ADVANCING TIME-FREQUENCY CLUSTERING TECHNIQUES FOR MULTICHANNEL BLIND SOURCE SEPARATION AND SOURCE NUMBER ESTIMATION

Ingrid Jafari

THE UNIVERSITY OF WESTERN AUSTRALIA

Presented for the degree of Doctor of Philosophy in the School of Electrical, Electronic and Computer Engineering

The University of Western Australia

2014
Abstract

The success of blind source separation systems relies on its ability to handle adverse real-world conditions, and audio applications of blind source separation systems are often faced with the cocktail party problem of multiple simultaneously active speakers. The convolutive nature of the mixing of the source signals and the relative number of the microphones and source signals also add to the difficulty of the blind source separation problem. This thesis focuses on the sparseness-based time-frequency masking approach for under-determined blind source separation. In particular, we direct our attention upon the automatic estimation of time-frequency masks based on the full-band clustering of spatial features as extracted from the microphone observations.

We explore recent advances of full-band clustering-based techniques to blind source separation and investigate and extend an existing technique termed the multiple sensors degenerate unmixing estimation technique (MENUET). We modify the MENUET by using an alternative clustering scheme for mask estimation, the fuzzy c-means, and present comprehensive evaluations in a range of environments to establish its feasibility. We then explore two extensions to the fuzzy c-means: firstly, the estimation of reliability weights for the spatial features, and secondly the inclusion of contextual information into the clustering objective function.

This thesis also investigates other full-band clustering techniques such as the Gaussian mixture model and Watson mixture model for mask estimation. We evaluate and compare the performance of all the full-band clustering techniques in this thesis, and conclusions are drawn as to which are the most robust for a variety of both simulated and real-world conditions, including international benchmark data sets. Finally, to remove any requirement on a priori knowledge on the number of source signals, we consider two novel approaches to source number estimation. The first uses an adaptive optimization scheme based on the fuzzy c-means clustering algorithm, whilst the second approach considers the full-band clustering of speech activity sequences with the Watson mixture model.
# Table of Contents

List of Figures vii

List of Tables xiii

Notations xvii

Acronyms and Abbreviations xix

Acknowledgements xxi

Declaration of Contribution xxiii

1 Introduction 1

1.1 Motivation .................................................. 1

1.2 Research objectives ......................................... 4

1.3 Scope of research ........................................... 5

1.4 Structure of thesis ......................................... 5

1.5 Thesis contributions ....................................... 6

1.6 Thesis publications ....................................... 7

2 Multichannel blind source separation 11

2.1 Introduction to BSS ......................................... 11

2.2 Independent component analysis ........................... 14

2.3 Nonnegative matrix factorization .......................... 16

2.4 Sparse component analysis ................................ 17
2.4.1 Sparseness-based approaches ........................................... 17
2.5 Computational auditory scene analysis ................................. 19
2.6 Distributed microphone arrays .......................................... 21
2.7 Summary .............................................................................. 21

3 Time-frequency masking approaches of blind source separation 23
3.1 Degenerate unmixing estimation technique (DUET) ................. 23
3.2 Beyond DUET ...................................................................... 24
3.3 Advancements into soft mask estimation ............................... 26
3.4 Frequency bin-wise masking approaches .............................. 28
3.5 Shortcomings ...................................................................... 30
3.6 Summary .............................................................................. 31

4 System overview .................................................................... 33
4.1 Signal processing and feature extraction .............................. 33
  4.1.1 Observation model ......................................................... 33
  4.1.2 STFT analysis .............................................................. 34
  4.1.3 Feature extraction ........................................................ 35
4.2 Clustering techniques .......................................................... 37
  4.2.1 HKM clustering algorithm .............................................. 37
  4.2.2 GMM clustering algorithm ............................................ 38
  4.2.3 FCM clustering algorithm .............................................. 41
4.3 Mask estimation .................................................................. 42
  4.3.1 HKM mask estimation .................................................... 43
  4.3.2 GMM mask estimation ................................................... 43
  4.3.3 FCM mask estimation .................................................... 43
4.4 Source resynthesis .............................................................. 44
4.5 Summary .............................................................................. 44

5 Experimental setup ............................................................... 47
5.1 Simulated conditions .......................................................... 47
5.2 Real room recordings ............................................. 48
5.3 Public benchmark data ............................................ 49
5.4 Algorithmic details of the HKM, GMM and FCM .................. 51
5.5 Performance evaluation criteria .................................. 53
5.6 Summary ......................................................... 55

6 Evaluations on HKM, GMM and FCM ................................. 57
6.1 Initial evaluations of FCM for mask estimation .................. 57
6.2 Simulated reverberant conditions ................................ 60
6.3 Real room recordings ............................................. 63
6.4 Benchmark data .................................................. 65
6.5 Summary ......................................................... 66

7 Advancements on the fuzzy c-means clustering ....................... 71
7.1 Introduction ....................................................... 71
7.2 Weighted FCM .................................................... 73
    7.2.1 Computation of weights .................................... 73
    7.2.2 wFCM clustering algorithm ................................ 74
7.3 Weighted spatial FCM ............................................ 75
    7.3.1 Contextual term ............................................ 75
    7.3.2 wsFCM clustering algorithm ................................ 76
7.4 Comparison of effect of weights and contextual information on mask estimation 77
7.5 Evaluations ....................................................... 79
    7.5.1 Simulated reverberant conditions .......................... 79
    7.5.2 Real room recordings ....................................... 83
    7.5.3 Benchmark data ............................................. 85
7.6 Summary ......................................................... 87

8 Watson mixture model clustering .................................... 91
8.1 Introduction ....................................................... 91
8.2 Proposed full-band WMM ........................................ 92
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2.1</td>
<td>Probabilistic model</td>
<td>92</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Estimation of parameters by EM algorithm</td>
<td>93</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Mask estimation and source recovery</td>
<td>95</td>
</tr>
<tr>
<td>8.3</td>
<td>Evaluations</td>
<td>96</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Simulated reverberant conditions</td>
<td>96</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Real room recordings</td>
<td>99</td>
</tr>
<tr>
<td>8.3.3</td>
<td>Benchmark data</td>
<td>102</td>
</tr>
<tr>
<td>8.4</td>
<td>Summary</td>
<td>102</td>
</tr>
<tr>
<td>9</td>
<td>Source number estimation via weighted adaptive FCM clustering</td>
<td>107</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction</td>
<td>107</td>
</tr>
<tr>
<td>9.2</td>
<td>Proposed source number estimation algorithm</td>
<td>109</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Calculation of weights</td>
<td>109</td>
</tr>
<tr>
<td>9.2.2</td>
<td>Adaptive FCM clustering algorithm</td>
<td>110</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Cluster quality measurement</td>
<td>111</td>
</tr>
<tr>
<td>9.3</td>
<td>Evaluations</td>
<td>113</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Experimental setup</td>
<td>113</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Source number estimation results</td>
<td>115</td>
</tr>
<tr>
<td>9.3.3</td>
<td>BSS results</td>
<td>115</td>
</tr>
<tr>
<td>9.4</td>
<td>Summary</td>
<td>116</td>
</tr>
<tr>
<td>10</td>
<td>Source number estimation via clustering of speech activity sequences</td>
<td>119</td>
</tr>
<tr>
<td>10.1</td>
<td>Introduction</td>
<td>119</td>
</tr>
<tr>
<td>10.2</td>
<td>Proposed source number estimation algorithm</td>
<td>120</td>
</tr>
<tr>
<td>10.2.1</td>
<td>Estimation of speech activity sequences</td>
<td>120</td>
</tr>
<tr>
<td>10.2.2</td>
<td>WMM for clustering sequences</td>
<td>121</td>
</tr>
<tr>
<td>10.2.3</td>
<td>EM algorithm for MAP parameter estimation</td>
<td>123</td>
</tr>
<tr>
<td>10.2.4</td>
<td>Adaptive mixture weight for source number estimation</td>
<td>124</td>
</tr>
<tr>
<td>10.2.5</td>
<td>Decision of source number</td>
<td>125</td>
</tr>
<tr>
<td>10.3</td>
<td>Evaluations</td>
<td>127</td>
</tr>
<tr>
<td>10.3.1</td>
<td>Experimental setup</td>
<td>127</td>
</tr>
</tbody>
</table>
10.3.2 Results .................................................. 128
10.3.3 Comparison of source number estimation methods ........ 129
10.4 Summary .................................................. 131

11 Conclusions and future work .................................. 133
  11.1 Conclusions ............................................... 133
  11.2 Limitations and future work ............................. 135
  11.3 Summary ................................................. 137

I Bibliography ................................................. 139

A Derivation of GMM, FCM and WMM clustering update equations for
  mask estimation ........................................... 159
  A.1 Derivation of GMM update equations .................... 159
  A.2 Derivation of FCM update equations .................... 161
  A.3 Derivation of wFCM update equations ................... 163
  A.4 Derivation of WMM update equations ................... 163

B Determination of optimal number of mixture model components .. 167
  B.1 Determination of optimal number of Gaussian components .... 167
  B.2 Determination of optimal number of Watson components ...... 176

C Calculation of speech activity sequences and WMM update equations for
  source number estimation ................................... 185
  C.1 Calculation of speech activity sequences ............... 185
    C.1.1 Model ..................................................... 185
    C.1.2 EM algorithm ........................................... 186
  C.2 Derivation of WMM update equations ................... 188
List of Figures

2.1 BSS schematic and signal notations for the convolutive mixing model. Adapted from Sawada et al. [18]. ......................................... 14
2.2 Typical CASA system architecture. Adapted from Wang and Brown [6]...... 19
3.1 General processing flow for clustering-based BSS, as used in the MENUET. Adapted from Sawada et al. [18]. .................................................. 26
3.2 Comparison of the clustering stage for mask estimation in BSS methods: (a) Full-band clustering and (b) Frequency bin-wise clustering. Adapted from Sawada et al. [18]. .................................................. 29
4.1 Spectrogram examples of various stages in the BSS scheme for a case with three microphones and four sources. (a) Sources, (b) Mixtures, (c) Time-frequency separation masks (soft) and (d) Estimated separated signals. ................. 45
5.1 Configuration of microphones and sources for experiments with simulated and real room recording data. .............................................. 49
5.2 Configuration of microphones and sources for experiments with SiSEC 2008 data. 50
5.3 Configuration of microphones and sources for experiments with SiSEC 2010 data. . 51
5.4 Location of noise in the SiSEC 2010 data. ........................................ 52
6.1 Configuration of microphones and sources for preliminary experiments using a variety of features for clustering. .................................. 58
6.2 Source separation results for $R = 50$ and $120$ cm and a microphone spacing of $4$ cm with simulated data, when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

6.3 Source separation results for $R = 50$ and $R = 120$ cm and a microphone spacing of $4$ cm with real data collected in an office environment, when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

6.4 Source separation results for $R = 50$ and $120$ cm and $R = 120$ cm and a microphone spacing of $8$ cm with real data collected in an office environment, when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

7.1 Comparison of estimated time-frequency masks for a case with three microphones and four sources. (a) FCM, (b) wFCM (c) sFCM and (d) wsFCM. Red areas indicate higher membership values (dominant source speaker) and blue areas indicate lower membership values (interference speakers). The reverberation was set to $\text{RT}_{60} = 0$ ms.

7.2 Comparison of estimated time-frequency masks for a case with three microphones and four sources. (a) FCM, (b) wFCM (c) sFCM and (d) wsFCM. Red areas indicate higher membership values (dominant source speaker) and blue areas indicate lower membership values (interference speakers). The reverberation was set to $\text{RT}_{60} = 300$ ms.

7.3 Source separation results for $R = 50$ and $120$ cm and a microphone spacing of $4$ cm with simulated data, when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
7.4 Source separation results for $R = 50$ and $120$ cm and a microphone spacing of 4 cm with real data collected in an office environment, when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation. 

7.5 Source separation results for $R = 50$ and $120$ cm and a microphone spacing of 8 cm with real data collected in an office environment, when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

8.1 Source separation results for $R = 50$ and $120$ cm and a microphone spacing of 4 cm with simulated data, when the WMM was used for mask estimation, compared against the GMM and wsFCM. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

8.2 Source separation results for $R = 50$ and $120$ cm and a microphone spacing of 4 cm with real data collected in an office environment, when the WMM was used for mask estimation, compared against the GMM and wsFCM. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

8.3 Source separation results for $R = 50$ and $120$ cm and a microphone spacing of 8 cm with real data collected in an office environment, when the WMM was used for mask estimation, compared against the GMM and wsFCM. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

9.1 Configuration of microphones and sources with labeling convention shown for source number estimation experiments.

10.1 General processing flow of the proposed source number estimation algorithm.
10.2 Depiction of clusters in the proposed speech activity sequence clustering. The speech activity sequences are clustered with the WMM into $K$ speech activity clusters, with $N$ classified as active and $K - N$ classified as 'garbage'.

10.3 Time-frequency mask representations of the resulting WMM clusters, where each mask shows the binary probability that the posterior probability sequence belongs to that cluster. Red areas indicate a 1 and blue areas indicate a 0.

10.4 Configuration of microphones and sources for source number estimation experiments.

B.1 Source separation results for $R = 50$ cm and a microphone spacing of 4 cm with simulated data when different $K$ values are used with the GMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

B.2 Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with real data collected in an office environment when different $K$ values are used with the GMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

B.3 Source separation results for $R = 50$ and 120 cm and a microphone spacing of 8 cm with real data collected in an office environment when different $K$ values are used with the GMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

B.4 Source separation results for $R = 50$ cm and a microphone spacing of 4 cm with simulated data when different $K$ values are used with the WMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
B.5 Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with real data collected in an office environment when different $K$ values are used with the WMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

B.6 Source separation results for $R = 50$ and 120 cm and a microphone spacing of 8 cm with real data collected in an office environment when different $K$ values are used with the WMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Experimental conditions for evaluations with simulated and real room recording data.</td>
<td>48</td>
</tr>
<tr>
<td>5.2</td>
<td>Experimental conditions for evaluations with SiSEC 2008 and 2010 data.</td>
<td>53</td>
</tr>
<tr>
<td>6.1</td>
<td>Source separation results for the configuration in Figure 6.1 for a reverberation time of $\text{RT}_{60} = 128 \text{ ms}$ when the HKM and FCM were used for mask estimation. Results given with respect to the SIR gain averaged across all sources.</td>
<td>59</td>
</tr>
<tr>
<td>6.2</td>
<td>Results of the Student’s $t$-test on the SDR and SIR results of Figure 6.2. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.</td>
<td>63</td>
</tr>
<tr>
<td>6.3</td>
<td>Results of the Student’s $t$-test on the SDR and SIR results of Figures 6.3 and 6.4. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.</td>
<td>64</td>
</tr>
<tr>
<td>6.4</td>
<td>Source separation results with the SiSEC 2008 and 2010 data when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.</td>
<td>68</td>
</tr>
<tr>
<td>7.1</td>
<td>Results of the Student’s $t$-test on the SDR and SIR results of Figure 7.3. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.</td>
<td>83</td>
</tr>
<tr>
<td>7.2</td>
<td>Results of the Student’s $t$-test on the SDR and SIR results of Figures 7.4 and 7.5. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.</td>
<td>86</td>
</tr>
</tbody>
</table>
7.3 Source separation results with the SiSEC 2008 and 2010 data when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface. 

8.1 Results of the Student’s $t$-test on the SDR and SIR results of Figure 8.1. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.

8.2 Results of the Student’s $t$-test on the SDR and SIR results of Figures 8.2 and 8.3. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.

8.3 Source separation results with the SiSEC 2008 and 2010 data when the GMM, wsFCM and WMM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.

9.1 The different quality measures included in this study in the estimation of the cluster quality (cf. Section 9.2.3). All measures are shown with our proposed weights $\delta_{\tau,f}$ included. The centroids, membership value and partition matrix are indicated by $v_k$, $u_{k_{\tau,f}}$ and $U$ respectively. $\bar{y}_{\tau,f}$ denotes the mean of the feature data set, and $|\cdot|$ denotes the cardinality.

9.2 Experimental conditions.

9.3 Source number estimation accuracy for various microphone and source configurations. The recordings were collected in an office environment with reverberation time $RT_{60} = 390$ ms. Results are given with respect to the accuracy measured over 20 trials. The standard deviation is shown in the brackets.

9.4 Source separation results for varying number of sources, $R = 120$ cm and a microphone spacing of 4 cm with real data collected in an office environment, when the FCM and FCM$_3$ were used for mask estimation. Results are averaged over all sources and 20 trials.

10.1 Experimental conditions.
10.2 Source number estimation accuracy for various source and microphone configurations. Six different room reverberations were tested for the cluster numbers as indicated in the leftmost column. Results are given with respect to the accuracy measured over 8 trials. .............................................. 130

B.1 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output SDR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface. .............................. 172

B.2 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output ISR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface. .............................. 173

B.3 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output SIR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface. .............................. 174

B.4 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output SAR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface. .............................. 175

B.5 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components $K$ in the WMM algorithm with respect to the output SDR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface. .............................. 180

B.6 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components $K$ in the WMM algorithm with respect to the output ISR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface. .............................. 181
B.7 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components $K$ in the WMM algorithm with respect to the output SIR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

B.8 Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components $K$ in the WMM algorithm with respect to the output SAR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.
Notations

Scalars, vectors and matrices

Scalars are denoted by regular lowercase letters, e.g. $v$, vectors by lowercase letters in the boldface, e.g. $\mathbf{v}$ and $\mathbf{v}$, and matrices by uppercase letters in the boldface, e.g. $\mathbf{V}$.

Indices

A list of indices that are often used in this thesis:

- $n$: Source index
- $m$: Microphone index
- $t$: Time index
- $\tau$: Time frame index
- $f$: Frequency bin index
- $i, k$: Cluster index

Constants

A list of constants that are often used in this thesis:

- $c$: Speed of sound
- $j$: Imaginary unit
- $\epsilon$: Arbitrarily small positive quantity
- $\lambda$: Lagrangian multiplier
Sets

For the positive integers $a$, $b$ and $c$, the following sets are used in this thesis:

- $\mathbb{R}$: Set of real numbers
- $\mathbb{R}^a$: Set of the $a$-dimensional real-valued column vectors
- $\mathbb{R}^{a \times b}$: Set of the $a \times b$ real-valued matrices
- $\mathbb{R}^{a \times b \times c}$: Set of the $a \times b \times c$ real-valued matrices
- $\mathbb{C}^a$: Set of the $a$-dimensional complex-valued column vectors
- $\Omega$: Set of time-frequency slots in short-time Fourier transform plane of interest
## Acronyms and Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSS</td>
<td>Blind Source Separation</td>
</tr>
<tr>
<td>CASA</td>
<td>Computational Auditory Scene Analysis</td>
</tr>
<tr>
<td>DMA</td>
<td>Distributed Microphone Array</td>
</tr>
<tr>
<td>DOA</td>
<td>Direction-Of-Arrival</td>
</tr>
<tr>
<td>DUET</td>
<td>Degenerate Unmixing Estimation Technique</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-Maximization</td>
</tr>
<tr>
<td>ESPRIT</td>
<td>Estimation of Signal Parameters via Rotational Invariance Technique</td>
</tr>
<tr>
<td>FCM</td>
<td>Fuzzy c-means</td>
</tr>
<tr>
<td>aFCM</td>
<td>adaptive FCM</td>
</tr>
<tr>
<td>sFCM</td>
<td>spatial FCM</td>
</tr>
<tr>
<td>wFCM</td>
<td>weighted FCM</td>
</tr>
<tr>
<td>waFCM</td>
<td>weighted adaptive FCM</td>
</tr>
<tr>
<td>wsFCM</td>
<td>weighted spatial FCM</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>HKM</td>
<td>Hard k-means</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>ISR</td>
<td>Image-to-Spatial-Distortion Ratio</td>
</tr>
<tr>
<td>LOST</td>
<td>Line Orientation Separation Technique</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum A Posteriori</td>
</tr>
<tr>
<td>MENUET</td>
<td>Multiple Sensors DUET</td>
</tr>
<tr>
<td>NMF</td>
<td>Nonnegative Matrix Factorization</td>
</tr>
<tr>
<td>SAR</td>
<td>Sources-to-Artifacts Ratio</td>
</tr>
<tr>
<td>SCA</td>
<td>Sparse Component Analysis</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>SDR</td>
<td>Signal-to-Distortion Ratio</td>
</tr>
<tr>
<td>SIR</td>
<td>Source-to-Interference Ratio</td>
</tr>
<tr>
<td>SiSEC</td>
<td>Signal Separation Evaluation Campaign</td>
</tr>
<tr>
<td>SPICA</td>
<td>Sparseness and ICA</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transform</td>
</tr>
<tr>
<td>TDOA</td>
<td>Time-Difference-of-Arrival</td>
</tr>
<tr>
<td>TIFROM</td>
<td>Time-Frequency Ratio of Mixtures</td>
</tr>
<tr>
<td>W-DO</td>
<td>W-Disjoint Orthogonality</td>
</tr>
<tr>
<td>WMM</td>
<td>Watson Mixture Model</td>
</tr>
</tbody>
</table>
Acknowledgments

This final thesis is the combined effort of many. I thank my supervisor, Roberto Togneri, for his continued supervision and assistance throughout my undergraduate studies and candidature. I also thank my co-supervisor Sven Nordholm for his supervision and discussions.

