common examples are high frequency information and sounds that are masked by louder sounds. This could have consequences for an un-mixing technique that attempts to identify sources by their spectral envelopes and when trying to recover a masked source using unmixing.
4.4.2 The art of mixing

The sound supervisors that were questioned view their work as an art form and some expressed concern about any application that would lead to audience members having too much control over the sound. It is still undecided whether the eventual application of this work will give the audience control over the mix or will provide enhanced tools to content producers.

4.5 Chapter summary

In this chapter this thesis’ third research question, “3. What are the BASS requirements of the broadcasting industry?”, has been answered. This has been done using four sub-questions:

3.1 What is the role of the sound supervisor?
3.2 How is noise controlled?
3.3 What are the implications for an un-mixing tool?
3.4 What are the opinions of the sound supervisors?

Sound staff are often responsible for tasks additional to the mixing of the programme audio. These have been seen to include management of the communication system and relaying information between different people working on the programme.

Noise is controlled using systems including the Cedar DNS and Shure’s Intellimix. The Cedar system provides a method of removing background noise sources from a recording. The Shure system uses automatic gating to close microphones which are not being addressed.

The implications for an un-mixing tool are that it is required to be easy to operate with as little user interaction required as possible. Existing noise control techniques mean there may be spectral alterations to sounds in recordings and a non-stationary mixture.
Sound staff were supportive of the project but raised concerns about the use of compression on a lot of the audio being processed. They were also uncomfortable with the idea of the audience being given too much control over their mixes.

Answering these sub-questions has shown that the requirements of the broadcasting industry are a tool which can be deployed in non-stationary, reverberant environments with unknown numbers of sound sources. The tool will be operated by a busy sound supervisor who will need it to be quick to set up and use.
Chapter 5

Developing source separation for the broadcasting industry

In Chapter 2, a number of techniques that are currently being developed for BASS were detailed. Chapter 4 detailed observations of the BBC at work and suggested how a BASS technique may operate in industry. This chapter will bring together these two lines of work and make the case for the experimental work detailed in the next chapter. This will answer this thesis’s fourth question: “4. What should be the subject of further investigation?” This will be achieved by answering three sub-questions:

4.1 For broadcasting applications, which area of investigation looks most promising?

4.2 What are the current problems in this area?

4.3 How might these problems be addressed?

This chapter is structured in three sections. Firstly, in Section 5.1, the case for focusing on the TF mask is made based on the needs of broadcasting, its compatibility with the BASS approaches from Chapter 2 and the area for improvement it presents. Secondly, in Section 5.2, the results of Araki et al. (2012) are observed and the reason for the poor performance of the binary mask in that study is discussed. Finally, in Section 5.3, possible approaches to improving the performance of TF masking are detailed.
Chapter 5. Developing source separation for the broadcasting industry

5.1 The TF mask for broadcasting

This section will answer this chapter’s first sub-question: “For broadcasting applications, which area of investigation looks most promising?”. This question will be answered using observations of the broadcasting industry and the similarities in BASS techniques dealing with under-determined mixtures. In Subsection 5.1.1, the case is made for using a TF masking approach to separate audio mixtures encountered within the broadcasting industry. Subsection 5.1.2 focuses specifically on the binary TF mask.

5.1.1 The case for the TF mask

In Chapter 3, it was established that audio separation problems encountered in broadcasting are often under-determined. Adding more microphones to provide more observations of the mixture would allow the problems to become exactly determined but would involve large changes to current broadcasting practice. This would also involve the use of more equipment causing the cost of making content to increase. There would be no way of applying this technique to archive material.

The three BASS approaches detailed in Chapter 2 can each separate under-determined problems. Each approach achieves this by creating a sparse representation of the mixture. A sparse representation is one in which most of the elements are near zero.

As described by Plumbley et al. (2010), sounds are produced “either by resonant systems or by physical impacts, or both”. Resonant systems and physical impacts produce sounds which are sparse in frequency and time respectively. Therefore, a representation that reveals both time and frequency information is likely to exhibit sparseness in at least one of these dimensions.

Many TF transforms have been used for audio separation including those mentioned earlier in this thesis: the short-time Fourier transform and the gammatone filterbank, and others including the discrete cosine transform (DCT)
and wavelet based transforms.

From a sparse representation, ICA can cluster data together (Zibulevsky & Pearlmutter 2001), CASA can extract features to group the data and NMF can factorise the data into its coding and bases matrices. Most commonly, the results of the clustering, feature extraction or factorisation then inform a process whereby each element in the sparse representation is assigned a weight dependent on its likelihood of being part of the target rather than the interferer. The weighted sparse representation then steers resynthesis of the target signal. Further information on the calculation of TF masks in CASA, ICA and NMF can be found respectively in Wang (2005), Pedersen et al. (2008) and Grais & Erdogan (2011).

Thus, separation of an under-determined mixture, using ICA, CASA or NMF, will commonly rely on appropriate weighting of a time-frequency representation of the mixture. The weighting of a TF representation is referred to as TF masking. For the under-determined source separation problems typically encountered in the broadcast industry then, regardless of the chosen separation technique, TF masking is likely to play a key role.

5.1.2 The binary mask

The previous subsection found masking in the TF domain to provide a way of separating under-determined audio as required in many situations in broadcasting. This subsection focuses on current implementations of TF masking methods by discussing the justifications of the most prevalent implementation: the binary mask.

The binary mask is defined by Li & Wang (2009) as

\[
M_{ij}^{\text{IBM}} = \begin{cases} 
1 & \text{if } X_{ij} > Y_{ij} \\
0 & \text{otherwise}
\end{cases} \tag{5.1}
\]

where \(i\) and \(j\) are time and frequency indices and \(X\) and \(Y\) represent TF representations of the target and interferer respectively. This weighting retains the
entirety of TF cells which contain more target signal than interferer and rejects others.

The binary mask was proposed as the goal of CASA by Wang (2005). His paper gives two main reasons for the use of binary masks: firstly, consistency with models of human ASA, and secondly, consistency with the aims of source separation, speech recognition and noise reduction. The use of the binary mask extends beyond CASA techniques. Pedersen et al. (2005) make use of the technique with ICA and Grais & Erdogan (2011) calculate a range of TF masks including the binary mask with the results of an NMF algorithm.

Wang’s assertion, that the binary mask is consistent with models of human ASA, refers to the auditory masking phenomenon, which describes the process whereby a sound is “rendered inaudible by a louder sound within a critical band” (Wang & Brown 2006). The argument of consistency with the aims of a number of audio signal processing tasks is based on the fact that the binary mask is retaining target data, which is useful in these tasks, and discarding interfering data, which hinders them.

A mathematical approach to justifying the use of the binary mask is given by Li & Wang (2009). In the case where the TF representation of the signal is calculated with overlapping windows, they argue that the IBM is near optimal and simpler to calculate than the alternative ideal ratio mask (IRM). The IBM has also been shown to improve speech intelligibility in noise (Roman et al. 2003).

5.2 The performance of binary masking

The previous section made the case for TF masking to be used for BASS in the broadcasting industry and went on to highlight the prevalence of the binary TF mask within existing literature. This section will detail reported problems with the binary mask and then discuss whether each of the sections of a typical single-channel BASS system, represented in the block diagram in Figure 5.1, could be the source of the error. Identifying the source of the error will allow an area for improvement to be outlined.
Chapter 5. Developing source separation for the broadcasting industry

Figure 5.1: Overview of a typical under-determined BASS system. A TF representation is calculated before the target and interferer are estimated. Based on these estimates, the mixture is masked allowing the target estimate to be resynthesised.

TF masking has been noted to introduce artefacts into the separated output (Araki et al. 2005). These artefacts are often described as musical noise for the reasons explained in the opening chapter. In the work of Araki et al. (2012), source separation algorithms are compared using PEASS and use of the IBM is shown to produce audio of a low perceptual quality. Table 5.1 shows some of the results from this paper.

<table>
<thead>
<tr>
<th></th>
<th>2 mic, 3 speech</th>
<th>2 mic, 3 music</th>
<th>2 mic, 4 speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR ISR SIR SAR</td>
<td>SDR ISR SIR SAR</td>
<td>SDR ISR SIR SAR</td>
</tr>
<tr>
<td></td>
<td>OPS TPS IPS APS</td>
<td>OPS TPS IPS APS</td>
<td>OPS TPS IPS APS</td>
</tr>
<tr>
<td>O1</td>
<td>10.8 20.1 21.7 11.1</td>
<td>10.4 18.0 18.8 12.5</td>
<td>9.1 17.6 20.0 9.3</td>
</tr>
<tr>
<td></td>
<td>38.9 61.8 70.5 37.7</td>
<td>33.3 48.5 64.8 34.2</td>
<td>27.1 57.7 71.8 21.9</td>
</tr>
<tr>
<td>O2</td>
<td>8.5 15.7 17.4 9.1</td>
<td>9.0 14.1 18.1 11.3</td>
<td>7.5 13.7 16.4 8.1</td>
</tr>
<tr>
<td></td>
<td>24.0 29.8 72.4 20.0</td>
<td>30.4 28.3 69.5 21.6</td>
<td>22.0 20.9 70.8 13.1</td>
</tr>
</tbody>
</table>

Table 5.1: Results from the study by Araki et al. (2012). The results shown are for separated audio using the IBM calculated over the STFT (O1) and the cochleagram (O2). For each system, the BSS Eval results are shown in the top row and the PEASS results are shown on the bottom row. The scores for artefact related metrics are particularly low. All PEASS metrics are between 0 and 100 with higher scores always indicating better perceptual quality.

The results in Table 5.1 show the quality of audio separated by the IBM is low. The OPS results appear to be held back in quality by the prevalence of artefacts in the separated audio.
5.2.1 Representation

The discussion in Section 5.1 highlighted the fact that a sparse representation is required to solve the under-determined BASS problem. In Araki et al. (2012), two TF representations were used and, in both cases, artefacts were observed in the separated audio. To reduce the artefacts it is necessary to understand whether the TF representation is a source of artefacts.

Currently the most common CASA model makes use of a gammatone filterbank where the filters are spaced on the equivalent rectangular bandwidth (ERB) scale. This is done for reasons of similarity to the spacing and bandwidth of human auditory filters in the cochlea (Wang & Brown 2006). The STFT is also used for its computational efficiency.

Unlike the STFT, the gammatone filterbank does not allow perfect reconstruction of a signal. This increasingly becomes a problem with fewer filters in the bank. A perfectly reconstructable auditory filterbank may improve the audio output by the system.

Audition of the inversion of a 128-channel filterbank without masking suggests that there is an amount of improvement to be made here but that it is small in comparison to areas for improvement later in the system. PEASS analysis of audio which has been converted to a cochleagram and back again also suggests that the loss is minimal; an OPS of 98 was achieved on a speech sample.

5.2.2 Feature extraction

This subsection discusses the possibility that the artefacts in separated audio are due to a flaw in the way features are extracted for the calculation of the IBM. As described in Chapter 2, a number of different techniques can be used to find spatial, temporal, spectral, and statistical features in audio.

The results O1 and O2, cited from Araki et al. (2012), use ideal masks implying a perfect feature extraction at the given resolution. Figure 5.2 shows a comparison of the APS results from Table 5.1 for separation algorithms using ideal and non-ideal
Chapter 5. Developing source separation for the broadcasting industry

Figure 5.2: A box plot comparison of APS for the ideal and non-ideal techniques reported in Araki et al. (2012). The diamonds mark the means. The whiskers mark 1.5 inter-quartile ranges from the upper and lower quartiles.

feature extraction. The scores for the non-ideal feature extraction are spread over a greater range with some being higher than those using ideal feature extraction. This suggests that algorithms using ideal feature extraction will not necessarily generate audio with a greater APS than those using non-ideal feature extraction. Thus, it does not appear that attempts to increase APS by improving feature extraction would be likely to be fruitful.

5.2.3 Masking

The findings of the previous two subsections make it seem increasingly likely that the artefacts in the separated audio are a result of the process used for the masking.
Chapter 5. Developing source separation for the broadcasting industry

The IBM is popular in literature as it is believed to produce the best possible SNR (Wang 2005). This reasoning does not take into account the potential for artefact introduction into the separated audio.

Araki et al. (2012) argue that their study sets the upper bound for the performance of the binary masking based algorithms. This is only true for the specific filterbank used in this study but, as discussed above, audition of a 128-channel cochleagram without a mask suggests the performance is mainly impeded by the masking artefacts.

The usefulness of a TF mask for the broadcasting application of audio source separation, alongside the poor performances of systems using this method, makes the case for improving the binary masking function to produce fewer artefacts as the goal for further work in this project.

5.3 How can the binary mask be improved?

This chapter has so far argued that calculation of a TF mask can be useful when separating under-determined audio mixtures, as often encountered in the broadcasting industry, and that current separation systems using the binary TF mask produce low quality separated audio. The limiting factor in the performance of binary masking is the artefacts that are introduced to the audio by the process. This section will answer this chapter’s final sub-question: “How might these problems be addressed?” by discussing switching functions and ways of redistributing or smoothing the error from a binary mask. The aim is to determine if pursuing either of these ideas may provide better quality audio.

5.3.1 Alternatives to the binary switching function

The switching function defines the relationship between the target-to-mixture ratio (TMR) at a given TF cell and the value of the mask for that cell. To improve the separation performance of TF masking some authors have proposed alternatives to the binary switching function. Here those belonging to the IRM and the sigmoidal
mask are detailed. Both of these functions are shown in comparison with the switching function of the binary mask in Figure 5.3. These functions will be discussed here to establish whether there is evidence that a non-binary switching function may provide better quality audio than a binary one.

![Figure 5.3: The relationship between the TMR and mask weighting for the binary, ratio and sigmoidal switching functions.](image)

The ideal ratio mask

The IRM is a prominent alternative to the IBM. The mixture is masked using the exact ratio of the target to interferer. The advantage of this is that TF cells, which contain both target and interferer information, can have their energy content split appropriately rather than allocating it, in its entirety, to the stronger of the target and interferer. This in turn reduces the switching severity. The IRM is a similar concept to the Weiner filter as detailed by Loizou (2013).
Li & Wang (2009) investigated the IRM and found “on average, the IRM gives a better SNR gain than that of the IBM.” The paper then continues to make the case for use of the IBM as the IRM’s improvement is small and the IBM can be simpler to calculate. No account is given of differences in perceptual performance between the two masks.

The sigmoidal mask

The idea of using a sigmoidal switching function for a TF mask has been suggested by both Araki et al. (2006) and Grais & Erdogan (2011). The sigmoidal mask exists between the ratio mask and the binary mask in that it holds a linear relationship with the signal ratio for target-to-interferer ratios close to zero and is close to one or zero for extremely negative and positive values respectively. Grais & Erdogan (2011) compare a series of sigmoidal masks with both the IRM and IBM. Their results show at each SIR tested a sigmoidal mask produces the highest SDR. There is a suggestion in the data that masks that are more binary produce better results when the TMR is higher. As with the IRM, no perceptually related metrics are available for the sigmoidal mask.

5.3.2 Smoothing of the binary mask

Smoothing of the binary mask is motivated by the hypothesis that the artefacts in the audio from binary masking systems are caused by switching frequencies on and off instantaneously. To prevent this either the severity of the switching can be reduced or the effect can be smoothed over time.

Time smoothing techniques include those described by Araki et al. (2005) which details a method using a fine-shift overlap-add technique to generate the mask. Their technique uses shorter time steps in the re-synthesis of the audio than in the analysis. This has the effect of smoothing the mask through time. Madhu et al. (2008) details a cepstral smoothing algorithm for the reduction of artefacts in binary masking systems. This transforms the binary mask into the cepstral domain and smooths each quefrency channel. The smoothed mask is then
recovered by inverting the cepstral transform. The process of dithering, where noise is added to a signal before quantisation, could also be of use.

5.4 Chapter summary

This chapter aimed to answer this thesis’s fourth research question: “4. What should be the subject of further investigation?”. The findings of Chapter 4 indicate that a tool for use in broadcasting must be able to separate under-determined mixtures. The BASS techniques discussed in this thesis can all do this provided that a sparse representation of the mixture audio is available. This is often achieved using TF representations which can then be masked. The IBM is a particularly common goal for source separation systems as it provides good objective performance.

Use of the binary mask is problematic in that it has been shown to introduce artefacts into separated audio. This was demonstrated in the results of Araki et al. (2012) where systems using binary masks were observed to have low APS scores. Looking at the details of the experiment has revealed that the low scores can not be the result of the representation or feature extraction methods used. The mask used is the most likely cause of artefacts and should therefore be the focus of further work in this project.

Evidence from Madhu et al. (2008) suggests that an alternative masking methods may lessen the masking artefacts. Further work will be performed to test alternative masks to find which most improves the quality of the separated audio.
Chapter 6

Perceptual quality improvement of time-frequency masks

Earlier in this thesis it was established that the binary mask can be used as a part of a number of different BASS techniques. Experimental work using the binary mask has shown that separated audio is laden with artefacts. Araki et al.’s (2012) investigation of the quality of audio separated by binary masks suggests that the binary mask may be inferior to other separation techniques. In particular, the work of Ozerov et al. (2012) which uses Wiener filtering, a technique closely related to ratio masking (Li & Wang 2009), is shown to be superior at a task separating under-determined non-convolutive mixtures.

The aim of this chapter is to answer this thesis’s fifth question: “5. Can binary masking performance be improved?”. This chapter will achieve this aim by detailing experimental work using a number of variations of the binary mask to perform audio separation on a set of synthetic mixtures. Any improvement given by a mask will be quantified so that they may be compared in terms of how much improvement they provide.

The remainder of this chapter is structured as follows: firstly, in Section 6.1, the experimental method is described; then, in Section 6.2, the experimental masks are explained; thirdly, in Section 6.3, the experimental masks are compared; and, finally, a summary and conclusions are given. The work in this chapter has previously been published as Stokes et al. (2013).
6.1 Method

To compare the separation performance of the IBM with other TF masks a number of audio mixtures were separated and the quality of separation was assessed. In Chapter 3, PEASS was found to be the best tool for quantifying the quality of separation and this is the technique that will be used for this experiment.

The experimental procedure was as follows:

- take target and interferer audio files, decimate to 24 kHz, normalise peak level to -3 dBFS and edit length to 10 s (240 × 10^3 samples);
- create mixtures of each permutation of target and interferer, applying unity gain to both;
- calculate masks and separate audio; and
- calculate PEASS metrics.

This section will now discuss the audio material used and the use of the PEASS metrics. The experimental masks used are detailed in Section 6.2.

6.1.1 Audio material

This experiment used six target sounds and six interferer sounds. Each permutation of target and interferer was used creating 36 different mixtures to be separated. Each audio extract is 10 s in length and decimated to a sampling rate of 24 kHz to reduce the computational load. This was deemed preferable to a shorter sample at higher sampling frequency as that would have given data representing a shorter period of time which means the editing of the audio has a greater effect on how it is perceived. A normalisation to -3 dBFS was applied to all files.
Target audio

The target audio was taken from two recordings of BBC Radio 4’s ‘The Bottom Line’. This audio was recorded in London during a visit to Broadcasting House and provides genuine broadcast material for use in the experiment.

Each extract featured a single speaker. Three were from male speakers and three from female speakers. These were taken from the studio audio and selected to cover exactly ten seconds without the start or end splitting a word.

Interferer audio

Interferers were taken from a number of different sources to try and reflect some of the range of interferers that may need separating. Two speech extracts were taken from the European Broadcasting Union (EBU)’s Sound Quality Assessment Material (SQAM) resource. Two musical interferers, also from SQAM, were used: one was a piece of pop music and the other a piece of violin music. Environmental noise was taken from the computational hearing in multisource environments (CHiME) resource (Barker et al. 2013). The CHiME corpus provides noise recorded in “real multisource noise” environments.

6.1.2 PEASS

PEASS provides four metrics for rating separated audio. Each of the metrics corresponds to a particular attribute of the separated audio on a scale of 0, the lowest quality, to 100, the highest quality. To establish whether the level of artefacts was reduced the APS was used. To understand whether any change in APS had affected the suppression of the interferer the IPS was used. The OPS was also used to demonstrate how the interaction between the two previous metrics relates to overall perceived quality.
Chapter 6. Perceptual quality improvement of time-frequency masks

6.2 The experimental masks

This experiment compared a number of mask improvement techniques to the IBM. The techniques tested were a mixture of novel techniques and suggestions from literature. This section will detail each technique and how it was implemented for this experiment. Each modified mask aims to reduce the artefacts in the separated audio by either adding noise to mask, smoothing the mask or removing small segments of the mask. The techniques tested were:

1. the ideal binary mask (IBM);
2. the dithered binary mask (DBM);
3. the noisy binary mask (NBM);
4. the cepstrally-smoothed binary mask (CBM); and,
5. the segmented binary mask (SBM).

For the purpose of visual comparison, the masks for a section of one specific mixture are shown in Figure 6.1.

6.2.1 TF representation

All the masks used are calculated from TF representations of the target and interferer. This TF representation was calculated using a fourth-order gammatone filterbank implementation from Ohio State University\(^1\) as detailed in Wang & Brown (2006).

The gammatone filterbank representation was used to create a cochleagram from 128 filters spaced on the ERB scale between 50 Hz and 12 kHz. This provided coverage of all frequencies up to the Nyquist limit. All other values were set to defaults; the window length was 320 samples.

\(^1\)available from http://www.cse.ohio-state.edu/pnl/software.html
Figure 6.1: The TF masks used in the experiment. They are shown here as calculated for one of the experimental mixtures.
Chapter 6. Perceptual quality improvement of time-frequency masks

6.2.2 The ideal binary mask

As described earlier in this thesis the IBM uses the ratio of target to interferer energy in a TF cell to assign it as part of the target signal or part of the interference. For this experiment the mask was calculated using a 0 dB threshold. This leaves the mask’s formulation as simply

\[
M_{ij}^{IBM} = \begin{cases} 
1 & \text{if } X_{ij} > Y_{ij} \\
0 & \text{Otherwise}
\end{cases}
\]  

(6.1)

where \(i\) and \(j\) are the time and frequency indexes respectively. \(X_{ij}\) and \(Y_{ij}\) are the values of the target and interferer cochleagrams at \(ij\).

6.2.3 The dithered binary mask

Dither is applied in systems where quantisation is about to occur. Either as a result of analogue to digital conversion or in a digital system where bit resolution is to be lost. Dither involves adding noise to mitigate the effects of quantisation distortion.

The dither can also be applied when calculating the binary mask. Similar to the calculation shown in Equation 6.1, the Dithered Binary Mask can be calculated as

\[
M_{ij}^{DBM} = \begin{cases} 
1 & \text{if } X_{ij} + \Delta > Y_{ij} \\
0 & \text{Otherwise}
\end{cases}
\]  

(6.2)

where \(\Delta\) is the triangularly-distributed dither noise and all other symbols retain their meanings from Equation 6.1. The mean of the triangular distribution was zero and the variance was chosen as described in the optimisation stage.

Optimisation

In order to find the optimal performance of the dithering technique the mask was calculated with an increasing noise ranges. Looking at the change in PEASS results allows the noise range which gives the maximum OPS value to be found...
and also provides more information about the effect of a dithered mask on the separation.

Figure 6.2 shows the PEASS metrics changing as the noise increases. The vertical dashed line marks the optimum OPS at 36 on the perceptual scale with noise of range 0.6 added. This value equates to 0.8 standard deviations of the TF target signals.

![Graph showing the change in PEASS metrics](image)

**Figure 6.2**: The change in the DBM’s APS, IPS and OPS as noise of increasing variance added. The dashed line marks the optimal OPS recorded.

6.2.4 The noisy binary mask

The NBM takes a binary mask and adds an amount of noise to it. This results in a mask with points triangularly distributed around 0 and 1 but not constrained to these values. By diminishing the severity of the step when the mask switches artefact salience may be reduced. The triangular noise distribution was used to
allow comparison with the DBM where the same noise is applied before each TF is compared to the threshold. The NBM is formulated as

\[
M_{ij}^{\text{NBM}} = \begin{cases} 
1 + \Delta & \text{if } X_{ij} > Y_{ij} \\
0 + \Delta & \text{Otherwise}
\end{cases}
\]  

(6.3)

Optimisation

The procedure for optimising the NBM was identical to the optimisation of the DBM. The noise range was allowed to increase from 0 and the PEASS results were recorded at each value.

Figure 6.3 shows the results of the optimisation. The dotted vertical line marks the optimum OPS value at noise range of 0.5 giving a mean OPS of 49.

6.2.5 The cepstrally-smoothed binary mask

Cepstral analysis allows a signal to be viewed as the spectrum of its frequency-domain representation. Artefacts can be removed in the cepstral domain as they are temporally short and occur at apparently random frequencies.

The method proposed by Madhu et al. (2008) will be used in this experiment. Madhu et al. divided the cepstrum into three regions to be processed differently. The low index queuefreqency bins define the spectral envelope. To avoid distortion of the spectral envelope the low index bins are given little or no smoothing (Jan et al. (2011) take this as 1000th of the sampling frequency). The bin that relates to the harmonic structure of a given spectrum is also smoothed lightly in comparison with the rest of the spectrum. Bins that are not related to the spectral envelope or harmonic structure are smoothed more. This allows smoothing of frequencies associated with musical noise without risking distortion of important features.
Chapter 6. Perceptual quality improvement of time-frequency masks

Figure 6.3: The change in the NBM’s APS, IPS and OPS as noise of increasing variance added. The dashed line marks the optimal OPS recorded.