I offer grateful thanks to Tomohiro Nakatani, Nobutaka Ito and Mehrez Souden of NTT Communication Science Laboratories in Kyoto, and especially to Nakatani-san for providing me the opportunity to conduct research under the guidance of himself and colleagues. My experiences at NTT proved invaluable to my knowledge and personal growth, and contributed to the completion of this thesis.

And last but absolutely not least, I owe the most gratitude to my family. Mum, Dad, Nu and Mon, thanks for sticking by me as I pursued this even if you did not understand why I did it. Mum and Dad, your support and encouragement mean a lot to me, from my days in prep to now ÷ Nu and Mon, your bountiful pictures and conversations kept me entertained and sane, and you guys make me laugh like no other. And to my other half, Dave, thank you for being my patient and understanding best friend. It’s your turn now!
Declaration of Contribution

The work contained in this thesis is solely that of the candidate except as otherwise indicated by due acknowledgement.

Signed:  
Date: 24 April 2014
Chapter 1

Introduction

1.1 Motivation

A remarkable skill that is often taken for granted is the ability of the human cognitive system to distinguish between multiple simultaneously active speakers. The attribute to focus upon a single speaker from a complex mixture of conversation and background noise has long been a subject of importance within the speech processing community. Naturally, the implementation of this capability into an automatic machine is of great interest. As proposed by Cherry in the seminal paper [1],

“How do we recognize what one person is saying when others are speaking at the same time? On what logical basis could one design a machine for carrying out such an operation?”

Cherry challenged the engineering community to design such a machine and many an endeavor at imitation of this human ability has since been attempted. Decades after this “cocktail party problem” was coined by Cherry, researchers from a range of disciplines have made significant progress into designing such a machine. One such field of research is blind source separation (BSS), where BSS is the recovery of the original source signals from a given set of observations without \textit{a priori} information on the environment. Successful BSS systems are applied to a variety of applications in fields from acoustics to medicine, and as such, methods of BSS have important repercussions.

There are numerous methods to BSS ranging from those with a highly statistical basis such as independent component analysis, nonnegative matrix factorization and sparse
component analysis [2–5] to those motivated by the human auditory system such as computational auditory scene analysis (CASA) [6]. In particular, the auditory phenomenon of masking, where components in the perceived speech mixtures with lower energy are suppressed whilst the higher energy components are emphasized, has been of interest to researchers of BSS. This phenomenon has been realized within BSS approaches in CASA, and more recently, within the time-frequency masking approach for BSS.

Over the last 15 years the concept of time-frequency masking has been of great importance and has emerged into its own field of research. A significant benefit of these approaches is their applicability to an arbitrary number of microphones relative to sources, dissimilar to other leading methods such as independent component analysis, which require at least as many microphones as sources. The origins of the time-frequency masking approach are generally credited to Yilmaz and Rickard, authors of the degenerate unmixing estimation technique (DUET) [7]. Yilmaz and Rickard introduced a separation mask that could be applied to stereo observations to filter out the target source. The fundamental premise of this approach was the assumption of sparseness between the sources, more formally, the W-disjoint orthogonality (W-DO), which implied that for each time-frequency slot at most one source was active. Whilst this research was limited in that it was only applicable for stereo configurations in anechoic environments, this notion of W-DO sparseness combined with the design of the separation masks consequently motivated a plethora of time-frequency masking approaches to BSS [8–11]. A significant development was brought about by Araki et al. in their multiple sensors DUET (MENUE) [10], where the DUET was extended to an echoic multichannel scenario and the mask estimation was automated by the clustering of suitable location features by the hard \( k \)-means algorithm. This clustering approach to mask estimation has since been applied in numerous algorithms [11–13].

Time-frequency clustering approaches to mask estimation can be broadly categorized into either full-band, where all frequency bins are classified at once, or frequency bin-wise, where the classification occurs within each frequency bin. The majority of full-band clustering approaches cluster based on the time-difference-of-arrival (TDOA) between microphones [7,10,14–16], and in ideal anechoic conditions such features form clusters, where each cluster corresponds to a specific source. The benefit of full-band clustering for mask
estimation lies in its simplicity in that all frequencies are dealt with at once. However, as
the reverberation increases the anechoic assumption is increasingly violated, and as such the
TDOA-based features become unreliable [17]. Furthermore, TDOA-based full-band clus-
tering suffers from the spatial aliasing problem when the microphone spacing is sufficiently
large. In a bid to overcome this, frequency bin-wise methods use a two-stage structure
where the classification is performed on an instantaneous mixing model at each frequency
bin, and is therefore robust against room reverberation and the spatial aliasing problem
provided the frame in the frequency analysis is long enough to cover the main part of the
impulse response [18].

However, despite the encouraging premise of frequency bin-wise clustering, such methods
introduce additional complexity with the inherent permutation ambiguity at each frequency
bin, and the permutation alignment stage is critical to the BSS performance. Whilst nu-
merous successful techniques for alignment have been proposed, most notably by Sawada
et al. [17,18], there has not been a lot of research conducted for the full-band approach.

Time-frequency clustering approaches to BSS can also be classified depending on the
nature of the mask estimation: hard or soft. The MENUET algorithm used binary masks,
however, binary masks have been known to introduce musical noise and artifacts [11,19].
Subsequent work to the MENUET introduced soft masking schemes, however, the majority
of these were evaluated using either a stereo/linear microphone array in even- and over-
determined conditions (i.e. where the number of microphones is equal to or exceeds the
number of sources) [9,11,13,15].

As such, there is a lack of clustering-based BSS methods which focus on the under-
determined environment with an arbitrary arrangement of microphones, where the number
of sources exceeds the number of microphones. The majority of methods which do consider
this challenging environment employ frequency bin-wise techniques [18]; however, we pro-
pose to investigate advancements in full-band clustering, and to resurrect this field of BSS.
We focus our primary research direction on improving the mask estimation stage.

Furthermore, the majority of BSS methods also assume that the number of sources
is known a priori. However, such information cannot be assumed to be readily available
in realistic conditions for a blind system in reverberant conditions. For the estimation
of the number of sources for speech signals, the literature is limited with only a handful of methods [12, 20, 21]. As such, we consider the automatic detection of the number of clusters. Our first approach considers modifications to the full-band clustering within a MENUET-based framework to adaptively determine the optimal number of clusters. Our second approach uses speech activity sequences as calculated from the frequency bin-wise clustering of the microphone observations, and we use this to measure the speech activity to deduce the number of speakers.

1.2 Research objectives

The first focus of this thesis is the investigation and evaluation of improved mask estimation techniques in the framework of the MENUET BSS algorithm. Specifically, we aspire:

- To evaluate existing techniques in the field of time-frequency clustering-based multi-channel BSS.
- To investigate alternative means for mask estimation via implementation of standard common and emerging clustering algorithms.
- To design, implement and evaluate novel modifications to the standard algorithms as deemed necessary to improve mask estimation.
- To discuss and compare the different proposed mask estimation techniques in a range of simulated and real-world environments.

Our second focus is the investigation of novel techniques of blind source number estimation. Specifically, we aspire:

- To investigate and customize existing techniques of cluster number determination, and evaluate their applicability to the field of blind source number estimation within the MENUET framework.
- To investigate and design a method of blind source number estimation that can be applied to the frequency bin-wise clustering BSS framework.
1.3 Scope of research

This thesis presents a comparison between different clustering-based methods of time-frequency mask estimation for the blind separation of simultaneously active speakers. We also introduce two novel approaches to the source number estimation problem. We confine our research to simulated and real-world environments, where the data from the real-world environments includes recordings from an office environment and international public benchmark data of the Signal Separation Evaluation Campaigns (SiSEC). We consider reverberation as the main source of interference; however, we also include the effects of background environmental noise in our evaluations with the benchmark data.

Due to the fact that this thesis focuses on under-determined BSS, i.e. the case where there are fewer available microphones than speakers, the microphone arrays in our work are limited to two or three microphones. Furthermore, all source speakers are assumed to be stationary in position. Despite many real-world scenarios violating such an assumption of stationarity this lies beyond the scope of this thesis. Moreover, all algorithms in this thesis were implemented and operated via batch-processing and not currently evaluated for real-time use.

1.4 Structure of thesis

This thesis is organized as follows. We begin with an introduction to multichannel blind source separation in Chapter 2. The BSS problem is formally defined with insight into the different assumptions on mixing conditions and their applications. We cover a wide variety of approaches to multichannel BSS, from the early techniques of independent component analysis to the latest state-of-the-art developments.

Chapter 3 provides a comprehensive overview of time-frequency clustering approaches to BSS. We begin with a description of the pioneering work, followed by the algorithms consequently motivated by such work. This leads on to a summary of the recent advances in clustering-based source separation, and we conclude with a discussion of current shortcomings.

In Chapter 4 we introduce the MENUET algorithm in detail and we provide the back-
ground information necessary to facilitate a deep understanding of the algorithm, including information on the hard \( k \)-means clustering (HKM) used for time-frequency mask estimation. We also introduce the Gaussian mixture model (GMM) and fuzzy \( c \)-means (FCM) clustering algorithms for mask estimation.

The experimental setup used to realize the system in Chapter 4 is presented in Chapter 5. All aspects of the experimental setup are covered, including the software used for simulations and details of the environments in which the recordings are collected. The performance evaluation metrics are also described in this chapter.

In Chapter 6 we present the results of the evaluations of the experimental setup as described in Chapter 5. The mask estimation abilities of three algorithms, the HKM, GMM and FCM, are compared and discussed. In Chapter 7 some advancements on the FCM are introduced and evaluated. We describe the advancements and modifications to the FCM and present the results of evaluations.

Chapter 8 introduces the Watson mixture model (WMM) for mask estimation, and we describe the details of the WMM and its application within the clustering-based source separation framework. We also present results of experimental evaluations.

In Chapters 9 and 10 we move beyond source separation and focus upon source number estimation. These chapters present two novel algorithms: the first based on an adaptive FCM algorithm for source number estimation, and the second employs a WMM for full-band clustering of speech activity sequences.

Finally, we conclude the thesis in Chapter 11 with a summary of the major contributions. This chapter also discusses the limitations of the research and provides suggestions to overcome such limitations. We also offer insight into future avenues of research.

1.5 Thesis contributions

The contributions of this thesis are:

- Investigation of the FCM for mask estimation with MENUET algorithm: we evaluate the ability of the FCM for mask estimation within the MENUET framework. We also implement the GMM and compare against the FCM. The algorithm is introduced
in Chapter 4 and the results are presented in Chapter 6. The work is also partially published in [22–25].

- Extension of the standard FCM via weights and contextual information: we extend the FCM via the integration of reliability weights and contextual information to improve the quality of the estimated masks. The algorithm is introduced and evaluated in Chapter 7. The work is also partially published in [26–28].

- Investigation of the WMM for mask estimation with MENUET algorithm: we evaluate the ability of the WMM for mask estimation within the MENUET framework. The algorithm is introduced and evaluated in Chapter 8. The work is also partially published in [28, 29].

- Development and investigation of an adaptive FCM clustering technique for automatic detection of the number of sources: we introduce a novel and customized version of an adaptive FCM clustering technique to automatically count the number of sources. The algorithm is introduced and evaluated in Chapter 9. The work is also partially published in [30].

- Development and investigation of the WMM for full-band clustering of speech activity sequences within a frequency bin-wise clustering framework: we introduce the clustering of speech activity sequences via the WMM for source number estimation. This is introduced and evaluated in Chapter 10. The work is also published in [31].

1.6 Thesis publications

The peer-reviewed publications arising from this thesis are listed below. For first-authorship papers, the candidate was responsible for the conception and implementation of ideas, as well as the construction of the manuscript. For second-authorship, the candidate was responsible for the conception of the main ideas, the supervision of the first author in the implementation, and the co-construction of the manuscript. All manuscripts were commented on by the co-authors. The work arising from this thesis comprises original contributions except where due reference is made.
CHAPTER 1. INTRODUCTION

Book chapters


Journal articles


Conference proceedings


Chapter 2

Multichannel blind source separation

This chapter provides an introduction to multichannel BSS techniques. There are a wide range of methods for BSS from those with deep statistical roots to those which are inspired by the human auditory system. We provide insight into such methods and the underlying assumptions of each. The rest of the chapter is organized as follows. We begin with the conception and introduction of the BSS problem in Section 2.1. We continue with a presentation of early BSS methods such as independent component analysis in Section 2.2, and related techniques in Sections 2.3 and 2.4. Methods of computational auditory scene analysis are discussed in Section 2.5 and the recent BSS field of distributed microphone arrays in Section 2.6.

2.1 Introduction to BSS

Whilst this thesis focuses on the use of BSS for the separation of speech signals, techniques of BSS are applied far beyond speech applications. For example, the separation of heart and lung sounds, the preprocessing of electroencelograms for Alzheimers detection, seismic processing, the extraction of fault information from engine fault data, and financial time series analysis [32–38]. Applications of BSS within the audio field include hands-free telecommunication, automatic meeting note-taking, cochlear implants, hearing aids and applications
in automobiles environments [39].

The genesis of source separation for audio applications can be attributed to the work of Hérault and Jutten by whom the problem was first formally modeled [40]. Early examples of BSS in speech processing subsequently arose [2,41–43], and over the last three decades, the interest in BSS has risen sufficiently to be considered a comprehensive research field in its own right.

The general BSS problem can be formally stated as follows. Given $M$ microphone observations and $N$ sources related by an unknown $M \times N$ mixing matrix $A$, we have the relation

$$X = AS,$$

where $X$ denotes the mixture matrix and $S$ the matrix of sources. The mixture matrix $X$ is an $M \times P$ matrix, where $P$ denotes the number of time samples, and consists of the microphone observations at each time frame as

$$X = \begin{bmatrix}
  x_1(t_1) & x_1(t_2) & \ldots & x_1(t_P) \\
  x_2(t_1) & x_2(t_2) & \ldots & x_2(t_P) \\
  \vdots & \vdots & \ddots & \vdots \\
  x_M(t_1) & x_M(t_2) & \ldots & x_M(t_P)
\end{bmatrix}. \quad (2.2)$$

The source matrix $S$ is an $N \times P$ matrix

$$S = \begin{bmatrix}
  s_1(t_1) & s_1(t_2) & \ldots & s_1(t_P) \\
  s_2(t_1) & s_2(t_2) & \ldots & s_2(t_P) \\
  \vdots & \vdots & \ddots & \vdots \\
  s_N(t_1) & s_N(t_2) & \ldots & s_N(t_P)
\end{bmatrix}. \quad (2.3)$$

The BSS problem deals with the estimation of the $M \times N$ mixing matrix $A$.

The selection of a suitable BSS method must consider the assumption of the mixing model. The mixing model can be classified into one of three models: instantaneous, anechoic and echoic (convolutive). The difficulty of the BSS problem is often dependent on the mixing assumptions of the signals in the acoustic environment.

The most primitive mixing model is the instantaneous model which assumes that the
signals arrive at the microphones instantaneously with different signal intensities or attenuation factors. Each microphone effectively receives a linear combination of the signals. Hence, each observation at the \( m \)th microphone can be expressed as

\[
x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t)
\]

where \( a_{mn} \) denotes the attenuation parameter, or mixing parameter, of the \( n \)th source signal received at the \( m \)th microphone. The mixing matrix \( A \) in the instantaneous case simply consists of the set of attenuation factors, i.e. \( \{a_{11}, \ldots, a_{1N}, \ldots, a_{M1}, \ldots, a_{MN}\} \).

Whilst the instantaneous mixing case is that which most early BSS algorithms were designed for and is useful for theoretical derivations, the assumptions are restrictive in realistic environments. An extension of the instantaneous model is the anechoic mixing model, where not only the different signal intensities between the microphones are considered but also the different time delays between microphones. The source signals assume single non-dispersive paths from each source to each microphone, and each microphone then captures a linear combination of scaled and time-delayed source signals. This is expressed as

\[
x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \delta_{mn})
\]

where \( \delta_{mn} \) represents the delay between the \( n \)th source and \( m \)th microphone. The mixing matrix estimation then amounts to the estimation of the set of attenuation and delay parameters, i.e. \( \{a_{11}, \ldots, a_{1N}, \ldots, a_{M1}, \ldots, a_{MN}\} \) and \( \{\delta_{11}, \ldots, \delta_{1N}, \ldots, \delta_{M1}, \ldots, \delta_{MN}\} \).

However, real-world acoustical paths between each source and microphone lead to convolutive mixing as a result of the multipaths generated by the sound reflections off space, and also due to the time delays between the microphones. Therefore, at each microphone, the received signal is a convolution of the original source signal and the impulse response between each signal and the microphone. We refer to this mixing model as the echoic mixing model. In this instance, the observation at each microphone is modeled as a convolutive sum as

\[
x_m(t) = \sum_{n=1}^{N} \sum_{p} h_{mn}(p)s_n(t - p)
\]
where $h_{mn}(p)$ represents the room impulse response between the $n^{th}$ source and the $m^{th}$ microphone. Therefore, the mixing matrix estimation for the echoic mixing model is equivalent to the estimation of the impulse responses between all $N$ sources and $M$ microphones, i.e. $\{h_{11}, \ldots, h_{1N}, \ldots, h_{M1}, \ldots, h_{MN}\}$.

Whilst both the anechoic and echoic mixing models involve a convolutive mixing procedure, the anechoic mixing case characterizes the impulse response through the attenuation and delay parameters. In this thesis we deal with echoic mixing models exclusively, and we refer to the echoic mixing model as the convolutive mixing model henceforth.

The convolutive mixing model is structurally much more challenging than the instantaneous model in (2.4) due to the delay and reflections. Typically in reverberant conditions at a sampling rate of 8 kHz, the impulse response $h_{mn}(p)$ can have thousands of taps which adds to the difficulty.

We depict the BSS problem and signal notations for the convolutive mixing model in Figure 2.1.

![Figure 2.1: BSS schematic and signal notations for the convolutive mixing model. Adapted from Sawada et al. [18].](image)

### 2.2 Independent component analysis

Early works in BSS were largely focused on the instantaneous mixing model [2, 41, 42], and as one of the earliest approaches to BSS, methods of independent component analysis (ICA) model the microphone observations as linear combinations of the source signals.
ICA is fundamentally motivated by the central limit theorem: the distribution of a sum of independent distributions tends to be Gaussian, with some conditions. In ICA, the sum refers to each microphone observation, and the independent distributions are the source signals.

This therefore imposes conditions upon any BSS system to which ICA is to be applied: that the mixing system be invertible, the constituent source signals are statistically independent and also as non-Gaussian as possible. Should these conditions be satisfied, effective separation can ensue and as such, ICA is generally sufficient in the separation of instantaneous mixtures due to the super Gaussian nature of speech [41]. The independent components are generally derived by the optimization of a contrast function, which provides a measure of independence between the components. For example, the minimization of the mutual information [41], maximum entropy [42] and non-Gaussianity [44].

The ICA has since been extended beyond the instantaneous mixing model, for example, its suitability for time-domain convolutive BSS has been investigated [45,46]. However, this is significantly more difficult than instantaneous ICA and can be computationally expensive due to the convolution operations. Alternatively, complex-valued ICA for instantaneous mixtures can be applied to each frequency bin individually [47–49]. However, whilst the latter employs the much simpler instantaneous ICA at each frequency bin, it results in the inevitable frequency permutation ambiguity problem over the frequency bins. This can be solved by permutation alignment of each bin [49].

The drawbacks of ICA include the scaling ambiguity, where the source signals are not properly scaled; the permutation ambiguity, where the correct source order cannot be recovered; and the inherent restriction on the relative number of microphones and sources, where the number of sources cannot exceed the number of microphones due to the matrix inversion in ICA (i.e. only even- or over-determined configurations). As such, ICA cannot be applied to all BSS scenarios, and we thus explore other approaches for separation.
2.3 Nonnegative matrix factorization

ICA can be viewed as a matrix factorization problem: the factorization of the observation matrix $X$ into the mixing matrix $A$ and source matrix $S$, where the rows of $S$ are as statistically independent as possible, by the search for the most non-Gaussian directions [50,51]. As just discussed, the applicability of ICA is conditional on whether all conditions are satisfied. However, nonnegative matrix factorization (NMF) seeks to factorize the observation matrix $X$ under less stringent conditions than ICA.

NMF is an emerging method in audio source separation [52]. In such approaches, an $M \times N$ observation matrix $X$ (cf. (2.1)) is factorized into two nonnegative matrices as

$$X = WV,$$  

(2.7)

where $W$ is an $M \times B$ matrix and $V$ is a $B \times N$ matrix. $B$ denotes the number of bases, which are generally empirically determined. NMF allows the exclusive additive combination of multiple basis signals to represent original signals, which implies that no subtractions can occur. This is in line with the intuitive notion of parts-based representations, i.e. the combination of components to form a whole signal or image [52].

NMF operates under the assumption that all matrices are nonnegative. For the application of NMF to BSS, the operations are performed in the magnitude spectrum domain, for example by squaring the short-time Fourier transforms (STFT) [53]. The factorization in (2.7) is executed via the minimization of the distance or divergence between $X$ and its factored form in $WV$. Suitable divergences include the $\alpha$ divergence [52], Itakura-Saito divergence [53] or the generalized Kullback-Leibler divergence [54].