The IBM is transformed to the cepstral mask, $M_{\text{cepst}}$, by

$$M_{\text{cepst}}^{i,j} = \text{DFT}^{-1}\{\ln(M_{i,j}^{\text{IBM}})\} \quad (6.4)$$

where $i$ is the time index, $j$ the frequency index and $l$ the quefrency index. The smoothing is then performed by

$$M_{\text{cepst}}^{i,j} = \gamma_l M_{\text{cepst}}^{i,j} + (1 - \gamma_l) M_{i,l}^{\text{cepst}} \quad (6.5)$$

with $\gamma_l$ the smoothing parameter in the $l$th quefrency bin. After the mask has been smoothed it is transformed back into the spectral domain by

$$M_{i,j}^{\text{CBM}} = \exp(\text{DFT}\{M_{i,l}^{\text{cepst}}\}) \quad (6.6)$$
In order to calculate $M_{CBM}$, zeros in $M_{IBM}$ must be replaced with near zero values to prevent $\ln(0)$ being used in the mask. In this experiment the value of 0.1 is used; this is consistent with the work of Madhu et al. (2008). $\gamma_l$ is the smoothing parameter its value is defined based on the value of $l$, according to

$$
\gamma_l = \begin{cases} 
\gamma_{env} & \text{if } l \in \{0, \ldots, l_{env}\}, \\
\gamma_{pitch} & \text{if } l = l_{pitch} \\
\gamma_{peak} & \text{if } l \in \{(l_{env} + 1), \ldots, K\}\backslash l_{pitch}
\end{cases}
$$

(6.7)

Optimisation

The CBM was optimised in terms of its three smoothing parameters. Allowing the values $\gamma_{env}$, $\gamma_{pitch}$ and $\gamma_{peak}$, to vary between 0 and 1, and enforcing the constraint: $\gamma_{env} \leq \gamma_{pitch} \leq \gamma_{peak}$. Optimal values of smoothing are found and information is gained about the effect of smoothing the different sections.

Figures 6.4, 6.5 and 6.6 show the variation of PEASS metrics with each smoothing parameter. Figure 6.4 shows that the APS is maximised by maximum smoothing of all three regions. Of the three smoothing coefficients $\gamma_{peak}$ appears to account for the most improvement in APS. This is probably due to it covering the largest part of the cepstrum.

Figure 6.5 shows that increasing the smoothing parameters decreases the IPS. Again the $\gamma_{peak}$ parameter appears to be the strongest factor in determining the IPS.

The OPS scores are shown in Figure 6.6. This metric again appears to decrease with increase smoothing. This is presumably driven by the decrease in IPS. The optimum value for OPS is 49, this value is obtained when all three smoothing parameters are equal to zero. Unlike optimisations described for previous techniques the zero case is not equivalent to the IBM. This is because even in the zero case the cepstral transform is still applied to the mask for which the lowest level in the mask is adjusted to 0.1.
Figure 6.4: Showing the APS obtained for the optimisation of the CBM. The $\gamma$ values were allowed to vary according to constraint $\gamma_{\text{env}} \leq \gamma_{\text{pitch}} \leq \gamma_{\text{peak}}$. Each mesh links data that share a common $\gamma_{\text{peak}}$ value. The largest mesh represents $\gamma_{\text{peak}} = 1$ and the smallest $\gamma_{\text{peak}} = 0$. 

92
Figure 6.5: Showing the IPS obtained for the optimisation of the CBM. The $\gamma$ values were allowed to vary according to constraint $\gamma_{\text{env}} \leq \gamma_{\text{pitch}} \leq \gamma_{\text{peak}}$. Each mesh links data that share a common $\gamma_{\text{peak}}$ value. The largest mesh represents $\gamma_{\text{peak}} = 1$ and the smallest $\gamma_{\text{peak}} = 0$. 
Figure 6.6: Showing the OPS obtained for the optimisation of the CBM. The $\gamma$ values were allowed to vary according to constraint $\gamma_{\text{env}} \leq \gamma_{\text{pitch}} \leq \gamma_{\text{peak}}$. Each mesh links data that share a common $\gamma_{\text{peak}}$ value. The largest mesh represents $\gamma_{\text{peak}} = 1$ and the smallest $\gamma_{\text{peak}} = 0$. 
6.2.6 The segmented binary mask

Image segmentation is the process of separating an image made up of pixels into sub images. In the case of a binary mask this can be performed by grouping connected pixels which share the same value. The size of these groups can then be used to remove those falling below a certain threshold from the mask by inverting them. A similar approach has been applied to speech perception in Cooke (2006).

Optimisation

The optimisation was performed by varying the threshold of the technique. This did not provide any variation in any of the perceptual metrics. Thresholds between zero and forty were evaluated but provided no variation in OPS when rounded to integer precision. The values taken through to the comparison stage are identical to all values in the range.

6.3 Comparison

The techniques will now be compared while configured with the optimal parameters discovered in Section 6.2. This comparison is shown in Figure 6.7. The bar chart shows the mean performance across all mixtures of the optimised techniques.

The results show that the dithered, noisy and cepstral masks provide measurable improvement from the IBM, in both the APS and consequently the OPS. These three techniques also reduce the IPS from the result obtained using the IBM.

6.3.1 Discussion

A number of matters arise from this experiment which will be discussed here and in some cases will form a basis for further work.
Chapter 6. Perceptual quality improvement of time-frequency masks

![Perceptual Score](image)

**Figure 6.7:** Each optimised technique’s mean PEASS scores for artefacts, interference and overall, across the mixtures.

**Binary and continuous masks**

The two masks with the highest OPS scores are both continuous. The other three masks in the comparison are binary. This suggests that better performance may be possible using continuous masks. The greatest improvement is 31 points on the OPS scale with the NBM.

This may be because little improvement is possible while maintaining a binary mask or it may be due to this particular processing of the mask.

**Cepstral smoothing**

The optimisation of the Cepstral Smoothing technique found the best performance when no smoothing was applied. The technique did however improve the
performance significantly from the IBM. This may be due to the changing of mask values of 0 to 0.1 to allow the cepstral transform to be made.

Transitions

The results appear to suggest the artefacts are the results of simultaneous transitions of multiple frequency bands from on to off. This is improved slightly by the dithering process as it randomises the transition to some extent. The cepstral and noisy masks reduce the severity of this transition which reduces the prevalence of the artefacts. Future work may focus on softening these transitions further.

6.4 Chapter summary

This chapter aimed to answer the question: “5. Can binary masking performance be improved?”. Experimental work detailed in this chapter has used the separation of 36 different mixtures as a means for comparison of TF mask improvement techniques. PEASS metrics were used to quantify the artefacts, interferer suppression and overall quality of the separated audio. An improvement of 31 points on the OPS scale has been recorded.

In all five masks have been tested and compared. These were:

- the IBM, which provided a mean OPS of 18;
- the DBM, which applied noise to the signal before calculating the binary mask and gave a mean OPS of 36;
- the NBM, which added noise to the binary mask and gave a mean OPS of 49;
- the CBM, which smoothed the mask in the cepstral domain and gave a mean OPS of 49; and
- the SBM which generated a mean OPS of 19.
This experiment suggests that continuous masks (NBM, CBM) produce better separated audio than binary ones (IBM, DBM, SBM). The unexpected result of the CBM optimisation suggests that it is not the smoothing process which provides the improvement but another part of the CBM process. The SBM shows no improvement over the range of values tested. Further work is required to establish the best way of limiting the transitions in the mask so they do not provide such severe switching but maintain optimal interferer suppression.
Chapter 7

Ideal sigmoidal masking

In a quest to improve the quality of audio separated by a binary TF mask, the previous chapter performed a number of modifications to the IBM. While various approaches led to an improved OPS, those that performed best were those that returned TF masks that after processing were no longer binary. The IRM, a continuous mask, is a commonly proposed alternative to the IBM. Work by Grais & Erdogan (2011) suggests a sigmoidal TF mask may be superior to both.

The aim of this chapter is to answer the question: “6. Which sigmoidal TF mask provides optimal separated audio quality?” This question will be answered by using a range of TF masks generated using different sigmoidal functions, to separate a corpus of mixtures, and using PEASS to measure the quality of the separated audio. Grais & Erdogan’s work shows that the binary and ratio mask can be seen as specific instances of sigmoidal masks. This information is used to develop a series of masks that vary in the way they use the ratio of target-to-mixture information.

The work in this chapter provides insight into two unexplored areas of the relationship between the binary and ratio masks. Firstly, the binary and ratio masks are treated as points on a continuum rather than a simple dichotomy. Secondly, whereas previous studies have focused on speech recognition (Srinivasan et al. 2006) or SNR (Li & Wang 2009), this study measures the perceived quality of the separated audio using PEASS.

A brief background on sigmoidal masking is given in Section 7.1. The experimental method is described in Section 7.2. Results are given in Section 7.3. A repetition of
the experiment at multiple TF resolutions is included in Section 7.4. The summary and conclusion of the chapter can be found in Section 7.5. The work in this chapter has previously been published as Stokes et al. (2014).

7.1 Background

The most common form of continuous TF mask is the ratio mask. In a given TF cell, the ideal ratio mask is defined by Srinivasan et al. (2006) as

\[ M_{ij}^{IRM} = \frac{X_{ij}}{X_{ij} + Y_{ij}} \]  

(7.1)

where \( X \) is the target energy and \( Y \) is the interferer energy.

Grais & Erdogan (2011) define the sigmoidal TF mask as

\[ M_{ij}^{ISM} = \frac{X_{ij}^p}{X_{ij}^p + Y_{ij}^p} \]  

(7.2)

where the parameter, \( p \), is used to control the linearity of the mask. Grais & Erdogan vary \( p \) between one and five, in this chapter \( p \) is varied between \( \frac{1}{32} \) and 32 with values evenly spaced on a series of powers of 2. This changes the amount of discrimination in the mask to vary in such way as to demonstrate the following properties:

- as \( p \to \infty \) the mask becomes binary;
- at \( p = 1 \) the mask is a ratio mask; and
- \( p = 0 \) produces a mask that is 0.5 at all values.

The range of switching functions used for the experiment in this chapter is shown in Figure 7.1.
Chapter 7. Ideal sigmoidal masking

Figure 7.1: The proposed range of sigmoidal switching functions to be used in this experiment. The approximation of the binary switching function is at $p = 2^5$ and the ratio switching function is the straight line through the middle of the set of sigmoids.

7.2 Method

This experiment was completed in three steps: firstly, audio mixtures were created and their TF overlap calculated to create a corpus containing an evenly distributed range of overlaps to ensure the difficulty of separation varied; secondly, for each mixture, for each of the sigmoidal switching functions an ideal mask was calculated using the known pre-mixture audio and each of the sigmoidal switching functions; and, finally, the audio was separated and then evaluated using the PEASS metrics. The software implementation was the same as that used for Chapter 6. This section describes the processes involved in each of the three steps.
7.2.1 Audio mixtures and overlap calculation

Twenty-two audio mixtures were generated from 10 second files with a sample rate of 24 kHz. Target signals were speech from a radio broadcast and SQAM\(^1\) (European Broadcasting Union 2008). Interferer signals were a range of background sound effects and ecological noise from the CHiME corpus (Barker et al. 2013). Twenty-two combinations of target and interferer were selected from a larger pool of mixtures according to how much they were deemed to overlap in the TF domain.

The TF overlap was measured using analysis of the histogram of the IRM. The ratio mask gives a good indication of overlap; in each element the extreme ratios, zero and one, indicate that there is no overlap between the sources whereas the central ratio, 0.5, indicates that the sources are entirely overlapping. This idea is the basis of the overlap metric used in this chapter.

To calculate the overlap, an 8192-point short-time Fourier transform was performed on the target and interferer signals. The IRM of these signals was then calculated and the following process was performed for the TF locations where the target signal exceeded the -96 dBFS noise floor of the 16-bit signals.

Firstly, the IRM was calculated as in Equation 7.1, then an eleven-bin histogram was calculated from the elements of \( M^{IRM} \),

\[
\mathbf{h} = \text{hist}_{11}(M^{IRM})
\]  

(7.3)

Next, \( \mathbf{h} \) was weighted in proportion to the amount of overlap represented by each bin. The sixth (middle) bin of the histogram contains IRM elements with values near 0.5, the maximum overlap; this was weighted at one. Either side of this mid point the weighting decreased linearly and symmetrically until reaching zero at bins one and eleven,

\[
\mathbf{w} = [0, 0.2 \ldots 1, 0.8 \ldots 0]
\]  

(7.4)

Finally, the weighted histogram was summed and divided by \( n \), the number of

\(^1\)http://tech.ebu.ch/publications/sqamed
target elements exceeding the 16-bit noise floor, to produce the final measurement,

\[ o = \frac{1}{n} \sum_{i=1}^{11} h_i w_i \]  

(7.5)

The above process measures the centralness of the histogram of the ratio mask. While this could also have been achieved using kurtosis the method employed had two distinct advantages. Firstly, kurtosis is calculated about the mean of the data; this means that two histograms with different means but similar shapes would have had similar overlap scores. The weighting of the metric used in this study was centred about 0.5 ensuring only the most severe overlap received the highest rating. Secondly, it has been noted that “kurtosis for bimodal distributions is not necessarily negative” (DeCarlo 1997). A histogram which represents little or no overlap is bimodal and a kurtosis based metric would have been difficult to interpret.