The $B$ bases encode speaker information and can also reflect the speakers’ speech patterns [4]. Whilst many methods of NMF were initiated for monaural source separation, multichannel processing has recently been proposed and explored to exploit the spatial properties captured in multiple microphones [53]. For the purposes of speech separation, upon estimation of the factorized matrices the bases can be clustered according to properties which exhibit similarities. In the case of convolutive multichannel BSS, often the phase difference between microphones is suitable and can be clustered with standard clustering
algorithms or customized top-down or bottom-up clustering approaches [53]. The sources can then be recovered using Wiener filters.

### 2.4 Sparse component analysis

A related area of research to ICA and NMF factorizes $X$ by searching for the matrix $S$ (cf. (2.1)) with as many zeros as possible, i.e., searches for a sparse representation of $S$. This matrix factorization is referred to as sparse component analysis (SCA). Methods of SCA seek to find the sparsest representation of the source signals in an over-complete basis through the optimization of a suitable criterion such as the $l_0$ or the $l_1$ minimization criterion [55].

In SCA, a sparsifying linear transform is applied to the mixture matrix $A$, such as the orthogonal wavelet or wavelet packet transforms, or STFT. After this transformation, the mixing matrix $A$ is estimated with a suitable technique such as via clustering the mixture correlation scatter plot [56] or with natural gradient ICA approaches [57]. A key hypothesis of the former approach is that for each point in the scatter plot, at most one source signal is dominant. Upon estimation of $A$ the source representations are estimated based on the sparseness assumption; for example, through estimation of the source signal coefficients via $l_q$ optimization for $q < 1$ [5], $l_1$ optimization [55] or a ML-based optimization [58].

Such sparseness-based approaches to source separation consequently led to the investigation of techniques which exploited the sparseness of speech. We discuss these in detail below.

#### 2.4.1 Sparseness-based approaches

SCA relies on the sparseness of the constituent source signals: the assumption that at most one source is dominant at each data point, and that the contribution of the others are insignificant. This notion of sparseness has been of increasing interest in the BSS research field due to its applicability to all BSS scenarios, in particular the under-determined case. Bermond and Cardoso specified that unlike even- and over-determined configurations, the under-determined BSS problem is not equivalent to mixing matrix estimation, and the
importance of sparse representations was emphasized [59]. There exists several definitions for sparseness in the literature; for example, to contain as “many zeros as possible” [50], that only a small number of source components different significantly from zero [58], or more quantifiable measures such as the fourth moment or kurtosis [60].

As also used in SCA, a sparse representation of speech mixtures can be acquired through the projection of the signals onto an appropriate basis, such as the Gabor, Fourier or Wavelet basis. In particular, Jourjine et al. initiated the concept of the W-DO of speech signals to formulate a method of BSS for an arbitrary number of sources with two microphones [14]. In this approach W-DO is assumed to hold if for a given window the supports of the windowed Fourier transforms of the source signals are disjoint. Despite speech signals often violating this assumption, satisfactory separation is achievable. A definition of approximate W-DO was then formalized by Rickard et al. where there was required to be only a percentage of the window for a given source to dominate by a specified threshold [61]. From this assumption of W-DO the mixing parameters were able to be estimated, from which the time-frequency separation masks were constructed. This eventually led to the formulation of DUET [7], which consequently motivated a plethora of time-frequency masking techniques based upon the sparseness as defined by the W-DO in the STFT domain in forthcoming years [8, 10–12, 62].

The DUET algorithm successfully recovers the original source signals from stereo microphone observations using estimates of the relative attenuation and delay parameters, where these estimates are used to construct time-frequency separation masks. Such time-frequency masks, applied to the microphone observations, are able to filter out the time-frequency slots classified as dominant to retrieve the target signal. This mimics the human auditory phenomenon of masking where the dominant signals tend to mask out the weaker signals. This notion of time-frequency masking is also extensively made use of and linked to computational auditory scene analysis.
2.5 Computational auditory scene analysis

Whilst the BSS approaches discussed thus far exploit mathematical assumptions and properties of the signals, another approach to source separation revolves around the study of the human auditory perception process. Auditory scene analysis, or ASA, is the term used to describe the phenomenon which allows humans to produce perceptual representations of different sources in an acoustic mixture [63]. It was proposed by Bregman that this scene analysis was enabled by the automatic organization of acoustic mixtures into streams, where each stream corresponds to a different sound source. This organization consisted of two stages: segmentation and grouping, where segmentation is the decomposition of the acoustic scene into segments, and grouping is the classification of the segments into streams according to their likelihood to have originated from the same stream, where the likelihood is computed using characteristic cues. The concept of ASA is summarized in the representative architecture model in Figure 2.2, as designed by Bregman [63].

The evident ability of ASA led to the investigation of automatic means of achieving ASA. Computational ASA (CASA) is the realization of ASA by computational means, and methods of CASA incorporate some degree of auditory modeling. One of the first CASA approaches to source separation was for single channel conditions, where the mixture was passed through a bank of bandpass filters to perform frequency analysis [64]. The interpeak interval in each filter channel was calculated to estimate the pitch periods of each source,
and each source was characterized by a state in a Markov model. This has since been extended to a more advanced system, for example to model the auditory periphery at a closer level [65] and to also allow resynthesis [66].

Other approaches to CASA attempt to emulate the human hearing system by employing perceptually-motivated psychoacoustic cues, as calculated by the microphone observations, in a signal processing framework. These include monaural (one ear) and binaural (two ears) cues. Monaural cues are a result of the acoustical properties of the outer ear and are significant for elevation source localization, whilst binaural cues are important for azimuthal source localization [67]. Examples of psychoacoustic binaural cues include interaural time differences and interaural phase differences, which have been successfully applied for BSS in a framework similar to the DUET algorithm [68]. Other examples of perceptually-motivated cues suitable for source separation include pitch, harmonicity, onset/offset and spatial locations.

At the heart of many CASA approaches lies the fundamental use of the auditory phenomenon of masking, where the dominant signals (i.e. those with stronger energy) mask out the weaker signals. As such, the formal definition of the goal of CASA has been proposed as the estimation of an ideal binary mask (IBM) for source separation [69]. After the transformation of the mixtures into the STFT domain the target source and interfering source energy are computed at each time-frequency slot \((\tau, f)\), and the IBM is defined as:

\[
M_{\text{IBM}}(\tau, f) = \begin{cases} 
1 & \text{for } 10 \log_{10} \left( \frac{\|s_t(\tau, f)\|^2}{\|s_i(\tau, f)\|^2} \right) > \epsilon, \\
0 & \text{otherwise.}
\end{cases}
\]  

where \(\|s_t\|^2\) is the energy of the target source, \(\|s_i\|^2\) is the energy of the interfering sources and \(\epsilon\) is a pre-defined threshold value. \(\epsilon\) is usually set to zero to satisfy a 0 dB criterion [69].

This psychoacoustic concept of masking has crossed over into other fields including image and source separation. Of particular interest is the time-frequency clustering approach to mask estimation, of which we present a comprehensive overview in the following chapter.
2.6 Distributed microphone arrays

Recent advances in BSS have employed the use of distributed microphone arrays (DMA) [70, 71]. In contrast to standard microphone arrays with co-located elements, a DMA consists of a set of microphones embedded in nodes, where each node is considered as a local processing unit, and each node communicates with the others to share information. Techniques that use the DMA have the advantage of scalability and spatial coverage [72]. The design of DMA algorithms is generally three-stage: first, the definition of tasks for each node to fulfill, second, the design of a communication protocol between the nodes, and finally, the processing of the exchanged data to achieve a global processing goal [72].

Existing techniques of BSS have been integrated into the DMA framework. For example, a distributed ICA algorithm was proposed, where each node performed ICA adaptations that were exchanged between the nodes with regularizing factors based on speech sparseness [73]. Another method of DMA also integrated the ICA into the array, where each node detected its local neighboring sources, performed a local separation with ICA and then shared the information between the nodes [74]. However, the use of ICA implied that each node was required to be over-determined, i.e. more microphones at each node than sources. This can be impractical in situations where not many microphones are available.

The sparseness time-frequency masking approach has also been applied within the DMA framework [72]. Each node calculated the posterior probabilities of the source clusters using features extracted from its own microphone observations, and this information was shared with the other nodes. The results of the individual probability calculations were combined to yield separation. The work was extended with the inclusion of inter-node and intra-node spatial features to form complementary features [75]. Despite the encouraging results of DMA, it is not suitable to some applications where spatial limits are in place.

2.7 Summary

This chapter presented an introduction to methods of BSS. We formalized the BSS problem and described three different mixing models. These mixing models along with the suitable assumptions to be made are taken into consideration when selecting the best BSS method.
For the instantaneous mixing model, ICA is sufficient for separation. However, the instantaneous model is largely unrealistic in practice and we thus seek alternative methods of separation. Advancements on the ICA which relax the instantaneous assumption include NMF, which is a recent topic of interest in BSS, and SCA, which assumes the sparseness of speech. Other methods which assume the sparseness of speech rely on time-frequency masking for separation, which draws its inspiration from the human auditory-motivated research area of CASA.

The primary benefit of time-frequency masking approaches to BSS is its applicability to an arbitrary configuration of microphones and sources, namely, the under-determined scenario. Given the large scope of research which followed the pioneering work in the DUET algorithm, we present an overview into such methods in the following chapter.
Chapter 3

Time-frequency masking approaches of blind source separation

In Chapter 2 we introduced BSS and various approaches to solve the problem. As discussed, the sparseness-based approach is an active research topic with significant performance benefits over traditional BSS methods. Of particular interest are the time-frequency masking approaches to BSS for their versatility and robust performance. As such, this chapter presents a review of time-frequency masking approaches to BSS commencing with its origins to the latest state-of-the-art algorithms. The rest of the chapter is organized as follows. In Section 3.1 we introduce the pioneering work behind the time-frequency masking approaches, and in Section 3.2 we describe the subsequent advancements. Section 3.3 details the development of soft mask estimation, while Section 3.4 introduces the latest frequency bin-wise approach to mask estimation. We conclude with the various shortcomings of current work in Section 3.5.

3.1 Degenerate unmixing estimation technique (DUET)

The advent of the time-frequency masking approach for BSS is generally credited to the authors of the DUET algorithm [7]. The DUET, designed for anechoic mixtures, transforms
CHAPTER 3. TIME-FREQUENCY MASKING APPROACHES OF BLIND SOURCE SEPARATION

The microphone observations into the STFT domain, where they are assumed to follow W-DO. The relative attenuation and delay mixing parameters at each time-frequency slot are then estimated from the microphone observations. If the sources exhibit sufficient W-DO, a histogram of the mixing parameter estimates in the parameter space will yield $N$ distinct peaks. As such, the mixing parameter estimates are used to compute a weighted histogram, the peaks of which are used for time-frequency separation mask estimation.

3.2 Beyond DUET

The formulation of W-DO and the development of the DUET algorithm consequently opened a new branch of research. Among the first extensions to the DUET was the time-frequency ratio of mixtures (TIFROM) algorithm, which relaxed the condition of W-DO and instead assumed that a source occurs alone in a small area of adjacent time-frequency slots in the feature space [8]. Therefore, in contrast to the DUET which computed the mixing parameter estimates at each time-frequency slot, the TIFROM computed the mixing parameter estimates from a small area of adjacent time-frequency slots. However, unlike the DUET which was designed for the anechoic mixing model, the TIFROM was limited to the linear instantaneous mixing model. As such, the TIFROM is highly impractical for acoustic environments.

The DUET was extended to the anechoic multichannel case with the adoption of a two stage optimization approach [62]. Firstly, the attenuation and delay mixing parameters were estimated by clustering a set of feature vectors, where the feature vectors were constructed from the STFT representations of the microphone observations. Secondly, the sources were estimated by an $l^q$ minimization based algorithm, where $q < 1$. This resulted in superior separation performance over the DUET, and also allowed more than two mixtures. However, it was still restrictive in its applicability to exclusively anechoic conditions.

The DUET was then extended to echoic conditions with the integration of the estimation of signal parameters via rotational invariance technique (ESPRIT) method [76] to form the DUET-ESPRIT algorithm [9]. The ESPRIT technique is a direction-of-arrival (DOA) method of estimation, which allowed a slight weakening on the W-DO assumption and
extension to a multichannel configuration. However, it was restricted to a linear microphone arrangement and was subjected to front-back confusions due to the natural constraint in spatial diversity from the microphone observations.

A related avenue of research by Araki et al. composed a two-stage algorithm which combined the sparseness principle in DUET with the established ICA to yield the sparseness and ICA (SPICA) algorithm [77], applicable to the convolutive mixing model. First, the sparseness of speech was used to estimate a binary mask which was applied to all microphone observations to remove the dominant source. The binary masks were calculated using the DOA as estimated from the phase difference between observations. After application of the mask, the remaining observation signals were therefore expected to be mixtures of \( N - 1 \) signals, and assuming the correct relative number of microphones to sources (cf. Section 2.2), frequency-domain ICA was then performed on the observations to extract the remaining source signals. Despite the promising separation results of the SPICA algorithm, it imposed a restraint on the number of sources relative to the number of microphones due to the ICA, and the algorithm was only proposed and evaluated for the specific configuration of \( M = 2 \) and \( N = 3 \). Furthermore, the frequency-domain ICA was also subject to the permutation alignment problem, which required an additional alignment stage in the SPICA.

However, despite such restrictions, this SPICA algorithm led the authors to expand their research to nonlinear microphone arrays with the introduction of the clustering of normalized observation vectors [19, 78]. This was also applied to an arbitrary number and arrangement of microphones and termed the multiple sensors DUET (MENUET) [10]. Whilst remaining similar in spirit to the DUET, the research proposed a novel normalization of the microphone observations in order to promote the formation of clusters. The normalization was performed in such a way that the normalized vectors embodied spatial information, therefore implying that each cluster was an estimate of the location of the corresponding source. Furthermore, the mask estimation was automated by application of the HKM clustering algorithm, where the final cluster memberships were interpreted as an indication of the dominant slots of each source in the time-frequency plane, and were consequently used as the separation masks. Evaluations in reverberant under-determined environments confirmed the viability of the MENUET scheme. The general scheme of the
CHAPTER 3. TIME-FREQUENCY MASKING APPROACHES OF BLIND SOURCE SEPARATION

3.3 Advancements into soft mask estimation

Whilst the masks estimated in the MENUET scheme were binary, a number of algorithms also considered soft clustering techniques for mask estimation. O’Grady and Pearlmutter proposed the soft line orientation separation technique (soft-LOST) and LOST algorithms to estimate an arbitrary number of sources from an arbitrary number of mixtures [79,80]. In this line of work, the density of the observations was assumed to be drawn from a Laplacian mixture model, where each mixture had an associated orientation vector on which each source was centered upon. The expectation-maximization (EM) algorithm was used to estimate the orientations of the sources.

In a similar approach to the soft-LOST and LOST algorithms, Araki et al. explored an EM algorithm for source separation where the histogram of the DOA was modeled with a GMM [12]. However, in an innovative approach to prevent each cluster being modeled by more than one Gaussian and to also enable source number estimation, a sparse distribution modeled by the Dirichlet distribution was used as the prior for the Gaussian mixture weights. This Dirichlet prior captured the assumption of sparseness in the directions of the source signals. The method was successful in real environments with low reverberation times.

Mandel et al. also used the EM algorithm with a probabilistic approach to source local-
CHAPTER 3. TIME-FREQUENCY MASKING APPROACHES OF BLIND SOURCE SEPARATION

ization and separation for stereo recordings [15]. The psychoacoustic cue of interaural phase difference was used for clustering under the assumptions that each time-frequency slot was dominated by a single source, and that a single delay/amplification caused the difference between the two receiving microphones (or ears) at a particular slot. The estimation of the probability distributions then uniquely encompassed both the direction of the source’s origin, and the time-frequency slots associated with that source. This approach could be applied for up to three sources with favorable localization and separation results.

Similarly, Izumi et al. investigated simultaneous source separation and localization in an EM framework to concurrently estimate the separation masks and DOA [16]. However, in contrast to the earlier work by Mandel et al. the EM algorithm was applied in the signal domain rather than a feature domain. A common objective function was optimized with the E-step corresponding to source localization and the M-step corresponding to source separation. The algorithm was evaluated in simulated under-determined conditions with successful results.

A novel line of research was proposed by Ito et al. to model the normalized observation vectors by a WMM with time-varying mixture weights [13]. Furthermore, inspired by previous work by Araki et al. [12], a prior based on the Dirichlet distribution was imposed on the model to prevent the degradation in separation performance from the time-varying mixture weights. However, this prior differed from that employed by Araki et al. as it was non-sparse. In contrast to other work which used the WMM for BSS [81], all frequency bins were simultaneously processed, thereby avoiding any frequency bin-wise permutation ambiguities (discussed below in Section 3.4). Although the experimental results of this algorithm were highly promising, they were only evaluated for a stereo even-determined case. Nevertheless, the introduction of the WMM for full-band clustering was significant and this is further investigated in Chapter 8 of this thesis.

Whilst previous mask estimation techniques made use of the k-means and EM-based clustering algorithms, the suitability of the FCM clustering for the DOA was also explored [11,82]. In this approach, the soft partitioning in the FCM was suggested to be preferable to hard clustering due to the inherent ambiguity surrounding the membership of each time-frequency slot to a cluster, where examples of contributing factors to ambiguity include the
effects of reverberation and environmental (background) noise. However, the investigations in these studies were limited to simulated over-determined configurations with a linear microphone array.

### 3.4 Frequency bin-wise masking approaches

Many of the aforementioned time-frequency clustering approaches to BSS rely on the clustering of spatial location features estimated from the ratio of microphone observations, particularly those based on the phase ratio, i.e. DOA/TDOA [10–12, 83]. However, such features as these have a tendency to be unreliable in echoic conditions due to the violation of the linear phase assumption in the impulse response $h_{mn}$ of the anechoic mixing model. Furthermore, when such features are estimated from widely-spaced microphones, spatial aliasing in the higher frequencies is inevitable [84, 85].

However, if the time-frequency clustering is executed in a frequency bin-wise manner then the clustering is not based on the DOA/TDOA but rather based on a frequency bin-wise instantaneous model [17]. Therefore, the assumption of the linear phase in the impulse response is not required and the microphone spacing is arbitrary [17]. Nonetheless, the cluster class order in each frequency bin is not necessarily equivalent and as such, a permutation ambiguity is unavoidable. This is alleviated using a permutation alignment procedure. Standard techniques of permutation alignment include those based on the DOA estimates of the bin-wise separated sources [49, 86], however these approaches are vulnerable to the effects of spatial aliasing. There are also techniques based on the temporal envelopes of the bin-wise separated sources [49, 87], however, these approaches require the data to be of a certain length [13].

In a bid to overcome these limitations in the existing approaches to permutation alignment, Sawada et al. proposed a novel approach to bin-wise clustering and alignment [17, 18]. The normalized microphone observation vectors were clustered in a frequency bin-wise manner based on the line orientation idea of O’Grady and Pearlmutter [79, 80], where the resulting posterior probabilities were representative of the speech activity in each cluster. The separated frequency bin components classified as originating from the same source were
CHAPTER 3. TIME-FREQUENCY MASKING APPROACHES OF BLIND SOURCE SEPARATION

Figure 3.2: Comparison of the clustering stage for mask estimation in BSS methods: (a) Full-band clustering and (b) Frequency bin-wise clustering. Adapted from Sawada et al. [18].

then grouped together based on the optimization of an elaborate score function using the correlation coefficients of the posterior probability sequences [18]. The frequency bin-wise clustering approach is summarized and compared against the full-band clustering in Figure 3.2.

The work of Sawada et al. consequently motivated other frequency bin-wise clustering approaches. For example, Reju et al. proposed the clustering of Hermitian angles between sample and reference feature vectors with the HKM and FCM algorithms [88]. The resulting masks were then clustered a second time for permutation alignment, and the algorithm was evaluated for various linear microphone arrangements with encouraging results.

A novel solution to the permutation alignment problem was proposed by Araki et al. by the simultaneous clustering of two features in a frequency bin-wise manner [83]. The first feature was a source locational feature of the phase difference between microphones to facilitate source separation, and the second feature was the spectral temporal envelope to enable permutation alignment. This algorithm also customized the EM cost function for source number estimation; however, the algorithm was only verified for even-determined stereo configurations. Furthermore, the source number estimation ability was subject to
the effects of spatial aliasing when the microphone spacing was increased [83].

The use of directional statistics for BSS was proposed in the context of the WMM by Tran Vu and Haeb-Umbach [81]. In this line of work, normalized observation vectors were modeled with a complex-valued WMM and clustered in a frequency bin-wise manner. The permutation alignment problem was solved by using the recent correlation coefficient based approach of Sawada et al. [18]. Despite the encouraging results, the work was only evaluated in an over-determined configuration with as many as eight microphones. This may be impractical in some applications of BSS. However, this work motivated further research as mentioned above in Section 3.3 by Ito et al. and also by Souden et al. where the WMM was employed for online ML-based separation [89], and each frequency bin was similarly aligned using the approach of Sawada et al. [18]. The online ML-based separation was realized by a recursive EM algorithm where the EM model parameters were adapted at each time frame to allow for variability in the speaker location. Encouraging results were achieved in an over-determined setting with a uniform linear array and two sources, where one source moved location throughout the recordings.