### 7.2.2 Estimates

Estimates of the target audio were generated in three steps: firstly, for the target, interferer and mixture a cochleagram was created. Secondly, the target and interferer cochleagrams were used to create the sigmoidal mask according to Equation 7.2. Finally, the mask was applied to the mixture cochleagram to obtain the estimate of the target audio.

The cochleagrams were generated using the process in Wang & Brown (2006). Each cochleagram was made using a bank of 128 fourth-order gammatone filters spaced on the equivalent rectangular bandwidth scale up to 12 kHz, the Nyquist frequency. The cochleagram was then generated from the gammatone filterbank using a rectangular window of length 320 ms and a 50% overlap.

To generate the range of sigmoidal masks. The \( p \) value in Equation 7.2 was varied to produce the series of different functions. Initially, 11 masks were used with the value of \( p \) scaled exponentially such that it takes values from the series of powers of two in the range \( 2^{-5} \) and \( 2^{5} \). These sigmoids are shown in Figure 7.1.
Each of the chosen mixtures was separated using ideal sigmoidal masks at each $p$ value. To allow comparison of the switching functions, analysis of the results was performed across mixtures at each $p$ value.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figures/figure7_2.png}
\caption{The mean TPS, IPS, APS and OPS values calculated over the 22 audio mixtures for different $p$ values. Dashed lines show 95% confidence intervals of the means and crosses mark the sample means. The results at $2^0$ represent a ratio mask and the results at $2^5$ approximate a binary mask.}
\end{figure}

### 7.3 Results

The separated audio obtained in the previous section was analysed using the PEASS toolbox. The results were collated across the range of $p$ values used so
Chapter 7. Ideal sigmoidal masking

Figure 7.3: The mean ISR, SIR, SAR and SDR values calculated over the 22 audio mixtures for different $p$ values. Dashed lines show 95% confidence intervals of the means and crosses mark the sample means. The results at $2^0$ represent a ratio mask and the results at $2^5$ approximate a binary mask.

the mean effect of changing the sigmoid could be analysed. This section will discuss PEASS results, BSS Eval results, and the relationship between the TF overlap and the results observed.

7.3.1 PEASS results

Figure 7.2 shows the mean and 95% confidence intervals for each PEASS metric. The results generally show expected behaviour: changing the switching function changed the amount of discrimination between TF elements, based
on the proportion of target energy they contained, and led to a worsening of artefacts, when a large amount of discrimination was applied, and to low interferer suppression, for low amounts of discrimination. This trade-off led to little variation in the OPS across large parts of the sigmoid range.

The area where the plot does not obey the artefacts-interference trade-off is around the \( p \) value \( 2^{-1} \). At this point the IPS was higher than at any of the values when \( p \geq 2^0 \). This, combined with the APS being above the low values it took when \( p \geq 2^0 \), gave a strong peak in the OPS scores. Due to the interesting results in this region, further results were generated at \( p \) values in intervals of \( 2^{\pm} \) between \( 2^{-3} \) and \( 2^0 \). The maximum improvement in the OPS, recorded at \( p = 2^{-4} \), was a
Chapter 7. Ideal sigmoidal masking

full 38 points over the IRM \((p = 2^9)\) and 49 points over the IBM (approximated at \(p = 2^5\)).

7.3.2 BSS Eval results

The BSS Eval metrics, shown in Figure 7.3, gave similar results with the SIR plateauing near \(p = 2^{-1}\). The highest SAR value, while the SIR was unchanging, occurred at \(2^{-1}\) giving the optimised point for the artefact-interferer trade-off.

7.3.3 The relationship between TF overlap and optimal sigmoid

Figure 7.4 shows the effect of TF Overlap on the location and height of the peak value for each mixtures at all \(p\) values. The correlation between overlap and the peak value is negative: as the overlap increased the peak OPS decreased. The correlation coefficient has been measured as -0.69. This is expected behaviour as mixtures displaying more overlap are likely to be harder to separate. There is little effect on the location of the peak: as the amount of overlap changed the peak value remained centred around \(p = 2^{-1}\). The correlation is lower; the coefficient is only -0.49. This may change at higher overlaps but further data would be required to investigate this.

7.4 Resolution

This section seeks to determine whether the results obtained previously are an effect of the TF resolution used. The artefact-interferer trade-off identified in both this work and previous studies may be affected by the TF resolution used for the analysis. The amount of switching that takes place is directly related to the window length as shorter windows entail more transitions. Conversely, the suppression of the interferer is reliant on short windows to allow localised attenuation of interfering sources.
To determine if the resolution has an effect, the OPS was measured at sixteen different TF resolutions using four of the sigmoidal values from the previous section: the flat, optimal, ratio and approximately-binary masks, found at the sigmoidal $p$ values of two to the powers of $-5, -\frac{4}{3}, 0$ and $5$ respectively. The time resolution was altered by changing the window length and the frequency resolution was changed by varying the number of gammatone filters. The number of filters was changed between 32, 64, 128 and 256, and the window length took the values 80, 160, 320 and 640 samples. The results of this work are shown in Figure 7.5. This shows that there is little variation in OPS due to the change in time or frequency resolution and that the optimum found in the previous stage of this study is still optimal at all resolutions tested. The sigmoid defined by $p = 2^{-\frac{4}{3}}$ gave the peak OPS value of 70 at all frequency resolutions and similar values were obtained across the range of window lengths.

7.5 Chapter summary

The aim of this chapter was to answer the question: “6. Which sigmoidal TF mask provides optimal separated audio quality?” To answer this question, a series of sigmoidal switching functions was defined and used to separate a corpus of audio. The PEASS toolkit showed that a sigmoidal switching function with $p = 2^{-\frac{4}{3}}$ provides the optimal OPS score, 70, for the series of sigmoids.

The results show a trade-off between artefacts and interference; the APS is highest when the mask is least varying and the IPS is higher when the mask is more varying. The point where this trade-off is optimised gives the peak OPS recorded.

The TF overlap of a mixture determines the best possible score that can be achieved for a given mixture. Mixtures with a higher degree of overlap have lower peak OPS while mixtures that overlap less give higher peak OPS scores. However, the amount of TF overlap appears to be unrelated to the optimal sigmoid for separating a given mixture.

The TF resolution that the TF mask has been calculated at does not change the
optimal sigmoidal switching function. The original setup, which uses 320 sample windows of a 128-channel gammatone filterbank’s output, produces the highest OPS of the resolutions tested.
Chapter 8

Non-ideal sigmoidal masking

The improvement in OPS for audio separated using a sigmoidal TF mask with $p = -\frac{4}{3}$ rather than the ratio or binary masks has so far only been demonstrated with ideal TF estimates of the target and interferer signals. This work will now measure the OPS for audio separated by the range of sigmoidal masks when the TF estimates of target and interferer signal have been provided by a state-of-the-art estimation algorithm.

The aim of this chapter is to answer the question “7. Which sigmoidal mask provides optimum quality under non-ideal conditions?” This question will be answered by repeating the experimental work in the previous chapter but using a state-of-the-art NMF algorithm to estimate the time-frequency representations of the target and interferer signals.

This chapter’s question must be answered as real-world unmixing problems do not have access to the known target and interferer signals used in the previous experiments in this thesis. Non-ideal target and interferer estimates may change the optimal switching function. For example, if the algorithm underestimates the amount of interferer then the point at which the artefact-interferer trade-off is optimised on the sigmoid scale might move. Chapter 7 shows the quality of audio that could be expected from an ideal TF mask estimation algorithm whereas this chapter provides a study of the separated audio quality that can be obtained using one of the best mask estimation algorithms currently available. The implications of the previous work on ideal sigmoidal masks for real-world algorithms will be better understood as a result of the study in this chapter.
The remainder of the chapter is structured as follows: Section 8.1 describes the state-of-the-art NMF algorithm used to generate target and interferer estimates; Section 8.2 explains the experimental procedure that was used; and Section 8.3 lists and discusses the results. The chapter is summarised in Section 8.4.

8.1 Gao & Woo’s algorithm

Chapter 2 established that a TF mask can be estimated by ICA, NMF or CASA. The TF estimation algorithm being used for this experiment takes the NMF approach described by Gao & Woo (2014)\(^1\). Of the two approaches in Gao & Woo’s paper, the work here uses the Quasi-EM NMF-2D algorithm. Gao & Woo’s approach is suited to this experiment due to three similarities with work in this thesis:

1. it is designed and tested on gammatone filterbank representations;
2. the method is aimed exclusively at the single-channel source separation problem; and,
3. the algorithm is designed for use when there is no prior knowledge of the sources.

The work of Gao & Woo extends the conventional NMF model,

\[ \|Y\|^2 \approx DH, \quad (8.1) \]

to a two-dimensional, convolutional one containing a time shift, \(\tau\), and a frequency shift, \(\phi\),

\[ \|Y\|^2 \approx \sum_{\tau,\phi} D^{\tau,\phi} H^{\phi}, \quad (8.2) \]

where \(D\) is an \(f \times i\) matrix and \(H\) is \(i \times t\), where \(i\) is the number of bases and \(f\) and \(t\) are the size of the signal in frequency and time respectively.

\(^1\)Matlab code for this algorithm is available at http://staff.uestc.edu.cn/gaobin/?page_id=5

111
Gao & Woo’s motivation in pursuing this approach is that it can model non-stationary sources using far fewer bases than a classic NMF algorithm. The use of fewer spectral bases also reduces the need for a clustering algorithm to group all the parts separated out by the NMF algorithm into coherent sources.

8.2 Experimental procedure

The experiment was designed to be as similar to that described previously in Subsection 7.2.2 (page 103) as possible. This was done to ensure any change in the results observed was due to the mask estimation method and no other factor. The experiment made use of the same mixture corpus and the same set of sigmoid functions. The only change was the use of Gao & Woo’s algorithm to provide estimates of the target and interferer signals.

Gao & Woo’s algorithm was applied to each of the twenty-two mixtures in the corpus, with the maximum number of iterations set at 50. As suggested in the authors’ code, the $\tau$ shifts were given the range zero to seven and the $\phi$ shifts given the range zero to thirty-two. In all cases, the algorithm was set to separate two sources from the mixture.

The unsupervised NMF algorithm separates a pre-defined number of sources. There is no classification of sources as target or interferer; a number of TF masks are generated, that number being equal to the prescribed number of sources in the mixture: in this case two. For each mixture, the two masks were applied and PEASS was used in order to ascertain which of the separations was the target. PEASS does require signals to be classified as target or interferer and the mask that returned the audio receiving the best OPS was deemed to be the target mask and used to generate the results.
8.3 Results

As with Chapter 7, results were analysed by averaging across all mixtures at a given \( p \) value. Both PEASS and BSS Eval results are given so that differences between the perceptually-motivated and the purely physical metrics can be discussed.

8.3.1 PEASS results

The PEASS metrics for the experiment are shown in Figure 8.1. The TPS and APS results show similar trends to those presented in Figure 7.2 (page 104): artefacts
Figure 8.2: BSS Eval metrics for Gao & Woo’s quasi-EM IS-NMF2D algorithm. Dashed lines show the 95% confidence intervals. For reference, the dotted lines show the means obtained in Chapter 7 when using ideal masks.

are least prevalent when the masking is near flat, $p = 2^{-5}$, and the TPS is again correlated with the artefact score.

The IPS score is reduced by the performance of the NMF algorithm that has been used. The OPS score is also reduced as a result of the lack of interferer suppression by the algorithm. However, the trend in the OPS data is similar to that observed previously: the peak result—70 on the previous experiment but 33 here—still lies in the region of $p = 2^{-1}$. 

114
8.3.2 BSS Eval results

The BSS Eval results are shown in Figure 8.2. Previously, in Figure 7.3 (page 105), the mean SDR values were all positive. This is no longer the case, again suggesting that performance has been significantly reduced by the error in the non-ideal TF masks.

8.4 Chapter summary

This chapter aimed to answer the question “7. Which sigmoidal mask provides optimum quality under non-ideal conditions?” Gao & Woo’s (2014) state-of-the-art algorithm provides an estimate of the time-frequency mask using Quasi-expectation maximisation Itakura-Satio two-dimensional NMF (Quasi-EM IS-NMF2D). Using the same experiment and corpus as in the previous chapter, the optimal sigmoid has again been found to lie in the region of $p = 2^{-1}$. The peak value is much diminished, compared to the ideal mask case, as the IPS of the non-ideal mask is much lower.
Chapter 9

Optimised sigmoidal masking

The PEASS results presented in the previous two chapters show that a sigmoidal TF mask can provide better quality separated audio than a binary or ratio mask. Sigmoidal masking was developed from the work of Grais & Erdogan (2011) and was extended in the previous chapter to include powers less than one. This chapter will apply further extensions to the concept of sigmoidal masking in order to determine if further perceptual quality improvements are possible.