3.5 Shortcomings

The advancements in time-frequency masking approaches of BSS beyond the DUET have their considerable benefits as discussed in this chapter. Such advancements include the extension of the DUET to convolutive and multichannel conditions, the use of soft mask estimation methods and the introduction of frequency bin-wise clustering methods. Despite such advancements in this field of BSS, there remains several shortcomings which must be addressed. For example, many of the soft mask estimation techniques discussed in Section 3.3 were limited to linear arrays with restrictions on the relative number of microphones to sources. Furthermore, despite the advantages and added robustness against reverberation and spatial aliasing of the frequency bin-wise clustering methods, such methods introduce additional complexity due to the compulsory permutation alignment stage. Therefore, full-band clustering approaches have the advantage over frequency bin-wise approaches in that the full-band clustering eliminates any requirement for frequency bin alignment. However,
full-band approaches which are able to handle nonlinear and arbitrary microphone arrangements are preferable due to their applicability in diverse settings, such as the MENUET algorithm by Araki et al. [10].

However, the simplicity encapsulated in the MENUET inevitably presents its own limitations. Most significantly, the HKM clustering utilized for mask estimation is not highly robust in the presence of outliers or interference in the data. This often leads to non-optimal localization and partitioning results, particularly for reverberant mixtures [90,91]. Furthermore, binary masking schemes have been shown to impede upon the separation quality due to musical noise distortions, and it was suggested that soft masking approaches bear the potential to significantly reduce the musical noise at the output [19]. This may be attributed to the fact that when a hard partitioning approach is implemented, abrupt changes in the recovered source estimate may introduce artifacts when resynthesized in the time domain.

As examined by Kühne et al. the FCM clustering for mask estimation was suggested to be preferable to binary clustering methods due to the inherent ambiguity surrounding the membership of time-frequency slots to a cluster, where examples of contributing factors to ambiguity include the effects of reverberation and environmental (background) noise [11]. However, the investigations to date which employ the FCM, as with many other soft masking approaches in the literature, have been restricted to a linear and/or over-determined microphone arrangement.

As such, the investigation of alternative mask estimation techniques in the context of the MENUET algorithm has potentially significant repercussions. Should a superior clustering technique be found, it would enable the separation of sources in an under-determined configuration with a nonlinear microphone array.

3.6 Summary

In this chapter, we presented an overview of time-frequency clustering approaches to BSS. We began with a description of the pioneering DUET algorithm and continued with the many algorithms that it consequently motivated. Whilst the initial algorithms such as TIFROM and DESPRIT imposed restrictions on the mixing model and/or microphone
arrangement, the approaches of Araki et al. moved time-frequency clustering approaches towards less restrictive conditions [10, 78].

The subsequent research saw the introduction of soft masking techniques using the EM algorithm, where the microphone observations and features were modeled with Gaussian, Laplacian and Watson mixture models. The recent advances in time-frequency clustering approaches to BSS have moved towards frequency bin-wise clustering for robustness against room reverberation and immunity against spatial aliasing. However, the frequency bin-wise clustering is at the expense of permutation ambiguity at each frequency bin. Whilst several approaches have successfully been proposed to overcome this problem [17, 18, 88], permutation ambiguities can be avoided by employing a full-band clustering technique.

As such, we focus our research onto an established full-band source separation technique, the MNUET algorithm. We seek to improve the mask estimation stage of the algorithm by investigating other clustering techniques and their suitability for source separation. The next chapter describes the MNUET in detail with several clustering algorithm options for mask estimation.
Chapter 4

System overview

As discussed in the previous chapter, the MENUET BSS scheme [10] has its benefits and shortcomings. We propose to improve the robustness of the basic MENUET scheme by investigating alternative methods of mask estimation. In this chapter, we describe the details of the MENUET algorithm and two alternative clustering algorithms for mask estimation. The rest of the chapter is organized as follows. In Section 4.1, we introduce the observation model and feature extraction techniques. In Section 4.2 we describe in detail the clustering algorithms, and in Section 4.3 we explain the application of time-frequency masks. Section 4.4 describes the method for source resynthesis.

4.1 Signal processing and feature extraction

4.1.1 Observation model

Consider a microphone array of $M$ identical and omnidirectional microphones in a reverberant enclosure where $N$ sources are present. As depicted in Figure 2.1, let $s_1(t), \ldots, s_N(t)$ denote the original source signals, and $x_1(t), \ldots, x_M(t)$ the microphone observations at time $t$. We assume a convolutive mixing model (cf. Section 2.1), where the observation at the $m^{th}$ microphone can be modeled as a mixture

$$x_m(t) = \sum_{n=1}^{N} s_{mn}(t)$$  \hspace{1cm} (4.1)
of source images at each microphone $m$

$$s_{mn}(t) = \sum_p h_{mn}(p) s_n(t - p), \quad (4.2)$$

where $h_{mn}(p)$ represents the impulse response between source $n$ and microphone $m$.

The goal of any BSS system is to recover the sets of the separated source signal estimates, \{\hat{s}_1(t), \ldots, \hat{s}_M(t)\}, \ldots, \{\hat{s}_N(t), \ldots, \hat{s}_N(t)\}$ where each set represents a source signal $s_1(t), \ldots, s_N(t)$. Ideally, the separation is performed without any information about $s_n(t)$, $h_{mn}(p)$ and $s_{mn}(t)$, and information should only be taken from the microphone observations $x_m(t)$.

### 4.1.2 STFT analysis

The time-domain microphone observations, sampled at a frequency of $f_s$ or sampling period $t_s$, are converted into their corresponding frequency domain time-series, $x_m(\tau, f) \in \mathbb{C}$ by the STFT. We utilize an $L$-point STFT window with an $S$-point shift as

$$x_m(\tau, f) = \sum_{p=0,1/f_s,\ldots,(L-1)/f_s} \text{win}_{\text{STFT}}(p)x_m(p + \tau) \exp^{-j2\pi fp}, \quad (4.3)$$

where $\tau$ denotes the time frame indices $\tau = 0, St_s, \ldots, T - 1$ and $f$ the frequencies $f = 0, (1/L)f_s, \ldots, ((L - 1)/L)f_s$. An appropriately selected window function is used in $\text{win}_{\text{STFT}}(p)$, where this analysis window is usually chosen such that sufficient information is retained within, whilst simultaneously reducing any signal discontinuities at the edges.

One which attenuates smoothly at each ends is suitable [10], such as the Hann window:

$$\text{win}_{\text{STFT}}(p) = \frac{1}{2} - \frac{1}{2} \cos \left( \frac{2\pi p}{L} \right). \quad (4.4)$$

The convolutive mixing model in (4.1) and (4.2) may be approximated as an instantaneous mixture at each frequency [47]

$$x_m(\tau, f) \approx \sum_{n=1}^N h_{mn}(f)s_n(\tau, f) + n_m(\tau, f) \quad (4.5)$$

where $h_{mn}(p)$ is the impulse response between source $n$ and microphone $m$, $s_n(\tau, f)$ is the time-frequency representation of $s_n(t)$ and $n_m(\tau, f)$ encapsulates any environmental noise and reverberation outside of $\text{win}_{\text{STFT}}(p)$ [18].
The W-DO (sparseness) of the speech signals in the STFT domain assumes at most one dominant source signal per time-frequency slot. This reduces (4.5) to

\[ x_m(\tau, f) \approx h_{mn}(f) s_n(\tau, f), \]  

(4.6)

assuming \( s_n(\tau, f) \) is the dominant source at time-frequency slot \((\tau, f)\). Whilst this assumption of sparseness holds true in anechoic conditions, as the reverberation and/or background noise in the acoustic scene increases it becomes increasingly unreliable. This is due to the effects of multipath audio propagation and multiple reflections [7].

### 4.1.3 Feature extraction

The time-frequency mask is computed by estimation of the time-frequency slots where a signal is assumed dominant. To estimate such time-frequency slots, the frequency-domain observation vectors are normalized and processed in such a way as to represent geometric information in sufficiently sparse conditions.

Previous research by Araki et al. identified level ratios and phase differences between the microphone observations as appropriate features, as such features retain information on the magnitude and argument of the time-frequency slots [10]. Further examples of suitable features are presented in Section 6.1.

The feature vector \( y(\tau, f) \) per time-frequency slot is estimated with normalized level information and phase differences between the microphone observations, represented by [10]:

\[ y(\tau, f) = [y^L(\tau, f), y^P(\tau, f)]^T. \]  

(4.7)

The level information is held in \( y^L(\tau, f) = [y^L_1(\tau, f), \ldots, y^L_M(\tau, f)] \) as

\[ y^L(\tau, f) = \left[ \frac{|x_1(\tau, f)|}{A_l(\tau, f)}, \ldots, \frac{|x_M(\tau, f)|}{A_l(\tau, f)} \right], \]  

(4.8)

and the phase information is captured in \( y^P(\tau, f) = [y^P_1(\tau, f), \ldots, y^P_M(\tau, f)] \) as

\[ y^P(\tau, f) = \left[ \frac{1}{A_p f} \arg \left( \frac{x_1(\tau, f)}{x_J(\tau, f)} \right), \ldots, \frac{1}{A_p f} \arg \left( \frac{x_M(\tau, f)}{x_J(\tau, f)} \right) \right], \]  

(4.9)

for a reference microphone observation \( J \). The level normalization term, \( A_l(\tau, f) = \sqrt{\sum_{m=1}^{M} |x_m(\tau, f)|^2} \), restricts the magnitude of \( y^L(\tau, f) \) to the interval \((0, 1)\). Similarly,
the phase normalization \( A_p = 4\pi c^{-1}d_{\text{max}} \), where \( c \) is the propagation velocity of sound and \( d_{\text{max}} \) the maximum distance between any two microphones in the array, realizes a similar range width as the values in \( y^L(\tau, f) \).

It is widely known that in the presence of reverberation, a greater accuracy in phase ratio measurements can generally be achieved with higher spatial resolution [92]; however, it should be noted that the value of \( d_{\text{max}} \) is upper bounded by the spatial aliasing theorem [10, 11, 18, 92]. In this study, microphone spacing is assumed to be sufficiently small in order to avoid spatial aliasing. However, if the exact value of \( d_{\text{max}} \) is not known, a positive constant may be used in its place [10]. This eliminates the requirement for knowledge of the precise spacing between microphones.

The frequency normalization in (4.9) ensures that the phase ratios in \( y^P(\tau, f) \) remain independent of the frequency, and thus permits full-band clustering. This assumes an anechoic mixing model, but it has been proven to approximately hold in convolutive mixing conditions [10], and is therefore applicable in this study. It is possible to cluster without such frequency independence by implementing frequency bin-wise clustering (cf. Section 3.4), however, the utilization of all the frequency bins avoids the frequency permutation problem and also permits data observations of short length [10].

The normalized observation vectors in (4.7) are modified using their complex representation

\[
y_m(\tau, f) = y^L_m(\tau, f) \exp^{jy^P_m(\tau, f)},
\]

where \( y^L_m(\tau, f) \) and \( y^P_m(\tau, f) \) denote the \( m^{\text{th}} \) components of \( y^L(\tau, f) \) and \( y^P(\tau, f) \) respectively. This modification is also equivalently effected by [10]

\[
\hat{y}_m(\tau, f) = |x_m(\tau, f)| \exp\left[j \frac{1}{A_p} \arg \left( \frac{x_m(\tau, f)}{x_J(\tau, f)} \right) \right],
\]

\[
y(\tau, f) \leftarrow \frac{\hat{y}(\tau, f)}{\|\hat{y}(\tau, f)\|},
\]

where \( \hat{y}(\tau, f) \) is an \( M \)-dimensional complex-valued vector \( \hat{y}(\tau, f) = [\hat{y}_1(\tau, f), \ldots, \hat{y}_M(\tau, f)]^T \). The unit-norm normalization in (4.12) retains spatial information and also projects the vector onto a unit hypersphere. This aids in the Euclidean
distance calculation for the subsequent clustering stages [10]. For the remainder of the discussion, any reference to feature vector $y(\tau, f)$ will henceforth be related to (4.12).

4.2 Clustering techniques

The three clustering techniques introduced in this section all belong to the family of partitional center-based clustering, and each has its own unique objective function. The common goal is the classification of the set of feature vectors, $Y = \{y(\tau, f)|y(\tau, f) \in C^M, (\tau, f) \in \Omega\}$ into $N$ clusters, where $\Omega$ is the set of time-frequency slots in the STFT plane. We refer to each feature vector $y(\tau, f)$ as $y_{\tau f}$ henceforth for conciseness. In the instance where the clusters are distinct, as with the HKM, each data point may only belong to one cluster. However, for the soft clustering techniques, each data element may belong to multiple clusters with a certain probability (membership).

4.2.1 HKM clustering algorithm

Previous mask estimation methods for source separation have employed binary clustering techniques such as the hard $k$-means (HKM) [10, 93]. The HKM is a discriminative clustering technique and much of its popularity lies in its ease of implementation, simplicity and efficiency. The HKM algorithm was independently introduced in numerous studies by Steinhaus [94], Ball and Hall [95] and MacQueen [96].

In this approach, the set of feature vectors $y_{\tau f}$ is clustered into $N$ distinct cluster sets $C = \{C_1, \ldots, C_N\}$. Each set from $C$ contains the feature vectors assigned to each of the $N$ clusters, and has an associated prototype vector, or centroid, denoted $v_{n, \text{HKM}}$. Clustering of the data is achieved through minimization of the objective function

$$J_{\text{HKM}} = \sum_{n=1}^{N} \sum_{y_{\tau f} \in C_n} \|y_{\tau f} - v_{n, \text{HKM}}\|^2.$$ (4.13)

Conditional on a set of initial centroids, this minimization is iteratively realized by the following alternating equations

$$C_n = \{y_{\tau f} | n = \arg\min_{n} \|y_{\tau f} - v_{n, \text{HKM}}\|^2\},$$ (4.14)
until convergence is met, where \( E\{\cdot\}_{y_{rf} \in C_n} \) denotes the mean operator for the feature vectors classified as within the cluster set \( C_n \) at that iteration. Convergence is typically defined as when the cluster assignments between successive iterations no longer change. The performance of the HKM clustering is sensitive to the initialization of the cluster centres, and it is therefore recommended to either design the initial centroids using an assumption on the microphone and source signal geometry as in [10], or to utilize the best outcome of a predetermined number of independent runs. The HKM is summarized in Algorithm 1.

**Algorithm 1** The hard \( k \)-means (HKM) clustering algorithm.

```
input: \( Y, N \)
output: \( V_{HKM} = \{v_{n,HKM}\}_{n=1}^N, C \)
1: initialize centroids \( V_{HKM}^{(0)} \)
2: repeat for \( r = 1, 2, \ldots \)
3: update cluster sets \( C^{(r)} \) using (4.14)
4: update centroids \( V_{HKM}^{(r)} \) with \( C^{(r)} \) using (4.15)
5: until Until cluster assignments between successive iterations no longer significantly change or predetermined number of runs reached
6: return \( V_{HKM} \leftarrow V_{HKM}^{(r)} \) and \( C \leftarrow C^{(r)} \)
```

### 4.2.2 GMM clustering algorithm

A number of studies in the literature for time-frequency clustering-based BSS have implemented the GMM for mask estimation [12, 15, 16] and it is therefore included in this study for comparative purposes. It is also included to further investigate the effects of soft masking on the MENUET system by providing the FCM with a fair comparison.

In the GMM-based clustering, each observation \( y_{rf} \) can be modeled as a weighted sum of \( K \) component Gaussian densities (clusters). Unlike the HKM (cf. Section 4.2.1) and the FCM (cf. Section 4.2.3), where the number of clusters is equal to the number of sources, the GMM-based clustering methods have the additional complexity in that the best fitting for the data in \( Y \) to a mixture model does not imply that \( K \) is equivalent to the number...
of sources [10]. The GMM also differs to the HKM and FCM in that it is a generative clustering algorithm, meaning that models (components) are generated to describe each of the clusters, and these models are then used to classify the data.

The data in $\mathbf{Y}$ is assumed to follow a Gaussian distribution as

$$p(\mathbf{y}_f|\Theta) = \sum_{k=1}^{K} \alpha_k p(\mathbf{y}_f|\mu_k, \Sigma_k),$$

which is a mixture of $K$ Gaussian distributions given by

$$p(\mathbf{y}_f|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{y}_f - \mu_k)^T \Sigma_k^{-1} (\mathbf{y}_f - \mu_k)\right),$$

where $\alpha_k$ denotes the mixture weight, $\mu_k$ the mean, and $\Sigma_k$ the covariance of the $k^{th}$ component. We denote the unknown mixture model parameter set as

$$\Theta = \{\theta_k\}_{k=1}^{K},$$

where the parameter set for each mixture is given by

$$\theta_k = \{\alpha_k, \mu_k, \Sigma_k\}.$$  

subject to the constraint

$$\sum_{k=1}^{K} \alpha_k = 1.$$  

The unknown parameter set $\Theta$ is estimated to optimize the log-likelihood of the mixture model, where this estimation is most commonly iteratively calculated using the expectation-maximization (EM) algorithm [16]. The data is then clustered around the maximum-likelihood parameters as determined from the EM algorithm by the final estimates of the $a$ posteriori probabilities at convergence.

We consider the maximum $a$ posteriori (MAP) estimation defined by

$$\max_{\Theta} \log p(\Theta|\mathbf{Y}) = \max_{\Theta} \left[ \sum_{\mathbf{y}_f} \log p(\mathbf{y}_f|\Theta) + \log p(\Theta) \right],$$

subject to the constraint in (4.20). We assume independence in the data $\mathbf{Y}$ and uniform prior distributions of $\mu_k$ and $\Sigma_k$. This optimization is efficiently executed via the EM algorithm, which iterates the E-step and M-step until convergence.
The E-step amounts to the calculation of the posterior probability, \( P(k|\mathbf{y}_{rf}, \Theta') \), where \( \Theta' \) denotes the current estimate of the parameter set. Noting Bayes’ theorem, this posterior probability \( \gamma_{krf} = P(k|\mathbf{y}_{rf}, \Theta') \) is computed as

\[
\gamma_{krf} = \frac{\alpha'_k p(\mathbf{y}_{rf}|\theta'^*_k)}{p(\mathbf{y}_{rf}|\Theta')} = \frac{\alpha'_k p(\mathbf{y}_{rf}|\theta'^*_k)}{\sum_{k=1}^{K} p(\mathbf{y}_{rf}|\theta'^*_k)}.
\]

(4.22)

In the M-step, we update the parameters in \( \Theta \) by maximizing the auxiliary function of the data likelihood, \( Q(\Theta, \Theta') \). Using the updated values in (4.22), the auxiliary function is given by

\[
Q(\Theta, \Theta') = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \log \alpha_k p(\mathbf{y}_{rf}|\theta_k).
\]

(4.23)

To obtain the update equations for each parameter in \( \Theta \), the equation in (4.23) is optimized with respect to each parameter. The details of this derivation are included in Appendix A.1.

The M-step then amounts to re-estimating each parameter according to the equations as follows. The update for the mean \( \mu_k \) is given by the weighted mean of the points in the data set as

\[
\mu_k = \frac{\sum_{\Omega} \gamma_{krf} \mathbf{y}_{rf}}{\sum_{\Omega} \gamma_{krf}}.
\]

(4.24)

The update for the covariance matrix is given by

\[
\Sigma_k = \frac{\sum_{\Omega} \gamma_{krf} (\mathbf{y}_{rf} - \mu_k)(\mathbf{y}_{rf} - \mu_k)^T}{\sum_{\Omega} \gamma_{krf}}.
\]

(4.25)

The mixture weight \( \alpha_k \) is updated as

\[
\alpha_k = \frac{\sum_{\Omega} \gamma_{krf}}{|\mathbf{Y}|}
\]

(4.26)

where \( |\cdot| \) denotes the cardinality. After a suitable initialization of \( \Theta \) (cf. Section 5.4), the E-step and M-step are iterated until convergence, where convergence is assumed to be reached when the differences between parameter set estimates in successive iterations is sufficiently small. The GMM is summarized in Algorithm 2.
Algorithm 2 The Gaussian mixture model (GMM) clustering algorithm.

input: $Y$, $K$, $\epsilon$
output: $\gamma_{k\tau f}$

1: initialize parameter set $\Theta$
2: repeat
3: E-step: compute posterior probabilities $\gamma_{k\tau f}$ using (4.22)
4: M-step: update mean $\mu_k$ using (4.24)
5: M-step cont.: update covariance matrix $\Sigma_k$ using (4.25)
6: M-step cont.: update mixture weight $\alpha_k$ using (4.26)
7: until difference between successive parameter set estimates less than termination threshold $\epsilon$
8: return $\gamma_{k\tau f}$

4.2.3 FCM clustering algorithm

The use of a soft clustering algorithm for mask estimation has improved the accuracy of the source separation in recent research [11, 22]. The origins of the FCM were proposed by Dunn [97] and later improved upon by Bezdek [98]. It is viewed as an extension of the HKM and as with the HKM, the feature set is clustered into $N$ clusters, where each cluster is represented by a centroid $v_{n, FCM}$. In contrast to the HKM, the FCM also has a partition matrix $U_{FCM} = \{u_{n\tau f} \in \mathbb{R}^{N \times T \times F}\}$ where $T$ and $F$ denote the number of time frames and frequency bins respectively. The partition matrix specifies the probability $u_{n\tau f}$, where $u_{n\tau f} \in [0, 1]$, to which a feature vector $y_{\tau f}$ belongs to the $n^{th}$ cluster.