This chapter’s research question is: “8. How might sigmoidal masking be further optimised?” This will be answered by a series of four experiments each of which tests a modification to the sigmoidal masking process described in the previous chapter. These four modifications are: an offset switching function, hysteresis, smoothing, and frequency dependence.

To offset the mask, the switching function of the TF mask is translated vertically and horizontally. The motivation for this is to question whether the sigmoid needs to pass through 0.5 or whether better quality can be achieved by offsetting the centre. The method and results of this work are presented in Section 9.1.

Hysteresis, described in Section 9.2, takes into account previous masking decisions in each channel. Hysteresis is explored as earlier chapters suggest that artefacts are related to the switching of the mask. Hysteresis provides separate switching processes for turning the mask on and off.

Section 9.3 explores the use of smoothing to utilise surrounding decisions in time and frequency to change masking coefficients. This is motivated by the idea that
smoothing can lessen the transitions in the mask which may reduce artefacts.

Frequency dependent masking is investigated in Section 9.4; this technique takes into account the frequency sub-band that is being analysed to inform masking decisions. The hypothesis motivating this work is that different sigmoids may be optimal at different frequencies.

In this chapter, ‘optimal sigmoidal mask’ (OSM) refers to the optimal mask, in terms of OPS, identified in Chapter 7 and \( p_{opt} \) refers to \( 2^{-\frac{4}{3}} \), the sigmoidal \( p \) value used to generate the OSM.

9.1 Offset sigmoidal masking

In this section, the OSM’s switching function will be offset horizontally and vertically to study the effect of this on separated audio quality. This will allow asymmetric variation from the IRM’s switching function.

9.1.1 Method

The definition of the OSM was extended by adding the offset parameters, \( a \) and \( b \), to give

\[
M^{OSM} = \frac{(X - a)^{p_{opt}}}{(X - a)^{p_{opt}} + (Y + a)^{p_{opt}}} + b
\]

(9.1)

In the case that \( a = b = 0 \) then the OSM remains as previously defined.

To ensure the switching function remained real, monotonic and limited between zero and one it was constrained by

\[
M^{OSM} = \begin{cases} 
0 & \text{if } M^{IRM} < a \text{ or } M^{OSM} < 0 \\
1 & \text{if } M^{IRM} > 1 + a \text{ or } M^{OSM} > 1 \\
M^{OSM} & \text{otherwise}
\end{cases}
\]

(9.2)

To test the perceptual quality of audio separated using the above idea, masks were generated using \( a \) and \( b \) values equal to \(-0.25\), \(0\), and \(0.25\). The nine combinations of
Chapter 9. Optimised sigmoidal masking

Figure 9.1: The offset sigmoidal curves used for mask generation. On each set of axes, the value of $a$ increases from the left curve to the right curve.

The three values for each of the two variables were used to calculate TF masks for the mixture corpus from Chapter 7. PEASS metrics were calculated and averaged across each combination of $a$ and $b$ to establish whether offsetting the switching function is advantageous. Figure 9.1 shows the switching functions that were tested.

9.1.2 Results

The mean PEASS scores for each combination of offsets are shown in ???. The mean TPS varies between 60, recorded when $a$ and $b$ are $-0.25$, and 76 recorded with no offset. The mean IPS is maximised with no offset at a value of 85. The mean APS reaches a peak of 85 when $a = 0.25$ and $b = -0.25$. The maximum mean OPS score, 67, is achieved with no offset applied to the sigmoidal switching function.
Table 9.1: The results of the offset sigmoidal experiment. $a$ and $b$ are the horizontal and vertical offset parameters respectively. The PEASS results quoted are averaged across all mixtures. The highest value in each column is shown in bold.

<table>
<thead>
<tr>
<th>$a$</th>
<th>$b$</th>
<th>TPS</th>
<th>IPS</th>
<th>APS</th>
<th>OPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.25</td>
<td>-0.25</td>
<td>73</td>
<td>55</td>
<td>70</td>
<td>38</td>
</tr>
<tr>
<td>-0.25</td>
<td>0</td>
<td>71</td>
<td>84</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>-0.25</td>
<td>0.25</td>
<td>60</td>
<td>77</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>0</td>
<td>-0.25</td>
<td>72</td>
<td>44</td>
<td>84</td>
<td>21</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>76</td>
<td>85</td>
<td>54</td>
<td>67</td>
</tr>
<tr>
<td>0</td>
<td>0.25</td>
<td>67</td>
<td>78</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>0.25</td>
<td>-0.25</td>
<td>71</td>
<td>39</td>
<td>86</td>
<td>18</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>75</td>
<td>59</td>
<td>83</td>
<td>33</td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>69</td>
<td>78</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

9.1.3 Discussion

These results show that the OPS is optimal when the OSM is applied with no offsetting. This condition has been compared to combinations of horizontal and vertical shifts of $\pm 0.25$. This finding suggests that offsetting the mask switching function will not provide improved quality separated audio.

The results in Table 9.1 again show the artefact-interferer trade off; the IPS is minimised and the APS maximised when there is the least area under the switching function, $(a, b) = (0.25, -0.25)$. When the area under the switching function is maximised, at $(a, b) = (-0.25, 0.25)$, the APS is at its minimum while the IPS is at 77 having passed through its maximum at $(0, 0)$.

9.2 Hysteresis

In this section, hysteresis is applied to the switching of the TF mask aiming to improve its quality. A hysteretic switch is one which turns on at a different rate to that at which it turns off. The application of hysteresis to the mask calculations may improve the audio quality by increasing the amount of target signal that is required to make the mask switch from ‘off’ to ‘on’ and the amount of interferer
that is required to make it switch from ‘on’ to ‘off’. A reduced amount of switching should, in turn, reduce artefacts. This hypothesis will be tested here using the audio corpus defined in Chapter 7.

Method

This experiment makes use of Preisach’s (1935) hysteresis model as described in Mayergoyz (1991). The Preisach model uses several non-ideal relays referred to as hysterons. The $n^{th}$ hysteron exhibits the following behaviour

$$h_n(x) = \begin{cases} 
1 & \text{if } x \geq \beta(n) \\
0 & \text{if } x \leq \alpha(n) \\
k & \text{if } \alpha < x < \beta 
\end{cases}$$

(9.3)

where $\alpha$ and $\beta$ represent the lower and upper switching thresholds of the hysterons and $k$ is the previous output of $h(x)$.

The outputs from the hysterons are summed together to give the hysteretic output. Whilst the Preisach model allows for weighting to be applied to each hysteron, this possibility is not considered here. This leaves the hysteretic sigmoidal mask in each TF cell as

$$M_{HSM}^{f} = \sum_{n} h_{nf}(x)$$

(9.4)

The Preisach model was applied to the optimal sigmoidal mask to give

$$\alpha(n) = \frac{n_{popt}}{n_{popt} + (1 - n)_{popt}} - g(-n^2 + n)$$

(9.5)

$$\beta(n) = \frac{n_{popt}}{n_{popt} + (1 - n)_{popt}} + g(-n^2 + n)$$

(9.6)

where $g$ controls the amount of hysteresis.

A system of one thousand hysterons was used to produce the hysteretic mask. This number was decided on by a preliminary study of the effect of mask coefficient
quantisation on the PEASS metrics that found that a mask calculated to more than two decimal places will be within a perceptual score scale point of the value calculated when double precision mask coefficients are used.

9.2.1 Results

The results of the hysteresis experiment are shown in Figure 9.2. There was little variance shown in the TPS scores from the application of hysteresis; they remained around 76. The IPS was observed to reduce, from 85 to 50, as more hysteresis was added to the system. The APS was increased, from 53 to 85, by increasing the hysteresis reducing the switching speed. The OPS was reduced by the application of hysteresis; at zero hysteresis, effectively the OSM, the OPS was 67, the same peak score recorded in Chapter 7.

9.2.2 Discussion

The results in Figure 9.2 show that hysteresis does not improve the quality of the separated audio. The optimisation of the artefact-interferer trade off is lost as more hysteresis is added. The APS is observed to increase with hysteresis as a result of the mask not switching as rapidly. However, this prevents the mask from suppressing as much interferer and this reduces the overall audio quality.

9.3 Smoothing

It is possible to smooth a TF mask in three different ways: in time, in frequency and in both time and frequency. Smoothing the mask may improve the transitions between TF units and thus improve the audio quality. Conversely, it may allow too much interferer into the separated signal and thus reduce the OPS.
Figure 9.2: The variation in PEASS metrics with increasing hysteresis. The solid line follows the change in the mean with × markers at the measurement points. The dashed lines mark the 95% confidence intervals.
Chapter 9. Optimised sigmoidal masking

9.3.1 Method

The smoothing applied in this section used a simple box filter with filter coefficients equal to the reciprocal of the filter length in the one-dimensional case (smoothing in just time or frequency) and the squared reciprocal in the two-dimensional case (smoothing in both time and frequency). Filter lengths were varied between zero frames and fifty frames in steps of five frames.

Smoothing was tested in two ways: firstly, over the mask itself to try to reduce the severity of the transitions. The second method attempted smoothed the cochleagram of the target signal, using the same kernels as before, to try to reduce unnecessary switching in the mask. The mask was calculated from the smoothed target and the unsmoothed interferer estimates. This second process aimed to account for the auditory masking of a listener which may remove some interferer. If the listener will not detect the interferer then TF masking does not need to be used to attenuate the signal. There are three different ways in which the dimensions of the TF representation are smoothed (just time, just frequency and both time and frequency) and two different places in which the smoothing is applied (over the mask and to the cochleagram of the target signal), giving a total six different methods of smoothing.

9.3.2 Results

The results for each of the six smoothing methods described previously are shown in Figure 9.3. Each of the smoothing techniques are optimal when the filter length is zero. The highest OPS is that of the OSM at 67. The minima reached during the experiment are just above 50 for the time smoothed mask and cochleagram but near 20 for the frequency and TF methods.

9.3.3 Discussion

The results shown in Figure 9.3 show that the smoothing reduces the OPS. The optimal results were recorded when the filter length was zero in all three cases. As
Figure 9.3: The results of the smoothing experiment showing from top to bottom: time, frequency and time-frequency smoothing. The solid line shows the results for the smoothed mask method. The dashed lines show results obtained for the smoothed target cochleagram method.
a filter length of zero returns the OSM this suggests that a smoothing filter should not be pursued as a method for enhancing the TF mask.

The smoothing performed in this section all takes place at the cochleagram level with the filter length calculated in frames. Pursuing a smoothing filter at the sample level may be a valid area of further investigation as shorter, in a temporal sense, filters can be created using this method. Consequently, this may allow smoothing of the switching transients without damaging the interferer suppression.

9.4 Frequency dependent masking

The optimisation of the artefact-interferer trade-off could conceivably be at different points on the sigmoid series in different frequency bands. This hypothesis will be tested in this section by separating the audio corpus defined in Chapter 7 allowing the mask to use a different sigmoid in each of four frequency bands.

The four frequency bands under test are defined by partitioning the 128 ERB-spaced gammatone filters into four groups of 32. This gave the centre frequencies of the filters at the edges of the bands as:

- Band 1: 50 Hz - 478 Hz
- Band 2: 494 Hz - 1590 Hz
- Band 3: 1645 Hz - 4488 Hz
- Band 4: 4630 Hz - 12000 Hz

TF masks were calculated where each of the above bands was allowed to be masked by a sigmoidal mask with \( p \) values of either \( 2^{-5}, 2^{-4}, 2^0 \) or \( 2^5 \). These represent the flat, optimal sigmoidal, ratio and binary masks respectively.
Figure 9.4: The OPS results of the frequency dependent masking test. The horizontal position of a set of axes denotes its value in band 1. The vertical position denotes the value in band 2. The x and y axes denote the values in bands 3 and 4 respectively. The flat (F), sigmoidal (S), ratio (R) and binary (B) masks are each represented by their initials.
9.4.1 Results

The OPS results of this experiment are shown in Figure 9.4. These results have been averaged across the mixture corpus as with previous experiments. The mean OPS scores range as high as 70 which again occurs with an OSM and as low as 14 when flat masking is used.

9.4.2 Discussion

The main finding of the results in Figure 9.4 is that the highest OPS was obtained when all four bands were using the OSM discovered in Chapter 7. This suggests that the optimisation of the artefact-interferer trade-off is not frequency dependent.

Beyond this main finding there are a number of combinations of the optimal sigmoidal mask and the ratio mask that appear to perform well. Were this area to be investigated further—not something that can be recommended by the results obtained here—then testing of these two masks over narrower bands would be the area of most interest. The experiment could also be extended by testing masks closer on the sigmoidal scale to the OSM. As with the full band sigmoidal experiments, the flat and binary masks do not provide good quality separation even when applied to specific sub-bands.

9.5 Chapter summary

This chapter has answered the question, “8. How might sigmoidal masking be further optimised?” This has been achieved by experimenting with four enhancements to sigmoidal masking. Offset sigmoidal masking allowed the mask’s switching function to be translated horizontally and vertically but demonstrated the best OPS scores when the offsets were zero. Hysteresis was added to the TF masking process but the hysteretic masks did not provide an improved OPS over the previous results demonstrated with the OSM. Both one and
two dimensional smoothing was applied to the OSM but the smoothing reduced interferer suppression thus lowering the IPS and OPS. Frequency dependent masking was also demonstrated by splitting the TF mask into four sub-bands and allowing each to take a different sigmoidal $p$ value. The OPS scores generated by the separated audio from this process, showed the OPS was highest when all four sub-bands took the same sigmoid as the OSM. Of the four attempts made to improve on the OSM in terms of OPS, none could provide an improved score.
Subjective assessment of separated audio quality

This thesis has identified TF masking as an audio separation method likely to have scope for perceptual quality improvement. Previous chapters have established that a continuous mask is preferable to a binary one and that a sigmoidal mask with a $p$ value approximately equal to $2^{-1}$ may provide the optimum quality. These findings have used PEASS to establish the quality of the separated audio.

This chapter aims to answer the question “9. Which TF mask do real listeners prefer?” The motivation for answering this question lies in the fact that PEASS represents a good model trained on specific data, and whilst it has been useful in guiding the investigation so far, separated audio may ultimately be consumed by real listeners; their opinions are important.

To answer this chapter’s question, two listening tests are reported. The first is a multiple stimuli with hidden reference and anchor (MUSHRA) style comparison of TF masks and the second is a paired comparison of two masks. Both studies are conducted by a panel of human sound assessors who are asked to perform blind ratings of the quality of audio separated by the different TF masking methods. The audio used is that separated under ideal circumstances in Chapter 7 as PEASS predicts larger perceptual differences for this audio.

This chapter is structured in four main sections followed by a summary. Section 10.1 will explain the MUSHRA experiment’s method. Section 10.2 presents the analysis of the data collected. Section 10.3 presents the second listening test, which was used for verification. Section 10.4 discusses the results obtained in the wider context of the thesis.
Chapter 10. Subjective assessment of separated audio quality

10.1 Method

The listening test followed a method similar to that found in recommendation ITU-R BS 1534-2 (2014) as this is the industry standard test for situations where “a high quality reference signal is used and the systems under test are expected to introduce significant impairments”.

The test comprised a panel of assessors each being asked to rate the overall quality of separated audio in comparison to a reference. Each of the assessors was asked to complete a familiarisation phase in which they listened to all of the stimuli in the test in a randomised order.

After familiarisation, assessors completed the rating part of the test, in which they were asked to rate the audio separated by each of the TF masking techniques. Stimuli were grouped into pages by programme item. The presentation order of both the pages and the stimuli on the pages was randomised.

10.1.1 Assessors

An invitation to participate in the test was sent to prospective assessors from both the University of Surrey and BBC R&D. The resulting panel consisted of 33 assessors: 25 from the BBC and 8 from the University.

The university assessors were all experienced in technical listening tests, while the BBC assessors had a wide range of listening experience. The range of listening experience necessitated the post-screening of assessors using their results to test their discrimination and reliability. The familiarisation stage was particularly important as even the experienced assessors may not have encountered a test rating separated audio before.

10.1.2 Test environment and set up

The environment in which a test is taken and the way the test is carried out can affect the results. This subsection details a number of factors that were considered
when designing this test.

Locations

The test was conducted in two locations. The use of multiple locations increased participation in the test. A comparison of results between locations can be made to see if location has a significant effect on the ratings collected.

An initial, smaller round of testing took place at the University of Surrey, using a ITU-R BS.1116-1 (1997) standard listening room. A larger round of testing was conducted at BBC R&D, in the user experience lab at their building in Shepherd’s Bush. This room is triple glazed to provide isolation from the surrounding city noise and while not a standardised testing room it provided a quiet enough environment for a headphone based test.

Replay method

To minimise the differences between the listening experiences at the different testing locations the assessors used the same pair of Sennheiser HD600 headphones with the same laptop and Focusrite VRM soundcard\(^1\). The matched hardware provided a consistent replay system while the use of headphones minimised any effect from the different rooms used.

User interface

The UI will be familiar to assessors that have previously been involved in MUSHRA listening tests. The interface was adapted from the BeaqleJS project\(^2\) \(^3\) (Kraft & Zölzer 2014), and is shown in Figure 10.1.

---

\(^1\)The VRM audio processing was switched off

\(^2\)Original: http://hsu-ant.github.io/beaqlejs/

\(^3\)Adapted Project: https://github.com/tobywstokes/beaqlejs/
Chapter 10. Subjective assessment of separated audio quality

Figure 10.1: The interface used for MUSHRA testing.
10.1.3 Stimuli

The stimuli for the test were arranged in pages by programme item. Each page featured five processings of the programme item: the hidden reference target signal, the three maskings of the target signal and an anchor signal. The ideal maskings from Chapter 7 were used as these displayed greater differences than the non-ideal cases. Stimuli were presented in a randomised order on each page. The order of the pages was also randomised.

Programme items

The programme items, detailed in Table 10.1, were chosen to represent a range of different target signals, are spaced over the previously defined overlap scale (Subsection 7.2.1, page 102), and produced a range of OPS results: 98, 78, 48 and 69.

<table>
<thead>
<tr>
<th>Target</th>
<th>Interferer</th>
<th>Overlap</th>
<th>Optimal OPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female speech</td>
<td>Environmental noise</td>
<td>0.461</td>
<td>98</td>
</tr>
<tr>
<td>Male speech</td>
<td>Female speech</td>
<td>0.734</td>
<td>78</td>
</tr>
<tr>
<td>Solo violin</td>
<td>Female speech</td>
<td>0.73</td>
<td>48</td>
</tr>
<tr>
<td>Male Speech</td>
<td>100 Hz LP filtered noise</td>
<td>0.933</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 10.1: The programme items to be used for the listening test, their measured overlap and the optimal OPS scores observed when separating each mixture in Chapter 7.

Anchors

The anchor in MUSHRA testing provides a lower limit of quality and should always be rated below the other stimuli by assessors. Emiya et al. (2011) proposed three anchors for use in their own listening tests:

1. The distorted target anchor is defined as the low-pass filtered target source signal, using a 3.5 kHz cutoff frequency, with 20% of the remaining time-frequency coefficients selected at random and set to zero.
2. The interference anchor is the mixture loudness matched to the target signal.

3. The artefacts anchor is defined as the target signal with 99% of the time-frequency coefficients selected at random to be set to zero.

The first anchor is the recommended anchor for MUSHRA testing procedure and the latter two are important as they contain degradations similar to those which will be encountered in this test. A preliminary test was conducted where assessors were asked to rate the three proposed anchors against the three masks and the known target. As a result of this test the artefacts anchor was chosen to be the anchor in the main test. This was because it had a lower mean rating than the other audio on test and also a lower variance than the other potential anchors. While this goes against the ITU-R BS 1534-2 (2014) recommendations for this test, the anchor chosen is more relevant to the audio on test than the distorted target anchor.

Loudness matching

The stimuli were loudness matched for the test. This was done by asking multiple assessors to adjust signals for equal loudness. The assessors results were then averaged for each stimulus and applied as gains to the signals.

10.2 Analysis

As with the design of the experiment, the analysis of the data collected follows the suggestions made in ITU-R BS 1534-2 (2014). Firstly, data from the assessors were post-screened to assess whether it was useful for analysis. Secondly, each assessor’s data were normalised to remove differences between subjects use and understanding of the scale. Thirdly, initial descriptive statistics were calculated for the remaining data. Finally, an ANOVA model was created to understand the distribution of variance within the results.
10.2.1 Post-screening of assessors

Due to the uncertainty about the technical listening ability of some of the assessors, it was necessary to post-screen the data collected to determine whether the assessors were capable of the task. Subjects were post-screened against three criteria specified in ITU-R BS 1534-2 (2014): rating of hidden references, reliability and discrimination. The rating of a hidden references is tested by a single calculation whereas the reliability tests were based on the eGauge tool as described by ITU-R BS.2300-0 (2014) and Lorho et al. (2010). eGauge measurements are based on an ANOVA of an individual assessor’s ratings.

Hidden reference ratings

ITU-R BS 1534-2 (2014) states that, “an assessor should be excluded from the aggregated responses if he or she rates the hidden reference condition for >15% of the test items lower than a score of 90.” This was easy to test and resulted in two subjects being removed from further analysis. The reference signal in this task was not particularly difficult to detect and failure to rate it at 100 as instructed suggests that the subject either did not understand the task or could not reliably detect the reference.

Reliability

Reliability is an assessor’s ability to repeatedly give the same rating for a given stimulus (Lorho et al. 2010). In the eGauge model, reliability is measured by taking the ratio of the average standard deviation of an assessor’s scores to the square-root of the mean square regression (MSR) for that assessor. The MSR is given by the sum of squared differences between assessors ratings of each TF mask divided by the degrees of freedom. The reliability of the $j^{th}$ assessor is given by

$$\text{Reliability}_j = \frac{\text{Span}_j}{\sqrt{\text{MSR}_j}}$$

(10.1)
Chapter 10. Subjective assessment of separated audio quality

Figure 10.2: The reliability calculations for each assessor with a line showing the 95% threshold. The leftmost six assessors have been removed from further analysis.

Discrimination

Lorho et al. (2010) describe discrimination as “the ‘signal-to-noise ratio’ of the repeated rating of a set of stimuli by an assessor”. The eGauge model measures the discrimination as the ratio of the mean sum of squares (MSS) to the mean square regression. The MSS is given by taking the sum of squared differences between each system and the assessor’s mean rating then dividing by the degrees of freedom. The discrimination ability of the $j^{th}$ assessor is given by

$$\text{Discrimination}_j = \frac{\text{MSS}_j}{\text{MSR}_j}$$  \hspace{1cm} (10.2)
Figure 10.3: The discrimination calculations for each assessor with a line showing the 95% threshold. The leftmost three assessors have been removed from further analysis.

Permutation testing

To measure the significance of the reliability and discrimination measures, eGauge uses permutation testing as described by Dijksterhuis & Heiser (1995). The permutation test randomly permutes the scores from each assessor and then calculates the eGauge metric under test. Continued permutation and calculation allows a distribution of ratings by chance to be collected. After a number of permutations—the analysis here uses 250 as recommended in the authors' implementation—a distribution of values has been collected from which the ninety-fifth percentile can be calculated. This value is used as the rejection threshold for each of the metrics.
Chapter 10. Subjective assessment of separated audio quality

Post-screening results

The statistical programming environment R was used to apply the above processing. The eGauge analysis identified a total of 7 assessors for removal from the test: four on the basis of reliability, one on the basis of discrimination and two on both bases. Assessors’ reliability and discrimination scores are shown with each 95% threshold in Figure 10.2 and Figure 10.3. With the two assessors removed by the hidden reference screening, a total of 9 assessors were removed leaving 24 assessors’ results to be used in further analysis.

10.2.2 Data pre-processing

Each assessor’s scores were normalised to remove differences in the way they may have perceived or used the scale. Normalisation is subtraction of the mean of the data set and division by the standard deviation. Each normalised score represents, in standard deviations, the difference between an assessor’s rating of an individual stimulus and the mean rating they gave across all stimuli. This departure from ITU-R BS 1534-2 (2014) accounts for differences in the way assessors may have perceived the scale. It can be shown that analysis of this experiment using either normalised or raw scores produces the same findings.

Normalisation does not include the reference or the anchor scores as these have served their purpose in marking the extremes of the scale and allowing assessors’ listening abilities to be tested. Consequently, the reference and anchor scores are not included in any of the forthcoming analysis. In total, 576 ratings are retained for analysis.

10.2.3 Initial analysis

Exploratory analysis was performed to visualise the dataset and understand how it is spread in different groupings. This included box plots, normality tests and

---

Subjective assessment of separated audio quality

multi-modality tests.

Box plots

Boxplots have been created for each combination of mask and programme item. These are shown in Figure 10.4. These show the medians and inter-quartile ranges for each combination of mask and programme item. The box plots also give a good visual indication of how skewed the distributions might be. In all cases, the binary masks ratings have lower medians than those of the ratio or sigmoidal masks.

![Box plots showing the spread of data for each of the programme items.](image)

**Figure 10.4:** Box plots showing the spread of data for each of the programme items.
Normality testing

Normality testing can help with later decisions about how the data should be used to inform a model. This can be tested visually and using numerical test of modality. Both of these approaches will be applied here.

Figure 10.5 shows histograms of each programme item’s ratings with an estimated normal distribution curve plotted for reference. The histograms suggest that, for a number of stimuli, the assessors’ ratings deviate from normal due to skew or kurtosis.

Multimodality testing

Each distribution in Figure 10.5 looks unimodal and this is also supported by the numerical multimodality test,

\[ b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}} \]

suggested in ITU-R BS 1534-2 (2014), where \( g \) represents the skew, \( k \) the kurtosis and \( n \) the number of elements in the data set. Values of \( b \) close to one indicate the distribution is likely to be multimodal. All values measured for these distributions were in the order of \( 10^{-3} \).