Clustering is achieved by the minimization of the cost function

$$J_{FCM} = \sum_{n=1}^{N} \sum_{\Omega} u_{n\tau f}^q \|y_{\tau f} - v_{n, FCM}\|^2,$$  

(4.27)

where $u_{n\tau f}$ is subject to the constraint $\sum_{n=1}^{N} u_{n\tau f} = 1$. The fuzzification parameter $q > 1$ controls the membership softness in $J_{FCM}$, and therefore controls the fuzziness of the generated time-frequency masks. Section 5.4 describes a suitable value for the fuzzification parameter in this BSS context.

The minimization problem in (4.27) can be solved using Lagrange multipliers and is
typically implemented as an alternating optimization scheme due to the open nature of its solution [11, 99]. Details of the derivation is included are Appendix A.2. The alternating update equations for the centroids and partition matrix is

\[
v_{n,\text{FCM}} = \sum_{\Omega} \frac{u_{n,\tau f} y_{\tau f}}{\sum_{\Omega} u_{n,\tau f}^q},
\]

(4.28)

\[
u_{n,\tau f} = \left[ \sum_{i=1}^N \left( \frac{\|y_{\tau f} - v_{n,\tau f,\text{FCM}}\|}{\|y_{\tau f} - v_{i,\tau f,\text{FCM}}\|} \right)^\frac{1}{q-1} \right]^{-1},
\]

(4.29)

until a suitable termination criterion is satisfied. Typically, for the FCM, convergence is determined as when the difference between successive partition matrices is less than some predetermined threshold \(\epsilon\) [98]. As is also the case with the HKM, it is known that the alternating optimization schemes such as these are prone to converge to a local, as opposed to global, optimum. Therefore, it is suggested to independently implement the algorithm with different start values several times prior to selecting the most fitting result [11]. The FCM algorithm is summarized in Algorithm 3.

**Algorithm 3** The fuzzy c-means (FCM) clustering algorithm.

1. input: \(Y, N, q, \epsilon\)
2. output: \(U_{\text{FCM}}, V_{\text{FCM}} = \{v_{n,\text{FCM}}\}_{n=1}^N\)
3. initialize partition \(U_{\text{FCM}}^{(0)}\)
4. repeat for \(r = 1, 2, \ldots\),
5. update centroids \(V_{\text{FCM}}^{(r)}\) with \(U_{\text{FCM}}^{(r-1)}\) using (4.28)
6. update partition matrix \(U_{\text{FCM}}^{(r)}\) with \(V_{\text{FCM}}^{(r)}\) using (4.29)
7. until \(\|U_{\text{FCM}}^{(r)} - U_{\text{FCM}}^{(r-1)}\| < \epsilon\)
8. return \(U_{\text{FCM}} \leftarrow U_{\text{FCM}}^{(r)}\) and \(V_{\text{FCM}} \leftarrow V_{\text{FCM}}^{(r)}\)

4.3 Mask estimation

In this work the source separation is effected through the estimation and application of time-frequency masks. The spatial source image estimate \(\hat{s}_{mn}(\tau, f)\) is retrieved through the application of the \(n^{th}\) mask to the \(m^{th}\) observation as [18]:
\[ \hat{s}_{mn}(\tau, f) = \mathcal{M}_{n,\ast}(\tau, f)x_m(\tau, f), \]  
(4.30)

where * denotes a method of mask estimation (HKM, GMM or FCM). These masks are the direct result of the clustering techniques as described below.

### 4.3.1 HKM mask estimation

For the HKM clustering, a binary mask for the \(n^{th}\) source is simply estimated as \cite{10}

\[
\mathcal{M}_{n,\text{HKM}}(\tau, f) = \begin{cases} 
1 & \text{for } y_{\tau f} \in C_n, \\
0 & \text{otherwise.}
\end{cases}
\]

(4.31)

where \(C_n\) denotes the set of feature vectors classified as belonging to the \(n^{th}\) cluster.

### 4.3.2 GMM mask estimation

For the GMM clustering, the mask is set as the final estimate of the posterior probabilities \(\gamma_{k\tau f}\) of the \(N\) Gaussian components, akin to previous works \cite{15}. If \(K > N\), we use the \(N\) dominant components, as indicated by their mixture weight values in \(\alpha_k\). This equates to

\[
\mathcal{M}_{n,\text{GMM}}(\tau, f) = \gamma_{n\tau f} = P(n|y_{\tau f}, \Theta'),
\]

(4.32)

where the index \(n\) pertains to one of the \(N\) dominant components.

### 4.3.3 FCM mask estimation

The output of the FCM clustering is the soft membership partition matrix \(U_{\text{FCM}}\), which indicates the degree of membership of each feature vector in the time-frequency plane \(\Omega\) to each of the \(N\) clusters. The membership values \(u_{n\tau f}\) are then interpreted as a collection of \(N\) time-frequency separation masks as \cite{11}:

\[
\mathcal{M}_{n,\text{FCM}}(\tau, f) = u_{n\tau f}.
\]

(4.33)
4.4 Source resynthesis

The estimated spatial source images \( \hat{s}_{mn}(\tau, f) \) are reconstructed in the time domain to obtain the estimates \( \hat{s}_{mn}(t) \). We follow related works [18] and apply an inverse STFT as

\[
\hat{s}_{mn}(t) = \sum_{\tau} \text{win}_{\text{ISTFT}}(t - \tau) \left[ \frac{1}{L} \sum_{f} \hat{s}_{mn}(\tau, f) \exp^{j2\pi f(t-\tau)} \right],
\]

where \( \text{win}_{\text{ISTFT}} \) is an appropriately chosen window that is defined as nonzero in the interval \([0, f_s(L-1)]\) and tapers smoothly to zero at each end to lessen the truncation at the edges. A perfect reconstruction may be realized if the windows used for the STFT and inverse STFT satisfy the condition \( \sum_{\tau} \text{win}_{\text{STFT}}(t - \tau) \text{win}_{\text{ISTFT}}(t - \tau) = 1 \), where \( \text{win}_{\text{STFT}} \) denotes the window used in STFT analysis (cf. Section 4.1.2) [18].

Figure 4.1 provides a pictorial overview of the system presented in this chapter to aid in an intuitive understanding of the general clustering-based approach to BSS in this thesis. Figure 4.1a depicts the spectrograms of the original four sources, recorded by three microphones as depicted in Figure 4.1b. The estimated time-frequency masks (soft masks, as estimated by the FCM algorithm) are in Figure 4.1c, whilst the estimated source signals, recovered through application to a microphone observation, are in Figure 4.1d.

4.5 Summary

In this chapter, a detailed overview of the MENUET BSS system was presented, which forms the fundamental framework on which much of the work in this thesis builds upon. We began with the processing of the microphone signals with details of the STFT analysis. The extraction of spatial features from the microphone observations was then explained with the combination of level and phase features into the complex-valued feature vector. We described the HKM clustering technique in detail, which is the method of mask estimation in the original MENUET system. We presented two alternative methods of mask estimation via the GMM and FCM clustering algorithms. The estimation of the masks was covered with the application to microphone observations and the recovery of the source signal estimates. We concluded with a pictorial overview of the time-frequency masking approach to BSS to
Figure 4.1: Spectrogram examples of various stages in the BSS scheme for a case with three microphones and four sources. (a) Sources, (b) Mixtures, (c) Time-frequency separation masks (soft) and (d) Estimated separated signals.

This chapter has provided the foundations for the initial experimental evaluations and variations on the MENUET scheme to follow in subsequent chapters. In the following chapter, we explain the experimental setup that is used to realize the system just described.
Chapter 5

Experimental setup

In Chapter 4 we provided an overview of the focal MENUET BSS system employed in this thesis. This chapter presents the setup and performance evaluation metrics for the experimental sections of this thesis. The rest of the chapter is organized as follows. In Section 5.1 we explain the experimental conditions for simulated data, and in Section 5.2 we describe the collection of real room data. Section 5.3 includes information on the public benchmark data utilized in this study. We explain the algorithmic details of the clustering methods in Section 5.4, and in Section 5.5 we describe the criteria used for BSS performance evaluation.

5.1 Simulated conditions

The first set of experiments were conducted in simulated conditions, designed to emulate that of the MENUET algorithm to as close a degree as possible. Furthermore, in simulated conditions a greater level of control over the room conditions was available, allowing a range of reverberation levels to be assessed. Table 5.1 summarizes the experimental conditions, and Figure 5.1 depicts the source and microphone layout. The wall reflections of the enclosure and the room impulse responses between each source and microphone were simulated using the image model method for small-room acoustics [100]. The room reverberation was quantified in the measure RT$_{60}$, where RT$_{60}$ is defined as the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound [100].
Table 5.1: Experimental conditions for evaluations with simulated and real room recording data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of microphones</td>
<td>$M = 3$</td>
</tr>
<tr>
<td>Number of sources</td>
<td>$N = 4$</td>
</tr>
<tr>
<td>Microphone spacing $d$</td>
<td>4 cm or 8 cm</td>
</tr>
<tr>
<td>Source to microphone spacing $R$</td>
<td>50 cm or 120 cm</td>
</tr>
<tr>
<td>Reverberation time</td>
<td>0 ms $\sim$ 450 ms</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>8 kHz</td>
</tr>
<tr>
<td>STFT window</td>
<td>Hann</td>
</tr>
<tr>
<td>STFT frame size</td>
<td>$L = 1024$ (128 ms)</td>
</tr>
<tr>
<td>STFT frame shift</td>
<td>$S = 256$ (32 ms)</td>
</tr>
</tbody>
</table>

The four speech sources were of randomly generated gender combinations, and were realized with phonetically-rich utterances from the TIMIT database [101] of length 6 s. The target-to-masker ratio between all of the sources was set to 0 dB. In avoidance of any spatial aliasing, the microphones were placed at a maximum distance of 4 cm apart. However, we also considered a microphone spacing of 8 cm to evaluate the effects of spatial aliasing (cf. Section 4.1.3). The distance $R$ between the microphone array and speakers was varied from 50 cm to 120 cm.

5.2 Real room recordings

The microphone and source setup in Figure 5.1 was then realized in a real-world office environment to validate the performance. The recordings were collected in an office of dimensions 6.00 × 5.20 × 2.65 m. An equilateral triangular array of three omnidirectional microphones was placed in the room, with the center of the array located at 2.00 × 1.75 × 1.00 m. The setup is shown in Figure 5.1, where the microphones are spaced 4 cm apart and the sources are at a distance of $R$ from the array center. As with the simulated conditions in Section 5.1, the value of $R$ was varied between 50 cm and 120 cm. The source signals were from the TIMIT database, and utterances were looped to a common length of 10 s. As with the simulated experiments, we also considered a larger microphone spacing of 8 cm to
investigate the effects of spatial aliasing on the source separation performance.

The room reverberation was measured using the method of recording the room response to an impulsive source [102, 103]. This technique records the response of the room to an impulsive-like source of excitation. In this study, the popping of a balloon was used as the source of excitation. The room reverberation was measured to be $\text{RT}_{60} = 390 \text{ ms}$.

### 5.3 Public benchmark data

To evaluate and compare our proposed methods with respect to the current state-of-the-art BSS algorithms, we also considered the publicly available benchmark data of the 2008 and 2010 Signal Separation Evaluation Campaign (SiSEC) [104, 105]. The development data sets were used as we required access to the contribution of each source for performance evaluation purposes.

From the 2008 set, the development data (dev1.zip) of the “Under-determined speech and music mixtures” set was used. We only considered the live recording “liverec” data. The room dimensions in this set were $4.45 \times 3.55 \times 2.50 \text{ m}$. The reverberation time $\text{RT}_{60}$ was either 130 ms or 250 ms, and the microphone spacing was 5 cm. The sources were placed at a distance of 100 cm from the microphones. The number of speakers was varied from
Figure 5.2: Configuration of microphones and sources for experiments with SiSEC 2008 data.

From the 2010 SiSEC data, we used the development data (dev.zip) in “Source separation in the presence of real-world background noise” set. In this data, two microphones were spaced at 8.6 cm and noise signals were recorded in real-world noise environments: ‘Cafeteria’ (Ca) and ‘Square’ (Sq). The ‘Cafeteria’ environment was stated as reverberant with an unspecified reverberation time, whilst the ‘Square’ had little or no reverberation. The noise signals were recorded at two different positions within the environment, center (Ce; where noise is more isotropic), and corner (Co; where noise may not be very isotropic) [105]. For each of the noise environments, two different locations of the same type of environment were considered (A and B). Figure 5.3 depicts the configuration of the microphones and source speakers, and Figure 5.4 illustrates the two different noise locations. As with the SiSEC 2008 data, the recordings were 10 seconds long, with mixed English and Japanese utterances of both genders. The number of sources was set to three. The distance between the sources and microphones was unspecified. The original samples were recorded at 16 kHz.

Due to the fact that both SiSEC databases employ a stereo configuration, we slightly modified the feature vectors to reflect this. The authors of the MENUET algorithm designed
Figure 5.3: Configuration of microphones and sources for experiments with SiSEC 2010 data.

the feature vector in (4.10) for \( M > 2 \), and presented an equivalent feature for when \( M = 2 \). We therefore employed feature F from Table 6.1, which was also utilized by Araki et al. for \( M = 2 \) [10]. It should be noted that for both the SiSEC 2008 and 2010, we did not alter the signals to account for the spatial aliasing in order to provide a fair comparison against the publicly published results [106, 107]. This is despite the inevitable aliasing due to the larger microphone spacing (i.e. \( d > c/f_s \), cf. Section 4.1.3). For easy comparison against the published results of the SiSEC the same evaluation criteria for the “Source spatial image estimation” task was used. This is outlined in Section 5.5. The conditions of the SiSEC recordings are summarized in Table 5.2.

5.4 Algorithmic details of the HKM, GMM and FCM

As mentioned in Sections 4.2.1 and 4.2.3, it is widely recognized that the performance of clustering algorithms is largely dependent on the initialization of the algorithms [91, 108]. If the initial parameters are not estimated with sufficient precision there is a high possibility of finding a local, as opposed to global, optimum. It has been recommended [91] to run the algorithms multiple times to reduce the degrading effects of its sensitivity; the effectiveness of this style of initialization was also described in [109]. In an effort to minimize
computational expense, we determined the smallest number of independent, single-iteration runs for initialization which would result in the best solution. Previous experiments by Kühne et al. [11] had implemented the best of 50 runs for FCM initialization; however, it was empirically confirmed that there was negligible difference in performance between 25 or 50 runs. Therefore, the initialization for the HKM and FCM was by the best solution of 25, independent, randomly initialized single-iterations runs. The ‘best’ solution was defined as the lowest cost function output of the independent runs.

Similar to the HKM and FCM algorithms, the GMM clustering approach also requires a suitable initialization. We follow a similar initialization as was implemented in other GMM-based clustering BSS schemes [18]. The initial centroids were selected as a subset of the clustering data and the mixture weights were set to $\frac{1}{K}$, where $K$ denotes the number of mixtures. The covariance matrices for all components were diagonal, where each element on the diagonal is the variance of the data along the corresponding dimension.

The GMM clustering approach is also sensitive to the selection of an appropriate number of mixture components in the model. It was observed in the experiments that an increase in the number of mixture components generally resulted in improved separation performance; however, this was at the expense of a considerable amount of experimentation and computational time. We therefore set $K = N$ for all evaluations, however, in Appendix B.1 we...
Table 5.2: Experimental conditions for evaluations with SiSEC 2008 and 2010 data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of microphones</td>
<td>$M = 2$</td>
</tr>
<tr>
<td>Number of sources</td>
<td>$N = 3, 4$</td>
</tr>
<tr>
<td>Microphone spacing $d$</td>
<td>5 cm (SiSEC 2008) or 8.6 cm (SiSEC 2010)</td>
</tr>
<tr>
<td>Source to microphone spacing</td>
<td>100 cm (SiSEC 2008)</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>16 kHz</td>
</tr>
<tr>
<td>STFT window</td>
<td>Hann</td>
</tr>
<tr>
<td>STFT frame size</td>
<td>$L = 2048$ (128 ms)</td>
</tr>
<tr>
<td>STFT frame shift</td>
<td>$S = 512$ (32 ms)</td>
</tr>
</tbody>
</table>

present the results of our comprehensive investigation into the optimal value of $K$ over a range of acoustic environments. The assignment of $K = N$ is in line with other studies that use the GMM within the time-frequency masking BSS framework [16].

Section 4.2.3 explains the role of the fuzzification parameter $q$ in the FCM clustering. Past research [11] identified that a value of $q$ in the range of $q \in (1, 1.5]$ results in source separation performance akin to hard clustering, and it was empirically determined that a value of $q = 2$ is optimal to achieve a balance between high separation performance with minimal artifacts [11]. This is consistent with other studies which also report an optimal value of 2 for the fuzzification parameter [110, 111]. Therefore, in this set of evaluations, the fuzzification $q$ is set to 2.

5.5 Performance evaluation criteria

We evaluated the source separation ability of the time-frequency masks with the widely used BSS EVAL toolkit [112]. This set of metrics is applicable to all source separation approaches, and no a priori information of the separation algorithm is required. It is also used for public benchmark evaluations, such as the SiSEC data.

Using a least-squares projection, the BSS EVAL toolkit assumes the decomposition of the estimated spatial images $\hat{s}_{mn}(t)$ (4.34) as
\[ \hat{s}_{mn}(t) = s_{mn}^{\text{img}}(t) + e_{mn}^{\text{spat}}(t) + e_{mn}^{\text{interf}}(t) + e_{mn}^{\text{artif}}(t), \] (5.1)

where \( m \) denotes the microphone observation index, \( s_{mn}^{\text{img}}(t) \) is the true source image, and \( e_{mn}^{\text{spat}}(t), e_{mn}^{\text{interf}}(t), e_{mn}^{\text{artif}}(t) \) are distinct error components representing spatial distortion, interference and artifacts respectively.

From this decomposition, we compute the image-to-spatial-distortion ratio (ISR) to provide an estimate of the relative amount of distortion present in the \( n \)th recovered signal as

\[ \text{ISR}_n = 10 \log_{10} \frac{\sum_{m=1}^{M} \sum_{t} s_{mn}^{\text{img}}(t)^2}{\sum_{m=1}^{M} \sum_{t} e_{mn}^{\text{spat}}(t)^2}. \] (5.2)

We compute the source-to-interference ratio (SIR) to provide an estimate of the relative amount of interference in the target source estimate. The SIR of the \( n \)th source is calculated as

\[ \text{SIR}_n = 10 \log_{10} \frac{\sum_{m=1}^{M} \sum_{t} (s_{mn}^{\text{img}}(t) + e_{mn}^{\text{spat}}(t))^2}{\sum_{m=1}^{M} \sum_{t} e_{mn}^{\text{interf}}(t)^2}. \] (5.3)

An estimate of the amount of artifacts is quantified in the sources-to-artifacts ratio (SAR) as

\[ \text{SAR}_n = 10 \log_{10} \frac{\sum_{m=1}^{M} \sum_{t} (s_{mn}^{\text{img}}(t) + e_{mn}^{\text{spat}}(t) + e_{mn}^{\text{interf}}(t))^2}{\sum_{m=1}^{M} \sum_{t} e_{mn}^{\text{artif}}(t)}. \] (5.4)

And finally, as an estimate of the total error in the \( n \)th recovered source, the signal-to-distortion ratio (SDR) is calculated as

\[ \text{SDR}_n = 10 \log_{10} \frac{\sum_{m=1}^{M} \sum_{t} s_{mn}^{\text{img}}(t)^2}{\sum_{m=1}^{M} \sum_{t} (e_{mn}^{\text{spat}}(t) + e_{mn}^{\text{interf}}(t) + e_{mn}^{\text{artif}}(t))^2}. \] (5.5)

It should be noted that the three error components in the overall quality measure in (5.5) have equal significance. In practice, it is recommended to weight the error components according to their application [112]. For example, in hearing aid applications, the presence of artifacts is crucial, and the corresponding error component \( e_{mn}^{\text{artif}}(t) \) should thus be given a higher weight.
Whilst each ratio encapsulates a unique aspect of the recovered source estimates, the most significant can generally be said to be the SDR, as it provides a measure of the overall quality of the source separation algorithm, and BSS systems in the literature often only report based on the SDR [13, 18]. However, many leading algorithms report on both the SDR and SIR as the SIR is an overall measure of separation ability, and is thus arguably equally as important [12, 81, 83]. The authors of the BSS EVAL algorithm also specify that spatial distortion is often of little importance for most applications of BSS, and as such several algorithms neglect this measure and report only on the other measures [72, 75, 89]. However, for the sake of completeness we include all four ratios to provide a complete picture of the performance of the algorithms in this thesis.

5.6 Summary

In this chapter, we outlined the experimental setup used for the evaluations in this thesis. The method used for collection of the simulated data was detailed, including the configuration of microphones and sources. The setup for real room recordings was also explained along with information on the benchmark data sets of the SiSEC. Algorithmic details on the HKM, GMM and FCM were also described. The chapter concluded with an explanation of the tools used for performance evaluation criteria, and provides the necessary background for the subsequent experimental evaluations.
Chapter 6

Evaluations on HKM, GMM and FCM

In Chapter 5 we described the experimental setup and performance evaluation metrics used to realize the MENUET-based BSS system of Chapter 4. This chapter presents the results of the evaluations conducted on the aforementioned system. The rest of the chapter is organized as follows. Section 6.1 presents the results for initial experiments on the FCM for a simple stereo configuration. Section 6.2 presents the results for simulated reverberant conditions on the three microphone setup of the MENUET algorithm with the two additional clustering algorithms as described in Section 4.2. Section 6.3 repeats the experiments of Section 6.2 in a real office environment, and Section 6.4 evaluates the system on the international benchmark data of the SiSEC.

6.1 Initial evaluations of FCM for mask estimation

Prior to evaluating the effectiveness of the FCM clustering for mask estimation in the three-microphone MENUET system, the FCM was evaluated for a simple stereo setup using a variety of feature sets in order to test its feasibility. Araki et al. presented a comprehensive review of suitable location features and their effectiveness at separation was evaluated using the HKM clustering for mask estimation [10]. We repeated these experiments using both the HKM and FCM clustering algorithms for mask estimation.
The experimental setup for these simulated evaluations was such as to replicate the original work by Araki et al. to as close a degree as possible. Figure 6.1 depicts the configuration. In an enclosure of dimensions \(4.55 \times 3.55 \times 2.50\) m with reverberation time of \(\text{RT}_{60} = 128\) ms, two omnidirectional microphones were placed a distance of 4 cm apart at an elevation of 120 cm. Three speech sources with a target-to-masker ratio of 0 dB were situated at 30\(^\circ\), 70\(^\circ\) and 135\(^\circ\) at a distance of 50 cm from the array at the same elevation as the microphones. The speech sources were randomly chosen from both genders of the TIMIT database in order to emulate the investigations by Araki et al. which utilized English utterances. The source separation performance was evaluated with respect to the improvement in the SIR, i.e. \(\text{SIR}_{\text{output}} - \text{SIR}_{\text{input}}\). The results are shown in Table 6.1.

![Figure 6.1: Configuration of microphones and sources for preliminary experiments using a variety of features for clustering.](image)

The original purpose of the evaluations by Araki et al. was to determine the effects of different normalizations upon the level and phase ratio features (cf. 4.1.3). As expected, separation performance generally increased as the features are of the same order of magnitude, due to the reduced variance between the features. It is additionally observed from the measured SIR gain that the FCM clustering is more robust than the original HKM for all but one feature set, and thus hints at the possibility of the FCM yielding similar results for related time-frequency masking BSS approaches. This confirms the suitability of the FCM for mask estimation in this BSS scheme, and also demonstrates its robustness against several types of spatial location features. The results of this investigation provide further
Table 6.1: Source separation results for the configuration in Figure 6.1 for a reverberation time of $RT_{60} = 128$ ms when the HKM and FCM were used for mask estimation. Results given with respect to the SIR gain averaged across all sources.

<table>
<thead>
<tr>
<th>Feature $\theta(\tau, f)$</th>
<th>SIR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HKM</td>
</tr>
<tr>
<td>A $\theta(\tau, f) = \left[ \frac{</td>
<td>X_2(\tau,f)</td>
</tr>
<tr>
<td>B $\theta(\tau, f) = \left[ \frac{</td>
<td>X_2(\tau,f)</td>
</tr>
<tr>
<td>A' $\theta(\tau, f) = \left[ \frac{</td>
<td>X_2(\tau,f)</td>
</tr>
<tr>
<td>C $\theta(\tau, f) = \frac{1}{2\pi} \arg \left( \frac{X_2(\tau,f)}{X_1(\tau,f)} \right)$</td>
<td>10.2</td>
</tr>
<tr>
<td>D $\theta(\tau, f) = \frac{1}{2\pi} \arg \left( \frac{X_2(\tau,f)}{X_1(\tau,f)} \right)$</td>
<td>10.1</td>
</tr>
<tr>
<td>E $\theta(\tau, f) = \left[ \frac{</td>
<td>X_2(\tau,f)</td>
</tr>
<tr>
<td>F $\theta(\tau, f) = \left[ \frac{</td>
<td>X_2(\tau,f)</td>
</tr>
<tr>
<td>G $\theta(\tau, f) = \left[ \frac{</td>
<td>X_2(\tau,f)</td>
</tr>
</tbody>
</table>

motivation to apply the FCM for mask estimation to other microphone arrangements and adverse acoustic conditions.

Araki et al. also compared the performance of the HKM against GMM clustering for a subset of the features in Table 6.1. The results of this demonstrated notable improvements in the SIR gain in comparison to the HKM, although this was at the burden of significantly greater computational expense. Furthermore, the selection of the number of Gaussian components proved to require a lot of trial and error, as mentioned in Section 4.2.2. However, in order to offer a fair comparison of the FCM against other clustering techniques, the GMM clustering was also implemented in all of our subsequent evaluations.
6.2 Simulated reverberant conditions

Evaluations were conducted on the MENUET system and its variations with the GMM and FCM for mask estimation using simulated data on the configuration as in Figure 5.1. The average performance ratios as measured by the BSS EVAL toolbox are depicted in Figure 6.2. It is immediately evident that the FCM offers considerable improvement over the HKM and GMM algorithms for all performance ratios, especially the SDR and SIR. We notice that the GMM outperforms the HKM with respect to the SAR, and this can be attributed to the musical noise distortion that arises when binary masking is used [113].

A generally smaller standard deviation for the FCM is also noted, for example in Figure 6.2c, at a reverberation time of RT_{60} = 300 ms and a distance of R = 50 cm, the standard deviation of the FCM is less than half of the other two clustering algorithms. This suggests that the FCM delivers more consistent and reliable separation of the sources.

However, we note that at the reverberation times of 300 ms and 450 ms when R = 120 cm, none of the clustering techniques are able to separate the sources efficiently, as indicated by the negative SIR values. This suggests that the standard off-the-shelf clustering algorithms are not suitable in more hostile conditions, and greater research into advancing the algorithms may prove beneficial.

To evaluate the statistical significance of the results, we conducted a Student’s t-test on the SDR and SIR results (Figures 6.2a and Figure 6.2c respectively). We selected this subset of the four performance ratios as they can be considered the most important ratios of the set (cf. Section 5.5). As the main contribution of this chapter is to evaluate the FCM against the HKM, we measured the p-values between these two clustering algorithms. A two-tailed distribution was assumed for each test, with unequal variances between the data. A 5\% significance level was used for all tests. Table 6.2 displays the p-values for each test.

At a glance we can see that for all the simulated conditions the p-value is below the 5\% significance level for both the SDR and SIR. We can therefore conclude that the results of using the FCM for mask estimation are encouraging as the FCM consistently delivers improvements in the SDR and SIR. This demonstrates that the performance of the FCM for mask estimation is largely unlikely to be due to chance in simulated environments.
Figure 6.2: Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with simulated data, when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Figure 6.2: cont. Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with simulated data, when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Table 6.2: Results of the Student’s $t$-test on the SDR and SIR results of Figure 6.2. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.

<table>
<thead>
<tr>
<th>Condition</th>
<th>SDR</th>
<th>SIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RT_{60} = 0$ ms, $R = 50$ cm</td>
<td>$p &lt; 0.001$</td>
<td>0.0016</td>
</tr>
<tr>
<td>$RT_{60} = 128$ ms, $R = 50$ cm</td>
<td>$p &lt; 0.001$</td>
<td>0.01</td>
</tr>
<tr>
<td>$RT_{60} = 300$ ms, $R = 50$ cm</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$RT_{60} = 450$ ms, $R = 50$ cm</td>
<td>$p &lt; 0.001$</td>
<td>0.02</td>
</tr>
<tr>
<td>$RT_{60} = 0$ ms, $R = 120$ cm</td>
<td>$p &lt; 0.001$</td>
<td>0.01</td>
</tr>
<tr>
<td>$RT_{60} = 128$ ms, $R = 120$ cm</td>
<td>$p &lt; 0.001$</td>
<td>0.002</td>
</tr>
<tr>
<td>$RT_{60} = 300$ ms, $R = 120$ cm</td>
<td>$p &lt; 0.001$</td>
<td>0.0076</td>
</tr>
<tr>
<td>$RT_{60} = 450$ ms, $R = 120$ cm</td>
<td>$p &lt; 0.001$</td>
<td>0.02</td>
</tr>
</tbody>
</table>

6.3 Real room recordings

The next step was the evaluation of the mask estimation abilities with data that was collected in a real room as described in Section 5.2. The results are shown in Figure 6.3 and follow a similar trend to that reported for the simulated conditions, with the FCM outperforming the other two clustering algorithms for most conditions. We note that at the distance of $R = 120$ cm, the GMM and FCM actually perform below that of the HKM with respect to the SIR. However, as with the simulated results, the FCM is able to achieve a higher average SDR, ISR and SAR for all reverberation times and distances.

We then evaluated with the same conditions but an increase in the microphone spacing to 8 cm to evaluate the effects of spatial aliasing. The results are shown in Figure 6.4a–6.4d. The results follow that in Figure 6.3 with an overall reduction in performance as expected due to the larger microphone spacing.

We also evaluated the statistical significance of the recorded data results using the Student’s $t$-test, and the values are shown in Table 6.3. We note the generally statistically significant results of the FCM over the HKM, except in the values of SIR as depicted in Figures 6.3 and 6.4.
Figure 6.3: Source separation results for $R = 50$ and $R = 120$ cm and a microphone spacing of 4 cm with real data collected in an office environment, when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

Table 6.3: Results of the Student’s $t$-test on the SDR and SIR results of Figures 6.3 and 6.4. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.

<table>
<thead>
<tr>
<th>Condition</th>
<th>SDR</th>
<th>SIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 4\text{ cm}, R = 50\text{ cm}$</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$d = 4\text{ cm}, R = 120\text{ cm}$</td>
<td>$p &lt; 0.001$</td>
<td>0.17</td>
</tr>
<tr>
<td>$d = 8\text{ cm}, R = 50\text{ cm}$</td>
<td>$p &lt; 0.001$</td>
<td>0.65</td>
</tr>
<tr>
<td>$d = 8\text{ cm}, R = 120\text{ cm}$</td>
<td>$p &lt; 0.001$</td>
<td>0.05</td>
</tr>
</tbody>
</table>
6.4 Benchmark data

Table 6.4 shows the results with the SiSEC 2008 “Under-determined speech and music mixtures” data. We can see from the individual results and the overall averages for each reverberation time that the FCM is superior to the HKM and GMM for separation in these real-world conditions. We recall that the sampling frequency of the SiSEC data is 16 kHz, and with a microphone spacing of 5 cm, the spatial aliasing theorem is violated with our full-band clustering (cf. Section 4.1.3).

We then evaluated our algorithm on the SiSEC “Source separation in the presence of real-world background noise” set, with the results shown in Table 6.4. We include all individual results, and the averages for each condition can easily be compared against the
publicly available “Average Results for 2 channels” table [107], with our proposed algorithm yielding varied results. For example, Duong et al. achieved an average SDR and SIR of 1.30 dB and 1.60 dB in the Cafeteria environment whilst we achieved 3.85 dB and 6.87 dB. However, in the Square environment Duong et al. achieve an average SDR of 3.70 dB whilst we measured 2.46 dB. However, it is clear that the FCM yields a better separation result than the HKM and GMM, whilst the HKM generally outperformed the GMM. We also note the higher average performance of all the algorithms in the Square environment over the Cafeteria environment; this is likely due to the negligible reverberation in the Square environment.

6.5 Summary

This chapter presented the preliminary experimental evaluations of this thesis. We conducted a range of experiments designed to comprehensively test the ability of the three clustering algorithms of HKM, GMM and FCM for time-frequency mask estimation. We firstly evaluated the FCM in the context of a simple three source stereo configuration and compared it to previously published results which utilized the HKM for the same configuration. We observed improvements in source separation performance from these initial experiments, in particular for the feature vector that is used in the MENUET.

Motivated by this, we then extended the system to the four source three microphone configuration of the MENUET. We included a soft masking scheme based on the GMM as a comparison, and evaluated with simulated data, real room recordings and public benchmark data. We evaluated with respect to standard current performance measures used in the BSS research field. The simulated results confirmed the superiority of the FCM against the HKM and GMM, however, in hostile conditions such as in high reverberation times, the performance of the FCM was compromised. This was reflected in the results we obtained in real room conditions, where again, the FCM did not surpass the baseline HKM/GMM when the source speakers were moved to a further distance. We also evaluated our work on two data sets from the international public benchmark data of the SiSEC, and our results demonstrated the viability of the FCM as a valid tool for mask estimation, where it
outperformed the HKM and GMM.

However, despite the encouraging results we obtained from these preliminary experiments, we can see the need to improve the performance in instances where the source speakers are further away from the microphones. Given that the MENUET framework exploits the sparseness property of speech, we consider other characteristics of speech signals to utilize. This moves us to the following chapter, where we investigate advancements on the FCM clustering algorithm and their applicability to the source separation task at hand.
Table 6.4: Source separation results with the SiSEC 2008 and 2010 data when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SDR (dB)</th>
<th>ISR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HKM</td>
<td>GMM</td>
</tr>
<tr>
<td>male3_liverec_130ms_5cm</td>
<td>1.13</td>
<td>-0.54</td>
</tr>
<tr>
<td>female3_liverec_130ms_5cm</td>
<td>1.28</td>
<td>0.22</td>
</tr>
<tr>
<td>male4_liverec_130ms_5cm</td>
<td>0.85</td>
<td>0.55</td>
</tr>
<tr>
<td>female4_liverec_130ms_5cm</td>
<td>0.70</td>
<td>-0.55</td>
</tr>
<tr>
<td>male3_liverec_250ms_5cm</td>
<td>0.61</td>
<td>-0.17</td>
</tr>
<tr>
<td>female3_liverec_250ms_5cm</td>
<td>0.34</td>
<td>0.74</td>
</tr>
<tr>
<td>male4_liverec_250ms_5cm</td>
<td>1.22</td>
<td>-0.35</td>
</tr>
<tr>
<td>female4_liverec_250ms_5cm</td>
<td>0.37</td>
<td>-0.19</td>
</tr>
<tr>
<td>Average for RT$_{60} = 130$ ms</td>
<td>0.99</td>
<td>-0.08</td>
</tr>
<tr>
<td>Average for RT$_{60} = 250$ ms</td>
<td>0.64</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th>SDR (dB)</th>
<th>ISR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HKM</td>
<td>GMM</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_A</td>
<td>2.31</td>
<td>0.15</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_B</td>
<td>3.55</td>
<td>-2.47</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_A</td>
<td>2.00</td>
<td>-1.22</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_B</td>
<td>2.40</td>
<td>0.12</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_A</td>
<td>0.66</td>
<td>-0.17</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_B</td>
<td>0.38</td>
<td>-1.23</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_A</td>
<td>3.16</td>
<td>-0.25</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_B</td>
<td>1.10</td>
<td>-0.05</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td>2.57</td>
<td>-0.86</td>
</tr>
<tr>
<td>Average for Square</td>
<td>1.33</td>
<td>-0.43</td>
</tr>
</tbody>
</table>
Table 6.4: cont. Source separation results with the SiSEC 2008 and 2010 data when the HKM, GMM and FCM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SISSEC 2008</th>
<th>SIR (dB)</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HKM</td>
<td>GMM</td>
</tr>
<tr>
<td>male3_liverec_130ms_5cm</td>
<td>1.08</td>
<td>-1.48</td>
</tr>
<tr>
<td>female3_liverec_130ms_5cm</td>
<td>1.95</td>
<td>-0.77</td>
</tr>
<tr>
<td>male4_liverec_130ms_5cm</td>
<td>1.91</td>
<td>-1.42</td>
</tr>
<tr>
<td>female4_liverec_130ms_5cm</td>
<td>3.66</td>
<td>2.05</td>
</tr>
<tr>
<td>male3_liverec_250ms_5cm</td>
<td>1.30</td>
<td>2.85</td>
</tr>
<tr>
<td>female3_liverec_250ms_5cm</td>
<td>2.85</td>
<td>0.85</td>
</tr>
<tr>
<td>male4_liverec_250ms_5cm</td>
<td>1.45</td>
<td>-2.03</td>
</tr>
<tr>
<td>female4_liverec_250ms_5cm</td>
<td>3.14</td>
<td>-1.99</td>
</tr>
<tr>
<td>Average for RT&lt;sub&gt;60&lt;/sub&gt; = 130 ms</td>
<td>2.15</td>
<td>-0.41</td>
</tr>
<tr>
<td>Average for RT&lt;sub&gt;60&lt;/sub&gt; = 250 ms</td>
<td>2.19</td>
<td>-0.08</td>
</tr>
</tbody>
</table>

SiSEC 2010

<table>
<thead>
<tr>
<th>SISSEC 2010</th>
<th>SIR (dB)</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HKM</td>
<td>GMM</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_A</td>
<td>4.72</td>
<td>-1.02</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_B</td>
<td>7.26</td>
<td>0.89</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_A</td>
<td>0.96</td>
<td>1.65</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_B</td>
<td>6.29</td>
<td>0.13</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_A</td>
<td>11.21</td>
<td>2.75</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_B</td>
<td>12.78</td>
<td>4.45</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_A</td>
<td><strong>14.62</strong></td>
<td>4.15</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_B</td>
<td>7.92</td>
<td>2.57</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td>4.81</td>
<td>0.41</td>
</tr>
<tr>
<td>Average for Square</td>
<td>11.63</td>
<td>3.48</td>
</tr>
</tbody>
</table>
CHAPTER 6. EVALUATIONS ON HKM, GMM AND FCM
Chapter 7

Advancements on the fuzzy c-means clustering

As demonstrated in the previous chapter, the standard FCM suffers reduced BSS ability in higher reverberation times when the source to microphone distance is increased. In this chapter, we explore advancements to the FCM for added robustness in such conditions. The rest of the chapter is organized as follows. We introduce the requirement for improvements to the FCM in Section 7.1 and present the proposed modifications in Sections 7.2 and 7.3. We compare the mask estimation abilities of the proposed modified FCM versions in Section 7.4, and in Section 7.5 we present the source separation results of the proposed algorithms for a variety of simulated and real acoustic environments.

7.1 Introduction

Despite the recent improvements of the work by Jafari et al. [22] by using the FCM for mask estimation it is not without its shortcomings, for example, as demonstrated in the previous chapter where the performance of the FCM was compromised at higher reverberation times. The standard FCM clustering is not very robust to outliers and this lack of robustness may manifest itself in the partitioning results: in this context, the estimated time-frequency masks.

Numerous studies have demonstrated that not all the location features of speech signal
mixtures correspond to the source locations in reverberant environments \([114, 115]\). However, the standard FCM treats all data at all time-frequency slots equally. In light of this, Kühne et al. measured the reliability of each location feature at all time-frequency slots in the plane and assigned weights to each as an indication of its usability \([11, 82]\). These indication weights were integrated into the standard FCM and mask estimation by modifying the objective function.

In addition to treating all data within the time-frequency space with equal importance, previous works also computed the separation masks by considering each time-frequency slot in isolation without any information of the immediate surrounding slots \([10, 22]\). However, it has been proposed that dominant segments of speech signals form localized patches within the time-frequency space, and that there exists a strong correlation between a time-frequency slot and its neighboring points \([11]\). This notion was incorporated into an FCM-based method for mask estimation and resulted in improvements in the source separation performance \([11]\). However, it was only evaluated in a simulated, over-determined environment with a uniform linear microphone array.

The use of surrounding and contextual information is also well-documented for robustness in image segmentation algorithms \([116–119]\). Of particular mention is the scheme by Chuang et al. where contextual information was integrated into the cost function of the FCM algorithm for image segmentation \([118]\). This was done by scaling each membership value \((4.29)\) with a weighted sum of the membership values of the immediate surrounding time-frequency slots. This served multiple purposes, including the promotion of homogeneous regions within the data space and robustness against noise in the data. Given the purpose of the membership value scaling, this may prove potentially beneficial to the BSS application at hand, especially because of the highly correlated nature of speech previously discussed. The resulting technique by Chuang et al. was termed the spatial FCM clustering algorithm, which we denote as sFCM.

Motivated by the previous use of weights for the feature data by Kühne et al. \([11]\) and the promising work of the sFCM for image segmentation by Chuang et al. \([118]\), we propose an extension to the MENUET-based BSS scheme described in Chapter 4. We first employ a slight alteration of the work by Kühne et al. and estimate weights for the feature data.
based on the variation in Euclidean distance of the surrounding time-frequency slots within a local neighborhood. We integrate this into the FCM update equations and evaluate its effect on mask estimation.

Moreover, to the best of our knowledge, the sFCM has not yet been investigated within the BSS framework for speech applications. We adapt the sFCM to suit our weighted FCM within the MENUET-based BSS scheme, and evaluate its ability for time-frequency mask estimation. Furthermore, in contrast to previous studies [11] we use a nonlinear microphone array in an under-determined setting, and evaluate the performance of the proposed scheme in both simulated and real-world conditions including public benchmark data, in the presence of reverberation and/or environmental noise.