Initial comparison

An initial comparison of the ratings of each mask can be made by looking at the confidence intervals of the scores ratings. Bootstrapped confidence intervals have been used as these are deemed more appropriate when the underlying distributions are not normal. These are plotted in Figure 10.6, these plots suggest that there is a clear preference for the continuous masks over the binary one. Amongst the continuous masks the ratio mask looks to be the preferred mask.
Chapter 10. Subjective assessment of separated audio quality

![Histograms of the normalised assessor responses for each combination of programme item and mask. For reference, a Gaussian curve is plotted over each distribution.](image)

**Figure 10.5:** Histograms of the normalised assessor responses for each combination of programme item and mask. For reference, a Gaussian curve is plotted over each distribution.
10.2.4 Model creation

A two-way repeated-measures ANOVA was conducted on the data post-screened ratings. Some violations of normality are suggested by Figure 10.5 but ANOVA is robust to normality violations “when group sizes are equal” (Field et al. 2013) as they are in the data here. The ANOVA can apportion the variance in the ratings data between the masks, programme items and any interaction between the two.

Mauchly’s test

The assumption of sphericity, which dictates that the differences between the conditions should have equal variance for a repeated-measures ANOVA to produce a valid F-ratio (Field et al. 2013), was tested using Mauchly’s test. The results, shown in Table 10.2, show that the sphericity assumption was violated for the effects of the mask and the programme item.
Chapter 10. Subjective assessment of separated audio quality

<table>
<thead>
<tr>
<th>Effect</th>
<th>W</th>
<th>p</th>
<th>p&lt;.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask</td>
<td>0.618</td>
<td>0.005</td>
<td>*</td>
</tr>
<tr>
<td>Programme</td>
<td>0.599</td>
<td>0.049</td>
<td>*</td>
</tr>
<tr>
<td>Mask:Programme</td>
<td>0.317</td>
<td>0.253</td>
<td></td>
</tr>
</tbody>
</table>

**Table 10.2**: Mauchly’s test output table from the R console. The output shows significant violations of sphericity for the mask and programme factors.

<table>
<thead>
<tr>
<th>Effect</th>
<th>DFn</th>
<th>DFd</th>
<th>SSn</th>
<th>SSd</th>
<th>F</th>
<th>p</th>
<th>p&lt;.05</th>
<th>ges</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1</td>
<td>25</td>
<td>5.76e5</td>
<td>5.63e4</td>
<td>255.62</td>
<td>1.23e-14</td>
<td>*</td>
<td>0.846</td>
</tr>
<tr>
<td>Mask</td>
<td>2</td>
<td>50</td>
<td>1.11e4</td>
<td>7.72e3</td>
<td>36.05</td>
<td>2.02e-10</td>
<td>*</td>
<td>0.096</td>
</tr>
<tr>
<td>Programme</td>
<td>3</td>
<td>75</td>
<td>6.63e4</td>
<td>2.82e4</td>
<td>58.60</td>
<td>1.29e-19</td>
<td>*</td>
<td>0.387</td>
</tr>
<tr>
<td>Mask:Programme</td>
<td>6</td>
<td>150</td>
<td>1.79e3</td>
<td>1.25e4</td>
<td>3.577</td>
<td>2.43e-03</td>
<td>*</td>
<td>0.017</td>
</tr>
</tbody>
</table>

**Table 10.3**: The output of the ANOVA process showing significant effects for the mask, programme and the interaction between them.

Main effects

The main effects are shown in Table 10.3. The effect of programme is adjusted due to the violation of sphericity.

10.2.5 Post hoc tests

Bonferroni post hoc tests were run on all groupings of mask and programme items. The results of this test are shown in Table 10.4. The values show, firstly, that the binary mask ratings are significantly different, at 95%, from the ratio mask ratings for all programme items. Secondly, the binary mask ratings are significantly different from the sigmoidal mask ratings for only the first programme item. Finally, the ratio and sigmoidal mask ratings are significantly different for only the fourth programme item.

In Table 10.5, the same Bonferroni-corrected t-test is shown for the individual mask groupings. These tests agree with the suggestions of the bootstrapped confidence intervals in Figure 10.6 that there are significant differences between all three masks.
<table>
<thead>
<tr>
<th></th>
<th>BinP1</th>
<th>BinP2</th>
<th>BinP3</th>
<th>BinP4</th>
<th>RatP1</th>
<th>RatP2</th>
<th>RatP3</th>
<th>RatP4</th>
<th>SigP1</th>
<th>SigP2</th>
<th>SigP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>BinP2</td>
<td>7.8e-13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BinP3</td>
<td>1.1e-15</td>
<td>1.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BinP4</td>
<td>&lt; 2e-16</td>
<td>0.05392</td>
<td>1.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RatP1</td>
<td>3.4e-05</td>
<td>&lt; 2e-16</td>
<td>&lt; 2e-16</td>
<td>&lt; 2e-16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RatP2</td>
<td>0.00180</td>
<td>9.4e-08</td>
<td>6.3e-07</td>
<td>4.4e-08</td>
<td>5.6e-13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RatP3</td>
<td>3.6e-07</td>
<td>1.00000</td>
<td>0.00031</td>
<td>0.00241</td>
<td>3.1e-15</td>
<td>1.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RatP4</td>
<td>0.00177</td>
<td>0.00020</td>
<td>9.5e-05</td>
<td>9.5e-09</td>
<td>6.5e-11</td>
<td>1.00000</td>
<td>1.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SigP1</td>
<td>0.00018</td>
<td>&lt; 2e-16</td>
<td>&lt; 2e-16</td>
<td>&lt; 2e-16</td>
<td>1.00000</td>
<td>8.9e-14</td>
<td>&lt; 2e-16</td>
<td>4.8e-15</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SigP2</td>
<td>0.00065</td>
<td>0.05056</td>
<td>0.00187</td>
<td>4.8e-05</td>
<td>4.2e-11</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>3.8e-15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SigP3</td>
<td>3.6e-08</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>2.6e-15</td>
<td>0.00494</td>
<td>0.16973</td>
<td>0.01564</td>
<td>&lt; 2e-16</td>
<td>0.03768</td>
<td>7-</td>
</tr>
<tr>
<td>SigP4</td>
<td>1.1e-09</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>&lt; 2e-16</td>
<td>3.1e-05</td>
<td>1.00000</td>
<td>7.7e-06</td>
<td>&lt; 2e-16</td>
<td>0.00814</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

Table 10.4: Results of the Bonferroni post hoc tests conducted between each pairing of mask and programme item
Chapter 10. Subjective assessment of separated audio quality

<table>
<thead>
<tr>
<th></th>
<th>Binary</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>&lt; 2e-16</td>
<td>-</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>8.0e-09</td>
<td>2.2e-05</td>
</tr>
</tbody>
</table>

Table 10.5: Showing significant differences between the ratings given to audio from each of the masks.

10.3 Paired comparison test

The MUSHRA test found assessors prefer a ratio mask to a sigmoidal mask and both these masks to the binary mask. There are two potential reasons to question this finding: firstly, it disagrees with the output of the PEASS model. Secondly, it is possible that instructing assessors to rate the quality of separated audio could have biased them to rate signals containing interferer lower than they would otherwise. For these reasons, a further paired comparison test was undertaken to establish assessors’ preferences for either the ratio or sigmoidal masks.

10.3.1 Method

The second test was a blind AB test; the stimuli were different separations of the same audio. The AB test was chosen to force the assessors to make a choice between the audio separated by each mask. The AB test was chosen over an ABX test, in which the assessor also has access to a reference signal, as the presence of the reference may diminish the differences between the two stimuli under test.

The test instructions did not explain the nature of the investigation; the task was explained purely as a test of quality with no mention of BASS. As assessors were required who did not know that BASS was being tested, they were all from a group who were unfamiliar with this research project.

Each assessor was asked to choose between the ratio mask and the sigmoidal mask to suggest which was of greater quality. The four programme items remained the same as the previous test. Each trial was replicated once meaning eight data were generated by each assessor.
User interface

The UI was again adapted from BeqleJS, which contains an ABX interface but not an AB testing interface. The UI is shown in Figure 10.7.

Figure 10.7: The UI from the AB test

10.3.2 Results

The test was sat by eleven assessors—studio managers and radio technologists, based at BBC Broadcasting House—giving a total of eighty-eight comparisons of the ratio and sigmoidal masks. The assessors favoured the ratio mask in 56 of these cases, preferring the sigmoidal mask in the remaining 32. Calculating binomial confidence intervals for these figures suggest this is a significant preference at the 95% level. These results are visualised in Figure 10.8.

10.4 Discussion

Results in previous chapters of this thesis suggested listeners would prefer audio separated by a sigmoidal mask to audio separated by a ratio mask and also prefer separated audio from a ratio mask to audio from a binary mask. The work in this chapter supports the second of these two claims but opposes the first.
Figure 10.8: The results of the paired comparison. The error bars show the 95% confidence intervals.
The preference for a continuous mask over a binary one is clear; the work in this chapter adds subjective assessment of quality to the arguments made earlier in the thesis from PEASS data. Once assessor data had been normalised, the mean rating of the binary masked audio was below the assessors’ averages. While many arguments have been made for the use of the binary mask for separating audio, the quality of the separated audio it produces cannot be one when continuous masking is also considered as an option.

The preference expressed for the ratio mask over the sigmoidal mask is contrary to the suggestion of PEASS results in Chapter 7 and Chapter 8. A possible reason for this is that the PEASS model was trained on stereo data whereas the audio in these tests was mono. The perceived level of interference can differ in mono and stereo as the human auditory system is able to distinguish and treat differently sounds from different locations. The results obtained seem to suggest that, for human listeners, the optimisation of the artefact-interferer trade-off is not located where PEASS would suggest.

Table 10.6 shows the rank order of the masking methods according to APS, IPS, OPS and the listener data. This analysis seems to suggest that the level of interference is under represented in the overall score by PEASS for these examples. Again, this could be due to the mono-stereo difference between the PEASS training set and the audio used in this test. To understand this further, a multi-stage listening test, where assessors rate overall quality and then target quality, interferer quality and artefact quality, would be required to allow a better comparison with Emiya et al.’s (2011) work.

<table>
<thead>
<tr>
<th>Mask</th>
<th>IPS</th>
<th>APS</th>
<th>OPS</th>
<th>Listener Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Ratio</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 10.6: Rank order comparisons of the three masks using four metrics: IPS, APS, OPS and the listener data from this chapter.
Chapter 10. Subjective assessment of separated audio quality

10.5 Chapter summary

This chapter aimed to answer the question: “9. Which TF mask do real listeners prefer?” The results of a MUSHRA style listening test clearly show that the ratio and optimal sigmoidal masks were preferred to the binary mask. They also suggest a preference for the ratio mask over the sigmoidal mask found previously to be optimal.

A paired-comparison test confirms the assessor preference for the ratio mask over the sigmoidal despite the disagreement of this finding with the results of previous studies using PEASS to assess the audio quality.

The differences between the PEASS results and the real assessor results may be due to a difference in the perception of artefacts in mono and stereo situations. Further studies would be required to test this hypothesis.
Chapter 11

Conclusions and further work

This thesis has identified BASS techniques suitable for broadcasting and made a number of discoveries in attempting to enhance the perceptual quality of the audio separated by these techniques. In this final chapter, the findings of each chapter are summarised, as are the answers to the research questions outlined in Chapter 1. Additional contributions to knowledge and areas for further work are then detailed. The thesis summary can be found in Section 11.1, contributions to knowledge in Section 11.2 and further work in Section 11.3.

11.1 Thesis summary

The research aims of this thesis were to answer the series of questions set out in Figure 1.1 on page 6. This section will summarise the report chapter by chapter answering the research questions as they arise.

11.1.1 Main aim

The main aim of this thesis was to answer the question “How can the perceptual quality of BASS be improved for broadcasting applications?” TF masking is a BASS technique suited to the audio mixtures encountered in broadcasting. TF masking can produce low-quality results, but the quality of the audio, as measured by the PEASS OPS metric, can be improved by up to 50 points by using a sigmoidal mask as defined in Equation 7.2. Formal listening tests suggest that a ratio mask
may be preferred in practice. Further tests documented in this thesis find no
evidence that modifications based on offsets, hysteresis, smoothing or frequency-
dependent masking can offer any additional quality improvement.

11.1.2 Chapter 1

In Chapter 1, BASS was introduced in general terms as a problem involving
separating acoustic energy related to one sound source from a multi-source
recording. TF masking was introduced as a method of performing BASS in the TF
domain. The main aim (summarised above) was broken down into nine specific
research questions, addressed in the subsequent chapters of the report.

11.1.3 Chapter 2

Chapter 2 aimed to answer the question, “1. What is BASS and how can
it be achieved?” BASS was found to be a term that relates to a subset of
audio processing problems involving treating parts of an audio mixture differently
according to the source that sound originates from. To answer this question, a
literature review was conducted answering two sub-questions:

1.1 How are BASS problems classified?

1.2 How can sounds from different sources be separated?

How are BASS problems classified?