7.2 Weighted FCM

7.2.1 Computation of weights

In the weighted FCM (wFCM) algorithm, we compute weights for each feature vector in \( Y = \{ y(\tau, f) | y(\tau, f) \in \mathbb{C}^M, (\tau, f) \in \Omega \} \) as a measure of its reliability. We base these weights on previous works by Kühne et al. and estimate a set of weights, \( W = \{ w_{\tau f} | w_{\tau f} \in \mathbb{R}^+, (\tau, f) \in \Omega \} \) using the variations in the feature values in the surrounding time-frequency slots for a nominated neighborhood \( N_{\tau f} \). These variations in feature values within local neighborhoods indicate whether it is dominated by a single source or multiple sources, as when there are multiple active speakers we expect to have a high variation in feature values [8, 11]. Furthermore, we also expect a high variation within the neighborhood when the features are unreliable due to sources of uncertainty such as high reverberation or background noise.

We define the neighborhood \( N_{\tau f} \) as

\[
N_{\tau f} = \{ (\tau', f') : |\tau' - \tau| \leq d_\tau, |f' - f| \leq d_f \},
\]

(7.1)

where \( d_\tau \) and \( d_f \) control the size of the neighborhood in the time and frequency directions respectively (i.e. number of time frames and frequency bins). We then compute the variation
\( v_{\tau f} \) within \( N_{\tau f} \) as
\[
v_{\tau f} = \frac{1}{|N_{\tau f}| - 1} \sum_{(\tau',f) \in N_{\tau f}} (y_{\tau' f} - y_{\tau f})^2,
\]  
(7.2)
where \(|\cdot|\) denotes the cardinality operator. A lower value in \( v_{\tau f} \) will yield more weight to that feature value during the clustering, as it implies greater reliability. The following empirically determined function was used for the weights [11]
\[
w_{\tau f} = 1 + \frac{1}{\max\{v_{\tau f}, \zeta\}},
\]
(7.3)
where \( \zeta \) is a constant to prevent division by zero, and which controls the upper limit of the weights. The weights function in (7.3) assigns large weights \( w(\tau, f) \gg 1 \) to regions with a low variance, whilst assigning a unity weight to regions with a high variance [11]. In our implementation, \( \zeta \) was set to \( 10^{-3} \). Furthermore, we introduce a weights exponent parameter \( p_w \) to control the influence of the variance weighting:
\[
w_{\tau f} \leftarrow w_{\tau f}^{p_w}.
\]
(7.4)
Note that a value of \( p_w = 0 \) sets the wFCM to the standard FCM.

### 7.2.2 wFCM clustering algorithm

We modify the cost function of the standard FCM algorithm as described in Section 4.2.3 to include the weights defined above. The cost function of the wFCM becomes
\[
\mathcal{J}_{\text{wFCM}} = \sum_{n=1}^{N} \sum_{\Omega} u_{n\tau f}^q w_{\tau f} \|y_{\tau f} - v_{n,w\text{FCM}}\|^2.
\]
(7.5)
As with the FCM, the minimization of \( \mathcal{J}_{\text{wFCM}} \) in (7.5) is solved using Lagrange multipliers, which results in the following alternating update equations:
\[
v_{n,w\text{FCM}} = \sum_{\Omega} u_{n\tau f}^q w_{\tau f} y_{\tau f},
\]
(7.6)
\[
u_{n\tau f} = \left[ \sum_{i=1}^{N} \left( \frac{\|y_{\tau f} - v_{i,w\text{FCM}}\|^2}{\|y_{\tau f} - v_{n,w\text{FCM}}\|^2} \right)^{\frac{1}{q-1}} \right]^{-1}.
\]
(7.7)
The details of this derivation are in Appendix A.3, and the wFCM algorithm is summarized below in Algorithm 4. We note that the update for the membership is unchanged to that of the standard FCM in (4.29). The addition of the weights in the centroid update equation allows for the inclusion of reliability information into the estimation of the centroid locations. The separation mask is then taken from the final estimates of the wFCM membership partition matrix, \( U_{wFCM} \), as in Section 4.3.3.

**Algorithm 4** The weighted FCM (wFCM) clustering algorithm.

```plaintext
input: \( Y, W, N, N_{\tau f}, q, p_w, \epsilon \)
output: \( U_{wFCM}, V_{wFCM} = \{v_{wFCM}\}_{n=1}^N \)

1: initialize partition \( U_{wFCM}^{(0)} \)
2: repeat for \( r = 1, 2, \ldots \)
3: update centroids \( V_{wFCM}^{(r)} \) with \( U_{wFCM}^{(r-1)} \) using (7.6)
4: update partition matrix \( U_{wFCM}^{(r)} \) with \( V_{wFCM}^{(r)} \) using (7.7)
5: until \( \| U_{wFCM}^{(r)} - U_{wFCM}^{(r-1)} \| < \epsilon \)
6: return \( U_{wFCM} \leftarrow U_{wFCM}^{(r)} \) and \( V_{wFCM} \leftarrow V_{wFCM}^{(r)} \)
```

### 7.3 Weighted spatial FCM

#### 7.3.1 Contextual term

As mentioned in Section 7.1, the mask estimation procedure using both the standard FCM and wFCM do not consider the contributions of surrounding time-frequency slots. We modify the wFCM to incorporate the time-frequency information of the surrounding slots via adaptation of the sFCM and we denote this as the weighted spatial FCM (wsFCM). We introduce a contextual term \( c_{n\tau f} \), where \( c_{n\tau f} \in \mathbb{R}^+ \), that provides a measure of the degree to which the slots in a local neighborhood \( N_{\tau f} \) around \((\tau, f)\) are assigned to the \( n^{th} \) cluster.

Using the same neighborhood as defined in (7.1), we compute the contextual term as

\[
c_{n\tau f} = \sum_{(\tau', f') \in N_{\tau f}} u_{n\tau'f'}, \tag{7.8}
\]

and we introduce the context weighting parameter \( p_c \) to control the relative degree of
influence of $c_{n\tau f}$:

$$c_{n\tau f} \leftarrow c_{n\tau f}^{p_c}. \tag{7.9}$$

Note that $p_c = 0$ will default the wsFCM to the wFCM.

As mentioned in Section 7.2.1, time-frequency regions with low variation in feature values indicate a higher likelihood of being dominated by a single source. The contextual term seeks to exploit this and represents the likelihood that the feature vector at the time-frequency slot $(\tau, f)$ belongs to the $n^{th}$ cluster.

### 7.3.2 wsFCM clustering algorithm

The membership partition update equation in (7.7) is modified from $u_{n\tau f}$ to $u'_{n\tau f}$ to incorporate the contextual information as follows:

$$u'_{n\tau f} = c_{n\tau f} u_{n\tau f}$$

$$= c_{n\tau f} \left[ \sum_{i=1}^{N} \left( \frac{||y_{\tau f} - v_{i,wsFCM}||^2}{||y_{\tau f} - v_{i,wsFCM}||^2} \right)^{\frac{1}{q-1}} \right]^{-1}. \tag{7.10}$$

The centroid update remains as in (7.6).

We can see from (7.10) that when the neighborhood has a high degree of homogeneity and thus a larger $c_{n\tau f}$ value, the updated membership value $u'_{n\tau f}$ will be strengthened from its original in $u_{n\tau f}$. Conversely, should the neighborhood exhibit differing values and a low $c_{n\tau f}$, the value of $u_{n\tau f}$ will be reduced.

The combined wsFCM clustering algorithm, summarized in Algorithm 5, is a two-stage process at each iteration. The first stage follows the wFCM and computes the centroids and partition memberships as in (7.6) and (7.7). The second stage computes $c_{n\tau f}$ and uses this to update the partition memberships using (7.10). These updated membership values are used for the following iteration. The final wsFCM partition memberships $u'_{n\tau f}$ in $U_{wsFCM}$ are then used as the separation masks.
Algorithm 5 The weighted spatial FCM (wsFCM) clustering algorithm.

<table>
<thead>
<tr>
<th>line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>initializes partition $U_{wsFCM}^{(0)}$</td>
</tr>
<tr>
<td>2:</td>
<td>repeat for $r = 1, 2, \ldots$</td>
</tr>
<tr>
<td>3:</td>
<td>update centroids $V_{wsFCM}^{(r)}$ with $U_{wsFCM}^{(r-1)}$ using (7.6)</td>
</tr>
<tr>
<td>4:</td>
<td>update partition matrix $U_{wsFCM}^{(r)}$ with $V_{wsFCM}^{(r)}$ using (7.10)</td>
</tr>
<tr>
<td>5:</td>
<td>compute contextual term $c_{nrf}$ using (7.8) and update $U_{wsFCM}^{(r)}$</td>
</tr>
<tr>
<td>6:</td>
<td>until $|U_{wsFCM}^{(r)} - U_{wsFCM}^{(r-1)}| &lt; \varepsilon$</td>
</tr>
<tr>
<td>7:</td>
<td>return $U_{wsFCM} \leftarrow U_{wsFCM}^{(r)}$ and $V_{wsFCM} \leftarrow V_{wsFCM}^{(r)}$</td>
</tr>
</tbody>
</table>

7.4 Comparison of effect of weights and contextual information on mask estimation

The integration of weights and contextual information as described in Sections 7.2 and 7.3 yields four variations of FCM-based clustering: the standard FCM ($p_c = p_w = 0$), weighted FCM (wFCM, where $p_w > 0$, $p_c = 0$), spatial FCM (sFCM, where $p_w = 0$, $p_c > 0$), and weighted spatial FCM (wsFCM, where $p_c > 0$, $p_c > 0$).

Figures 7.1 and 7.2 demonstrate an example of the mask estimation ability of the four FCM algorithms. The masks were estimated for the same mixtures of four speech signals with the four source and three microphone configuration in Figure 5.1, and the figures demonstrate the estimated separation mask for the same source. In Figure 7.1 the reverberation time was set to $RT_{60} = 0$ ms, while in Figure 7.2 the reverberation was $RT_{60} = 300$ ms. The microphone to source distance was set to $R = 120$ cm. In the color scale of the masks the red extreme denotes the dominant speech signal ($u_{nrf} \approx 1$) and the blue denotes the interfering speakers ($u_{nrf} \approx 0$). The colors between these two extremes denote the data points with uncertain memberships, where the cause of uncertainty is possible due to non-ideal effects such as reverberation or a lack of robustness in the clustering algorithm.

As evident in Figure 7.1, all four FCM-based clustering algorithms had accurate mask estimation abilities in anechoic conditions. Given that there are no sound reflections and therefore minimal inaccuracies in the feature extraction method of Section 4.1.3, we expect
these clustering algorithms to perform well. However, there is a distinction between the clustering methods without contextual information (FCM and wFCM) and those with the contextual information (sFCM and wsFCM), as the masks of the FCM and wFCM appear slightly more flecked than the sFCM and wsFCM.

When the reverberation time is increased to 300 ms in Figure 7.2 the difference in mask estimation quality between the clustering algorithms is heightened. Due to the reduction in accuracy in the feature extraction in the presence of reverberation, we expect that many of the values in $Y$ will differ from their true values. As such, the standard FCM will struggle with classifying it to its correct cluster, as shown in Chapter 6. However, the sFCM and wsFCM have much less flecked masks than the FCM and wFCM, and as such, the inclusion of the contextual information results in more accurate estimated separation masks. We explore this further in the experimental sections to follow.

Figure 7.1: Comparison of estimated time-frequency masks for a case with three microphones and four sources. (a) FCM, (b) wFCM (c) sFCM and (d) wsFCM. Red areas indicate higher membership values (dominant source speaker) and blue areas indicate lower membership values (interference speakers). The reverberation was set to $RT_{60} = 0$ ms.
Figure 7.2: Comparison of estimated time-frequency masks for a case with three microphones and four sources. (a) FCM, (b) wFCM (c) sFCM and (d) wsFCM. Red areas indicate higher membership values (dominant source speaker) and blue areas indicate lower membership values (interference speakers). The reverberation was set to RT$_{60}$ = 300 ms.

7.5 Evaluations

7.5.1 Simulated reverberant conditions

The experiments reported in Chapter 6 were repeated in order to compare the proposed wsFCM fairly against the standard FCM. Simulations were conducted for all variants of the FCM as described in Section 7.4. A wide variety of parameter combinations were tested against the simulated data, resulting in a combination which was empirically determined to provide robust results in differing reverberation conditions: $p_w = 0.25$, $p_c = 0.75$, $d_r = 2$, $d_f = 2$. The fuzzification parameter $q$ was set to 2 as in the previous experiments. These parameters were used for each of the experiments in this chapter.

The results for simulated conditions are shown in Figure 7.3. As immediately evident, there is a general increasing trend as we add in the improvements to the FCM for all performance ratios. In all cases the combined wsFCM is the leader, with the difference
between the wsFCM and baseline FCM increasing as the room reverberation is increased. For example, when $R = 50$ cm and $RT_{60} = 0$ ms, the difference between the FCM and wsFCM with respect to the SIR is approximately 2 dB, but when $RT_{60} = 450$ ms, the difference rises to nearly 4 dB.

We also note that in the hostile conditions of higher reverberation times at a distance $R = 120$ cm the SDR when $p_c > 0$ results in an unfavorable SDR value. Furthermore, the SAR at those conditions is also reduced, and the ISR has negligible improvement. Referring to the definitions of the BSS EVAL ratios in Section 5.5, it appears that the increase in SIR is at the expense of the spatial distortion error and artifacts error. Given that the purpose of the contextual term $c_{ntf}$ is to promote homogeneity this decrease in SDR and SAR is likely caused by insufficient smoothing. This could be remedied by increasing the value of the context weighting parameter $p_c$ [118]. However, care must be taken to not over-smooth or blur the resulting masks if the $p_c$ is increased.

To evaluate the statistical significance of the results on the simulated data we conducted a Student’s $t$-test on the results. As in the previous chapter, we employed a two-tailed distribution with unequal variances and assumed a 5% significance level. We compared the standard FCM against the wsFCM for the SDR and SIR, and the results are shown in Table 7.1. We can see that for most conditions, the results of the wsFCM are significant at the 5% significance level. We note that the high $p$-value for the SDR is when the performance of the FCM and wsFCM were approximately equal in anechoic conditions (cf. Figure 7.3a).
Figure 7.3: Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with simulated data, when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Figure 7.3: cont. Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with simulated data, when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Table 7.1: Results of the Student’s $t$-test on the SDR and SIR results of Figure 7.3. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.

<table>
<thead>
<tr>
<th>Condition</th>
<th>SDR</th>
<th>SIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RT_{60} = 0$ ms, $R = 50$ cm</td>
<td>0.50</td>
<td>0.004</td>
</tr>
<tr>
<td>$RT_{60} = 128$ ms, $R = 50$ cm</td>
<td>0.017</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$RT_{60} = 300$ ms, $R = 50$ cm</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$RT_{60} = 450$ ms, $R = 50$ cm</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$RT_{60} = 0$ ms, $R = 120$ cm</td>
<td>0.28</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$RT_{60} = 128$ ms, $R = 120$ cm</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$RT_{60} = 300$ ms, $R = 120$ cm</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$RT_{60} = 450$ ms, $R = 120$ cm</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
</tbody>
</table>

### 7.5.2 Real room recordings

The variants on the FCM for mask estimation were then tested on the real room recordings, collected as described in Section 5.2. We considered two microphone spacings: 4 cm and 8 cm, where the 8 cm spacing was deliberately included to evaluate the effects of spatial aliasing (cf. Section 4.1.3). Figure 7.4 shows the various performance ratios over the 20 combinations of 4 sources, each of length 10 s from the TIMIT database. It is clear that for all ratios and all values of $R$, both the sFCM and wsFCM systems offer significant improvement over the baseline FCM system. This is especially true when $R = 120$ cm, where the standard FCM and the wFCM completely fail at separation, as indicated by the negative SIR.

Figure 7.5 shows the performance ratios when the microphone spacing is increased to 8 cm, with the other conditions remaining unchanged. Comparing this figure to Figure 7.4, it is clear that the violation of the spatial aliasing theorem has a negative impact on the separation performance. However, the sFCM and wsFCM are able to achieve encouraging performance even when the sources are at a distance of $R = 120$ cm, suggesting that the contextual term offers increased robustness to spatial aliasing effects.
Figure 7.4: Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with real data collected in an office environment, when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

We also evaluated the statistical significance of the results in Figures 7.4 and 7.5 with the Student’s $t$-test. The $p$-values are shown in Table 7.2 and reflect the results in Figures 7.4 and 7.5, and again demonstrate the benefit of the contextual information and the minimal likelihood that the results are due to chance.
7.5.3 Benchmark data

We then evaluated the proposed clustering algorithms on benchmark data sets of the SiSEC: the 2008 “Under-determined speech and music mixtures” and the 2010 “Source separation in the presence of real-world background noise” development data sets as described in Section 5.3.

For the SiSEC 2008 data in Table 7.3, we note a similar trend as reported in the simulated and recorded evaluations, where the sFCM and wsFCM consistently outperformed the baseline FCM and even the wFCM. However, we note that the SDR values with the wFCM is very similar to, and even surpassing at times, that of the combined wsFCM. As
Table 7.2: Results of the Student’s $t$-test on the SDR and SIR results of Figures 7.4 and 7.5. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.

<table>
<thead>
<tr>
<th>Condition</th>
<th>SDR</th>
<th>SIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 4\text{ cm}, R = 50\text{ cm}$</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$d = 4\text{ cm}, R = 120\text{ cm}$</td>
<td>$p &lt; 0.001$</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$d = 8\text{ cm}, R = 50\text{ cm}$</td>
<td>0.002</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$d = 8\text{ cm}, R = 120\text{ cm}$</td>
<td>0.147</td>
<td>$p &lt; 0.001$</td>
</tr>
</tbody>
</table>

mentioned in Section 7.5.1 above, this can possibly be alleviated by increasing the value of $p_c$. Furthermore, as with the simulated recordings, the contextual term offers little benefit with respect to the artifacts error for the SiSEC 2008 data.

Nonetheless, the results we obtained from these recordings are comparable to the publicly available results from the campaign [106]; for example, the highest achieved average SDR and SIR for the “male4_liverrec_250ms_5cm” was 3.13 dB and 4.05 dB, by El Chami et al. [120], which employed a frequency bin-wise clustering approach to soft mask estimation using the interaural phase and level differences measured at the microphones. Our proposed wsFCM achieved an average SDR and SIR of 1.37 dB and 4.07 dB.

Similarly, the results of the proposed wsFCM system on the SiSEC 2010 data are comparable to the available results of the evaluation campaign [107]. Although we do not have access to the results in the development data set, we can compare our results to that of the test set (which were recorded in the same conditions). For example, Duong et al. reported an average SDR and SIR of 3.70 dB and 6.00 dB for the Square environment, whilst we reported an average of 2.25 dB and 15.13 dB. The generally higher results in the Square environment for the SIR can be attributed to the characteristics of the environment; the Cafeteria environment has reverberation, whilst the Square environment has little to no reverberation [105].
7.6 Summary

In this chapter, we examined extensions to the standard FCM algorithm for the purposes of time-frequency mask estimation. Previous clustering-based approaches to BSS have successfully used standard clustering techniques to estimate these masks, however, the approaches have generally disregarded the structured nature of speech signals. Motivated by the homogenous behaviour of speech signals, we modified the standard FCM algorithm to bias the clustering results in favor of cluster membership homogeneity within localized neighborhoods in the time-frequency space. We presented a two-stage algorithm: first, the estimation of observation weights, where these weights were calculated based on the variation in Euclidean distance within the time-frequency neighborhoods of a fixed size. Secondly, inspired by work in the image segmentation field, we integrated the contextual information of the time-frequency points with the introduction of a contextual term into the cluster membership update equation. We evaluated our proposed weighted spatial FCM algorithm on simulated data, real recordings and public benchmark data. We recorded notable improvements in the source separation ability, particularly in the higher reverberation instances. Our results were comparable against publicly published algorithms on international campaigns such as the SiSEC.