BASS problems are classified in terms of the type of mixture and the desired output
of the BASS process. Mixtures can be categorised in terms of stationarity, the ratio
of sources to mixture observations and whether convolutional effects are present.
Outputs required from the BASS process can vary from information retrieval tasks
to extraction of the entire source from a mixture.
Chapter 11. Conclusions and further work

How can sounds from different sources be separated?

Separation of a sound source from an audio mixture is possible by CASA, ICA or NMF. CASA and NMF rely on extraction spectro-temporal cues whereas ICA focuses on statistical cues.

11.1.4 Chapter 3

The aim of Chapter 3 was to answer the question, “2. How can BASS techniques be evaluated?” A literature review was used again to answer this question. BASS techniques can be evaluated either subjectively or objectively. An objective metric is a computational measurement of the difference between the estimate signal and the true target signal. A subjective metric asks listeners to audition these two signals and rate the differences. The PEASS metrics are appropriate for work in this thesis as they aim to predict listener opinion without the prohibitive time overheads associated with listening tests.

11.1.5 Chapter 4

A study of audio production techniques in the broadcasting industry was conducted in Chapter 4 aiming to answer the question, “3. What are the BASS requirements of the broadcasting industry?” This was achieved by documenting a series of conversations and observations with BBC sound supervisors. Sound staff have many responsibilities beyond the programme audio and hence are too busy to dedicate time to a specific audio separation tool. Noise control techniques already in place include automatic gating and stationary noise removal tools. An audio separation tool for deployment in industry must be able automatically deal with an unknown number of sources within a mixture and non-stationary mixtures.
11.1.6 Chapter 5

TF masking is a method of source separation that would suit the criteria of the broadcasting industry. The TF mask tends to introduce artefacts in the process of suppressing interfering audio. The answer to this chapter’s question, “4. What should be the subject of further investigation?”, is the improvement of the perceptual quality of audio separated by a TF mask.

11.1.7 Chapter 6

The widely documented binary TF mask was the subject of Chapter 6. The chapter aimed to answer the question, “5. Can binary masking performance be improved?” A study of multiple modified ideal binary masks showed that continuous masks produced better PEASS OPS scores than binary ones.

11.1.8 Chapter 7

Chapter 7 built on the findings of the previous chapter by answering the question, “6. Which sigmoidal TF mask provides optimal separated audio quality?”. A range of sigmoidal functions were used to create TF masks on a continuum between binary, ratio and a flat mask. A trade-off exists between the artefacts and interferer suppression. Optimising this trade-off gives the optimal sigmoidal mask. The optimal sigmoidal mask was obtained when the target and interferer energy in each cell was raised to the power of $2^{-\frac{4}{3}}$. This result has been demonstrated at multiple TF resolutions.

11.1.9 Chapter 8

The aim of Chapter 8 was to answer the question, “7. Which sigmoidal mask provides optimum quality under non-ideal conditions?” This was achieved by repeating the experiment in Chapter 7 with a non-ideal estimate of the target and interferer signal. Gao & Woo’s (2014) algorithm provided estimates of the
target and interferer signals. The results showed that while the OPS scores were reduced, by the reduced accuracy of target and interferer detection, the peak score still lay in the same region of the sigmoid set as in the previous experiment. This led to the conclusion that the previous finding is valid in both ideal and non-ideal cases.

11.1.10 Chapter 9

In Chapter 9, the aim was to answer the question, “8. How might sigmoidal masking be further optimised?” This chapter presented a number of studies looking to further the work on sigmoidal TF masking in the previous chapter. Of the four techniques studied—offset masking, hysteresis, smoothing and frequency dependent masking—none could improve the separation of the beyond the mask identified in Chapter 7.

11.1.11 Chapter 10

Chapter 10 aimed to answer the this thesis’ final research question, “9. Which TF mask do real listeners prefer?” This was achieved by two sets of subjective testing to establish the views of a panel of listeners on the quality of audio separated by binary, ratio and optimal sigmoidal masks. Two separate groups of listeners were shown to favour the quality of audio separated by the ratio mask. The first group also significantly preferred the optimal sigmoidal mask to the binary mask.

11.2 Contributions to knowledge

The research in this thesis has made a number of contributions to knowledge. These are briefly summarised in this section.
11.2.1 The sigmoidal continuum between the flat and binary masks

The formula for sigmoidal masking from Grais & Erdogan (2011) can be used to provide a continuum of TF masks that encompasses masks from flat through ratio to binary. In this thesis this was achieved using a series of powers of two.

11.2.2 Sigmoidal optimisation of artefact-interferer trade-off

Good quality TF mask based separation must optimise the artefact-interferer trade-off demonstrated in Chapter 5. Chapter 6 demonstrated this is possible using a sigmoidal mask with a $p$ value in range of $2^{-\frac{4}{3}}$. This was again demonstrated using non-ideal masking in Chapter 8.

11.2.3 Perceptual preference for ratio mask over a binary mask

There has been much discussion in the literature regarding the merits of binary (Wang 2005) and ratio (Hummersone et al. 2014) masks for source separation. A number of objective and subjective metrics have been used to argue for the use of each of these masks but a listening test where quality is assessed has not. The listening tests presented in Chapter 10 show that, in terms of quality, listeners prefer sounds separated by a ratio mask.

11.2.4 Negative findings

Chapter 9 introduced four novel methods to attempt to improve TF masking: offset sigmoids, hysteresis, filtering the mask and frequency dependent masking. The results indicate that none of these methods can offer any improvement to audio quality; the formulations of these approaches are available for future research.
11.2.5 Data to question PEASS

Listener preference has been shown to not match well with the PEASS predictions of perceptual quality for the audio corpus used in this thesis. This suggests PEASS is not a reliable predictor of perceptual audio quality for certain audio mixtures. These results may lead other researchers to question further the PEASS model with the aim of producing better models in the future.

11.3 Further work

Research work often raises as many, if not more, questions than it can answer. This thesis is no different. This section will introduce some topics for further work. These will be limited to areas arising directly from the work contained within this thesis.

11.3.1 Understanding the relationship between the ideal and non-ideal cases

The results in Chapter 7 and Chapter 8 complement each other by showing the same experiment under ideal and non-ideal conditions. However, no measurements have been taken about just how different these two cases are. The perceptual scores have been compared but the differences in the masks have not been quantified.

Further work could take one of two approaches: firstly, quantifiable amounts of error could be introduced into ideal TF masks and the change in perceptual quality measured against error. Secondly, non-ideal masks could be calculated and the amount of error measured using an adapted form of the IBMR.

11.3.2 An improved perceptual model

In Chapter 3, the literature strongly suggested that PEASS was the most advanced perceptual model for measuring the quality of BASS. However, the results in
Chapter 10, when contrasted with earlier chapters, suggest that PEASS is not entirely suited to measuring the quality of the audio separated in this thesis. An improved perceptual model could be pursued either by re-training PEASS for more specific tasks or by building a new understanding of how listeners perceive the quality of perceptual audio.

11.3.3 The optimal TF basis

A strong case can now be made for using a continuous mask to separate audio in the TF domain. This thesis has worked with the cochleagram as its TF representation for all experimental work. An audio signal can be decomposed into a number of different TF bases. An experiment could be devised where mixtures are separated using TF masking of a number of different bases with the perceptual quality of the outputs compared.

11.3.4 Sample level mask smoothing

Chapter 9 demonstrated that filtering the TF mask directly did not improve the quality of the separated audio. However it also suggested that this smoothing could have taken place at the sample level. This may reduce the extent to which the interferer suppression is reduced by the smoothing. Consequently, this may improve the optimisation of the artefact-interferer trade-off.
List of acronyms

AAC  advanced audio coding
ACF  auto-correlation function
ANOVA analysis of variance
APS  artefacts-related perceptual score
ASA  auditory scene analysis
BASS blind audio source separation
BBC British Broadcasting Corporation
BSS  blind source separation
CASA computational auditory scene analysis
CBM cepstrally-smoothed binary mask
CCF  cross correlation function
CHiME computational hearing in multisource environments
DAB  digital audio broadcast
DBM  dithered binary mask
DCT  discrete cosine transform
DDF  double difference function
DNS  dialogue noise supressor
DVB  digital video broadcast
eaptX enhanced aptX
EBU European Broadcasting Union
ERB  equivalent rectangular bandwidth
FFT  fast Fourier transform
IBM  ideal binary mask
IBMR ideal binary mask ratio
ICA independent component analysis
IID  inter-aural intensity difference
ILD  inter-aural level difference
IPD  inter-aural phase difference
IPS  interference-related perceptual score
IRM  ideal ratio mask
ISDN  integrated services digital network
ISM  ideal sigmoidal mask
ISR  image to spatial distortion ratio
ITD  inter-aural time difference
MADI  multi-channel audio digital interface
MIR  music information retrieval
MOS  mean opinion score
MPEG  Moving Pictures Experts Group
MSR  mean square regression
MSS  mean sum of squares
MUSHRA  multiple stimuli with hidden reference and anchor
MUSIC  multiple signal classification
NBM  noisy binary mask
NMF  non-negative matrix factorisation
OPS  overall perceptual score
OSM  optimal sigmoidal mask
PDF  probability density function
PEAQ  perceptual evaluation of audio quality
PEASS  perceptual evaluation for audio source separation
PSM  perceptual similarity measure
Quasi-EM IS-NMF2D  Quasi-expectation maximisation Itakura-Satio two-dimensional NMF
List of acronyms

**RMS** root mean square

**SACF** summary auto-correlation function

**SAR** sources to artefact ratio

**SBM** segmented binary mask

**SCCF** summary cross-correlation function

**SDR** source to distortion ratio

**SIR** source to interferer ratio

**SNR** sources to noise ratio

**SQAM** Sound Quality Assessment Material

**STFT** short-time Fourier transform

**TF** time-frequency

**TMR** target-to-mixture ratio

**TPS** target-related perceptual score

**UI** user interface
List of symbols

\( \Delta \) A random value selected from a triangular distribution

\( \gamma_l \) The smoothing parameter for the CBM

\( \lambda \) The sum of IBM cells correctly identified as containing target energy

\( A^{-1} \) The un-mixing matrix

\( A \) The mixing matrix

\( M^{CBM} \) The cepstrally-smoothed binary mask (CBM)

\( M^{DBM} \) The dithered binary mask (DBM)

\( M^{HSM} \) The hysteretic sigmoidal mask

\( M^{IRM} \) The ideal ratio mask (IRM)

\( M^{ISM} \) The ideal sigmoidal mask (ISM)

\( M^{NBM} \) The noisy binary mask (NBM)

\( M^{OSM} \) The optimal sigmoidal mask obtained when \( p = 2^{-\frac{4}{3}} \)

\( M^{IBM} \) The ideal binary mask (IBM)

\( \rho \) The number of cells incorrectly identified in the IBM

\( s \) The sources to be mixed

\( x \) The mixtures of the source signals

\( s_j \) The \( j^{th} \) signal from the set \( s \)

\( x_i \) The \( i^{th} \) mixture signal from the set \( x \)

\( \tau_{ij} \) A time lag

\( \text{MSR}_j \) The mean square regression of the \( j_{th} \) listening test assessor
MSS\textsubscript{j} The mean sum of squares for the \textit{jth} listening test assessor

\( \theta \) An angle

\( a_{ij} \) The attenuation of a signal travelling between point \( i \) and \( j \)

\( acf(\tau) \) The auto-correlation function

\( b \) The indicator of a distributions multi-modality

\( c \) The speed of sound in air

\( ccf(n, c, \tau) \) The cross correlation function

\( g \) A distribution’s skew

\( G'(t, \sigma) \) The first differential of the Gaussian function

\( h \) The histogram an IRM

\( H(x) \) The entropy of a variable \( x \)

\( J(x) \) The negentropy of the variable \( x \)

\( k \) The kurtosis of a distribution

\( o \) The TF overlap of two mixtures

\( p \) The parameter of the sigmoidal mask

\( P_{s,n} \) The orthogonal projector onto the interfering sources and noise

\( P_s \) The orthogonal projector onto the interfering sources \( s'_j \)

\( P_{s,j} \) The orthogonal projector onto the source \( s_j \)

\( r \) The radius of a listener’s head

\( w \) The weighting vector used in TF overlap calculations

\( E\{\cdot\} \) The expectation operator

\( \hat{s}_j \) An estimate of the source \( s_j \)
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_j^{\text{interf}}$</td>
<td>The separation error due to the presence of interfering sources</td>
</tr>
<tr>
<td>$s_j$</td>
<td>The $j^{\text{th}}$ source in an un-mixing problem $s_j$</td>
</tr>
<tr>
<td>$e_j^{\text{artif}}$</td>
<td>The separation error due to artefacts</td>
</tr>
<tr>
<td>$e_j^{\text{noise}}$</td>
<td>The separation error due to the presence of noise</td>
</tr>
<tr>
<td>$e_j^{\text{target}}$</td>
<td>The separation error due to distortion of the target</td>
</tr>
</tbody>
</table>
References


References