Whilst this chapter has explored advancements to the FCM, recent advances in clustering-based BSS have explored alternative generative clustering techniques, such as the GMM discussed in Chapters 4 and 6, or the WMM. The WMM has a small but significant usage within this time-frequency clustering-based BSS framework, and it is for this reason that we explore its suitability within the MENUET source separation scheme. This thus forms the focus of the following chapter.
Table 7.3: Source separation results with the SiSEC 2008 and 2010 data when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SDR (dB)</th>
<th>ISR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FCM</td>
<td>wFCM</td>
</tr>
<tr>
<td>male3_liverec_130ms_5cm</td>
<td>2.78</td>
<td>2.88</td>
</tr>
<tr>
<td>female3_liverec_130ms_5cm</td>
<td><strong>2.80</strong></td>
<td><strong>2.80</strong></td>
</tr>
<tr>
<td>male4_liverec_130ms_5cm</td>
<td>1.63</td>
<td><strong>2.43</strong></td>
</tr>
<tr>
<td>female4_liverec_130ms_5cm</td>
<td>2.05</td>
<td><strong>2.16</strong></td>
</tr>
<tr>
<td>male3_liverec_250ms_5cm</td>
<td>2.52</td>
<td><strong>2.69</strong></td>
</tr>
<tr>
<td>female3_liverec_250ms_5cm</td>
<td>1.94</td>
<td><strong>2.00</strong></td>
</tr>
<tr>
<td>male4_liverec_250ms_5cm</td>
<td>1.37</td>
<td><strong>1.88</strong></td>
</tr>
<tr>
<td>female4_liverec_250ms_5cm</td>
<td><strong>2.08</strong></td>
<td>0.95</td>
</tr>
<tr>
<td>Average for RT_{60} = 130 ms</td>
<td>2.31</td>
<td><strong>2.57</strong></td>
</tr>
<tr>
<td>Average for RT_{60} = 250 ms</td>
<td><strong>1.98</strong></td>
<td>1.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th>FCM</th>
<th>wFCM</th>
<th>sFCM</th>
<th>wsFCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev_2ch_3src_Ca_Ce_A</td>
<td>3.31</td>
<td>3.17</td>
<td><strong>3.35</strong></td>
<td>3.32</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_B</td>
<td>5.17</td>
<td>5.21</td>
<td><strong>5.51</strong></td>
<td><strong>5.51</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_A</td>
<td>3.80</td>
<td>3.82</td>
<td><strong>3.93</strong></td>
<td>3.92</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_B</td>
<td><strong>3.11</strong></td>
<td>3.08</td>
<td>3.08</td>
<td>3.06</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_A</td>
<td><strong>1.62</strong></td>
<td>1.62</td>
<td>1.40</td>
<td>1.39</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_B</td>
<td>1.54</td>
<td><strong>1.55</strong></td>
<td>1.31</td>
<td>1.32</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_A</td>
<td>3.92</td>
<td><strong>3.96</strong></td>
<td>3.77</td>
<td>3.78</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_B</td>
<td>2.76</td>
<td><strong>2.79</strong></td>
<td>2.50</td>
<td>2.51</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td>3.85</td>
<td>3.82</td>
<td><strong>3.97</strong></td>
<td>3.95</td>
</tr>
<tr>
<td>Average for Square</td>
<td><strong>2.46</strong></td>
<td><strong>2.48</strong></td>
<td>2.25</td>
<td>2.25</td>
</tr>
</tbody>
</table>
Table 7.3: cont. Source separation results with the SiSEC 2008 and 2010 data when the FCM, wFCM, sFCM and wsFCM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>Source</th>
<th>Duration</th>
<th>SiSEC 2008</th>
<th>SIR (dB)</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FCM</td>
<td>wFCM</td>
<td>sFCM</td>
</tr>
<tr>
<td>male3_liverec_130ms_5cm</td>
<td></td>
<td>5.15</td>
<td>5.84</td>
<td>6.40</td>
</tr>
<tr>
<td>female3_liverec_130ms_5cm</td>
<td></td>
<td>6.25</td>
<td>6.73</td>
<td>7.44</td>
</tr>
<tr>
<td>male4_liverec_130ms_5cm</td>
<td></td>
<td>3.08</td>
<td>4.31</td>
<td>4.63</td>
</tr>
<tr>
<td>female4_liverec_130ms_5cm</td>
<td></td>
<td>3.69</td>
<td>3.95</td>
<td>4.38</td>
</tr>
<tr>
<td>male3_liverec_250ms_5cm</td>
<td></td>
<td>4.57</td>
<td>4.76</td>
<td><strong>6.08</strong></td>
</tr>
<tr>
<td>female3_liverec_250ms_5cm</td>
<td></td>
<td>4.61</td>
<td>4.57</td>
<td><strong>6.22</strong></td>
</tr>
<tr>
<td>male4_liverec_250ms_5cm</td>
<td></td>
<td>2.84</td>
<td>1.58</td>
<td>1.13</td>
</tr>
<tr>
<td>female4_liverec_250ms_5cm</td>
<td></td>
<td>3.63</td>
<td>3.25</td>
<td>0.99</td>
</tr>
<tr>
<td>Average for RT&lt;sub&gt;60&lt;/sub&gt; = 130 ms</td>
<td></td>
<td>4.54</td>
<td>5.21</td>
<td>5.71</td>
</tr>
<tr>
<td>Average for RT&lt;sub&gt;60&lt;/sub&gt; = 250 ms</td>
<td></td>
<td>3.91</td>
<td>3.54</td>
<td>3.61</td>
</tr>
<tr>
<td>SiSEC 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_A</td>
<td></td>
<td>5.01</td>
<td>4.08</td>
<td>8.33</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_B</td>
<td></td>
<td>8.97</td>
<td>9.19</td>
<td><strong>12.47</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_A</td>
<td></td>
<td>6.56</td>
<td>6.68</td>
<td>8.69</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_B</td>
<td></td>
<td>6.92</td>
<td>7.04</td>
<td><strong>9.06</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_A</td>
<td></td>
<td>11.84</td>
<td>11.88</td>
<td><strong>13.46</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_B</td>
<td></td>
<td>13.09</td>
<td>13.23</td>
<td>15.78</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_A</td>
<td></td>
<td>14.52</td>
<td>15.03</td>
<td>17.60</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_B</td>
<td></td>
<td>11.65</td>
<td>12.05</td>
<td>13.48</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td></td>
<td>6.87</td>
<td>6.75</td>
<td>9.64</td>
</tr>
<tr>
<td>Average for Square</td>
<td></td>
<td>12.77</td>
<td>13.05</td>
<td>15.08</td>
</tr>
</tbody>
</table>
CHAPTER 7. ADVANCEMENTS ON THE FUZZY C-MEANS CLUSTERING
Chapter 8

Watson mixture model clustering

In Chapter 4 we described the generative GMM clustering algorithm, and its source separation ability with respect to time-frequency mask estimation was presented in Chapter 6. In this chapter, we consider another generative clustering technique based on the WMM that has gained recent attention in the field of BSS. The rest of the chapter is organized as follows. We introduce this mixture model and discuss its previous use within the BSS field in Section 8.1. In Section 8.2 we explain the algorithm in detail, with evaluations for a range of acoustic conditions presented in Section 8.3.

8.1 Introduction

Recent advances in the clustering-based approach to BSS have considered generative clustering techniques, such as the GMM [12, 16, 18]. A variant on the GMM based on the line orientation idea of O’Grady and Pearlmutter was employed by Sawada et al. for the frequency bin-wise clustering of normalized multichannel recordings [18]. This variant on the GMM can be considered to be approximately equivalent to a WMM [13, 89].

The WMM is a distribution defined on the unit hypersphere that is frequently employed in directional statistics [121], and has a small but documented usage in the clustering-based BSS framework. One of the first was by Tran Vu and Haeb-Umbach, where the complex WMM was used to model the whitened and normalized observation vectors [81]. It was proposed that due to the spatial diversity of the sources, the feature vectors would form
clusters on the unit hypersphere. This notion was consequently extended and evaluated in conjunction with beamforming/postfiltering for simultaneous BSS and noise suppression [122]. However, in both instances the WMM was used for frequency bin-wise clustering, and thus an additional permutation alignment stage was required prior to the postfiltering.

In subsequent work, Souden et al. utilized the WMM for online speech separation with a recursive EM algorithm [89]. Ito et al. used the WMM on normalized observation data in conjunction with time-varying, frequency-independent mixture weights for permutation-free BSS. This latest method is in contrast to the other WMM-based techniques [81,89,122] which all required an additional stage of permutation alignment following the clustering.

To the best of our knowledge, the WMM has yet to be applied within the BSS scheme detailed in Chapter 4. In light of the increasing interest of the WMM within this clustering-based BSS framework, we propose to evaluate the performance of the WMM for mask estimation and compare its performance against the other clustering techniques that have been used with the MENUMET, such as those outlined in Chapters 4 and 7 and in related work [22]. In contrast to some previous studies, we employ full-band clustering which eliminates any need for postprocessing of permutation alignment. Furthermore, unlike the recent work which utilized an over-determined setting with as many as eight microphones [81], and those which focused on even-determined environments with a stereo setup [13], we evaluate our work in under-determined settings with a nonlinear microphone array. We evaluate the proposed combined MENUMET-based scheme and WMM algorithm in a range of simulated and real-world conditions as described in Chapter 5.

8.2 Proposed full-band WMM

8.2.1 Probabilistic model

The features in $Y$ are mapped on the unit hypersphere in the $M$-dimensional feature space by the feature extraction method described in Section 4.1.3. We whiten and normalize our features (cf. Appendix C.1.1) in accordance with other studies that use the WMM for time-frequency clustering [13,81]. We propose to model the conditional statistics of the feature vectors in $Y$ by a complex-valued WMM as
\[ p(y_{rf} | \Theta) = \sum_{k=1}^{K} \alpha_k p(y_{rf} | a_k, \kappa_k) \] (8.1)

where \( \Theta \) is the parameter set which contains all the mixture model parameters. The complex Watson distribution is given by

\[ p(y_{rf} | a_k, \kappa_k) = \frac{(M - 1)!}{2\pi^M \mathfrak{M}(1, M, \kappa_k)} \exp(\kappa_k |a_k y_{rf}|^2) \] (8.2)

where \( M \) is the dimension of the clustering data (i.e. the number of microphones), \( \alpha_k \) is the mixture weight (probability), \( a_k \) is the mean orientation vector (centroid) of the \( k^{th} \) distribution and \( \kappa_k \) is the concentration parameter. The Hermitian transpose is denoted by \((\cdot)^H\) and \( \mathfrak{M}(\cdot, \cdot, \cdot) \) is Kummer's confluent hypergeometric function \([121,123]\). We denote the unknown mixture model parameter set as

\[ \Theta = \{ \theta_k \}_{k=1}^{K}, \] (8.3)

where the parameter set for each mixture is given by

\[ \theta_k = \{ \alpha_k, a_k, \kappa_k \} \] (8.4)

and the parameters \( \alpha_k \) and \( a_k \) satisfy the following constraints:

\[ \sum_{k=1}^{K} \alpha_k = 1, \| a_k \| = 1. \] (8.5)

### 8.2.2 Estimation of parameters by EM algorithm

To estimate the parameters in \( \Theta \), we consider the MAP estimation of the log-likelihood of the mixture model defined by

\[ \max_{\Theta} \log p(\Theta | Y) = \max_{\Theta} \left[ \sum_{\Omega} \log p(y_{rf} | \Theta) + \log p(\Theta) \right] \] (8.6)

subject to the constraints in (8.5). We assume independence between the vectors in \( Y \) and uniform prior distributions for the parameters in \( \Theta \).

We implement the iterative EM algorithm to optimize (8.6), which iterates the E-step and M-step until convergence.
In the E-step, we compute the posterior probability \( \gamma_{krf} = P(k|y_{rf}, \Theta') \) noting Bayes’ theorem:

\[
\gamma_{krf} = \frac{\alpha'_k p(y_{rf}|\theta'_k)}{\sum_{k=1}^{K} \alpha'_k p(y_{rf}|\theta'_k)}.
\]  

(8.7)

using the current estimate of the parameter set, \( \theta'_k = \{\alpha'_k, a'_k, \kappa'_k\} \).

In the M-step, we update \( \Theta \) by maximizing the auxiliary function of the data likelihood, \( Q(\Theta, \Theta') \), using the updated posterior probability values from the E-step in (8.7):

\[
Q(\Theta, \Theta') = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \log \alpha_k p(y_{rf}|\theta_k).
\]  

(8.8)

We expand (8.8) and neglect constants to yield:

\[
Q(\Theta, \Theta') = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \left( \log \alpha_k \right) - \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \log(2\pi^M \mathcal{M}(1, M, \kappa_k))
\]

\[
+ \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \kappa_k a_k^H R_k a_k
\]  

(8.9)

where \( R_k \) is defined as

\[
R_k = \frac{\sum_{\Omega} \gamma_{krf} y_{rf} y_{rf}^H}{\sum_{\Omega} \gamma_{krf}}.
\]  

(8.10)

To derive the update rules, we optimize \( Q(\Theta, \Theta') \) with respect to each parameter in \( \Theta \). The details are given in the Appendix A.4.

The update for the mean orientation \( a_k \) is given by the normalized eigenvector corresponding to the principal eigenvalue of \( R_k \).

The mixture weight update is shown to be

\[
\alpha_k = \frac{\sum_{\Omega} \gamma_{krf}}{|Y|},
\]  

(8.11)

where \(|.|\) denotes the cardinality.

For the update of the concentration parameter, we observe the following equation (cf. Appendix A.4):
where $r_k$ is the principal eigenvalue of $R_k$. We can approximately solve the above as [123]

$$
\kappa_k \approx \frac{M r_k - 1}{2 r_k (1 - r_k)} \left( 1 + \sqrt{1 + \frac{4(M + 1) r_k (1 - r_k)}{M - 1}} \right). \tag{8.13}
$$

After a suitable initialization (cf. Section 8.3) we iterate between the E-step and M-step until convergence is reached, where convergence is typically defined as an insignificant difference between successive parameter set estimates. The WMM is summarized in Algorithm 6.

**Algorithm 6** The Watson mixture model (WMM) clustering algorithm.

```
input: $Y$, $K$, $\epsilon$
output: $\gamma_{k\tau f}$
1: initialize parameter set $\Theta$
2: repeat
3: E-step: compute posterior probabilities $\gamma_{k\tau f}$ using (8.7)
4: M-step: update mean orientation vector $a_k$ using principal eigenvector of $R_k$ in (8.10)
5: M-step cont.: update mixture weight $\alpha_k$ using (8.11)
6: M-step cont.: update concentration parameter $\kappa_k$ using (8.13)
7: until difference between successive parameter set estimates less than termination threshold $\epsilon$
8: return $\gamma_{k\tau f}$
```

### 8.2.3 Mask estimation and source recovery

Since the features in $Y$ are mapped onto the unit hypersphere in the $M$-dimensional complex feature space, we expect that there will be $N$ distinct clusters formed on the complex hypersphere. Given that the features hold geometric information, each cluster should correspond to a source speaker.
Upon convergence of the EM algorithm the final estimate of the posteriors $\gamma_{krf}$ are used as the mask estimates. If $K > N$, we use the $N$ dominant components, as indicated by the mixture component weight values $\alpha_k$. The masks are then estimated as $[13, 89]$

$$M_{WMM,n}(r,f) = \gamma_{nrf},$$

(8.14)

where the index $n$ pertains to one of the dominant $N$ components.

### 8.3 Evaluations

The evaluations here followed the setup as in Chapter 5. Whilst the performance of the WMM clustering varies with the selection of an appropriate number of mixture components in a manner akin to the GMM, in order to provide a fair comparison against the GMM and other clustering techniques we set the number of clusters $K = N$ for all evaluations. This assignment of $K = N$ is also in line with other BSS studies which employ the WMM $[13, 89]$. However, we include the results of different $K$ values in the Appendix B.2 for the interested reader. The parameter set was initialized as follows, similar to other works that employ the WMM $[13]$: the mean orientation vectors $a_k$ were selected randomly from the data $Y$, the concentration parameters $\kappa_k$ were set to 20 and the mixture weight were set to $\alpha_k = \frac{1}{K}$. Convergence was considered to be achieved when the difference between the old and new mean orientation vectors was less than a predetermined threshold $\epsilon$ or when 100 iterations was reached.

In this chapter we chose to directly evaluate the WMM mask estimation against the other generative clustering technique of this thesis, the GMM, as well as the most advanced FCM from the previous chapter, the wsFCM.

#### 8.3.1 Simulated reverberant conditions

We firstly evaluated the WMM for simulated conditions, using the same set of utterances as used in Chapters 6 and 7. The results are given in Figure 8.1. We can see from here the superiority of the WMM with respect to the SDR, ISR and SIR. The SAR yields mixed results, with the wsFCM and WMM performing at a comparable level.
Figure 8.1: Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with simulated data, when the WMM was used for mask estimation, compared against the GMM and wsFCM. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Figure 8.1: cont. Source separation results for $R = 50$ and $120$ cm and a microphone spacing of 4 cm with simulated data, when the WMM was used for mask estimation, compared against the GMM and wsFCM. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Table 8.1: Results of the Student’s \( t \)-test on the SDR and SIR results of Figure 8.1. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.

<table>
<thead>
<tr>
<th>Condition</th>
<th>SDR</th>
<th>SIR</th>
<th>SDR</th>
<th>SIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT(_{60}) = 0 ms, ( R = 50) cm</td>
<td>( p &lt; 0.001 )</td>
<td>( p &lt; 0.001 )</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>RT(_{60}) = 128 ms, ( R = 50) cm</td>
<td>( p &lt; 0.001 )</td>
<td>( p &lt; 0.001 )</td>
<td>0.39</td>
<td>0.63</td>
</tr>
<tr>
<td>RT(_{60}) = 300 ms, ( R = 50) cm</td>
<td>( p &lt; 0.001 )</td>
<td>0.005</td>
<td>0.63</td>
<td>0.52</td>
</tr>
<tr>
<td>RT(_{60}) = 450 ms, ( R = 50) cm</td>
<td>( p &lt; 0.001 )</td>
<td>( p &lt; 0.001 )</td>
<td>( p &lt; 0.001 )</td>
<td>0.35</td>
</tr>
<tr>
<td>RT(_{60}) = 0 ms, ( R = 120) cm</td>
<td>( p &lt; 0.001 )</td>
<td>( p &lt; 0.001 )</td>
<td>0.30</td>
<td>0.016</td>
</tr>
<tr>
<td>RT(_{60}) = 128 ms, ( R = 120) cm</td>
<td>( p &lt; 0.001 )</td>
<td>0.002</td>
<td>( p &lt; 0.001 )</td>
<td>0.64</td>
</tr>
<tr>
<td>RT(_{60}) = 300 ms, ( R = 120) cm</td>
<td>( p &lt; 0.001 )</td>
<td>0.28</td>
<td>( p &lt; 0.001 )</td>
<td>0.60</td>
</tr>
<tr>
<td>RT(_{60}) = 450 ms, ( R = 120) cm</td>
<td>( p &lt; 0.001 )</td>
<td>0.06</td>
<td>( p &lt; 0.001 )</td>
<td>0.04</td>
</tr>
</tbody>
</table>

We evaluated the statistical significance of the SDR and SIR results and these are shown in Table 8.1. The \( p \)-values indicate the significant performance of the WMM over the other generative clustering technique of GMM, where nearly all the \( p \)-values are above the 5% significance level. However, the WMM results are only significant above the wsFCM in the particularly hostile conditions, for example, when the distance between the speakers and microphone is increased to \( R = 120\) cm and in echoic conditions.

### 8.3.2 Real room recordings

We then evaluated the WMM performance for real room recordings, with the setup outlined in Section 5.2. Figure 8.2 depicts the achieved ratios averaged across all recordings for different distances between the speakers and microphone array. Again, we notice a highly comparable performance against the wsFCM, with significant improvements over the GMM. This is especially notable in the achieved SIR. Furthermore, when the microphone spacing is increased to 8 cm, we see that the WMM is able to give satisfactory separation performance despite the spatial aliasing effects.

The statistical significance of the results are shown in Table 8.2. These values reflect the
Figure 8.2: Source separation results for $R = 50$ and 120 cm and a microphone spacing of 4 cm with real data collected in an office environment, when the WMM was used for mask estimation, compared against the GMM and wsFCM. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

trends in the results in Figures 8.2 and 8.3, with the WMM exhibiting statistical significance over the GMM in all conditions.
Figure 8.3: Source separation results for \( R = 50 \) and 120 cm and a microphone spacing of 8 cm with real data collected in an office environment, when the WMM was used for mask estimation, compared against the GMM and wsFCM. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.

Table 8.2: Results of the Student’s \( t \)-test on the SDR and SIR results of Figures 8.2 and 8.3. A two-tailed distribution was assumed for each test with unequal variances between the data sets. The sample size for each data set was 20.
8.3.3 Benchmark data

We evaluated the WMM for the public benchmark data of the SiSEC 2008 and 2010 as described in Section 5.3. Whilst the WMM was able to outperform the wsFCM in the simulated and real room recordings, we note that for the SiSEC data, it generally performs below the wsFCM, or at a comparable level. For example, in the SiSEC 2008 data, the average SDR for the RT$_{60}$ = 250 ms was equivalent for both the wsFCM and WMM at 1.54 dB. In the SiSEC 2010 data, the average SIR for the Cafeteria environment achieved by the WMM was 5.67 dB, whilst the wsFCM achieved an average SIR of 9.64 dB.

The SiSEC results of Chapter 6 where the HKM, GMM and FCM were evaluated also demonstrate the low performance of the GMM generative clustering algorithm compared to the discriminative algorithms of HKM and FCM. For example, for the SiSEC 2008 data in Table 6.4, the average SDR for the RT$_{60}$ = 250 ms when the HKM is used for mask estimation is 0.64 dB, whilst the GMM achieves just 0.01 dB. This trend is repeated in the SDR of the SiSEC 2010 data set, where the average SDR for the HKM over the Cafeteria data is 2.57 dB, whilst the GMM averages only -0.86 dB.

Given that the SiSEC data is recorded in real environments and with background noise, it is recommended to run further evaluations comparing the generative and discriminative models to investigate whether the presence of noise is an inhibiting factor in the performance of models such as the GMM and WMM.

8.4 Summary

In this chapter, we evaluated the WMM for clustering of observation data for time-frequency mask estimation. Recent developments in clustering-based BSS have implemented the WMM with success, however, not in the context of the MENUET system. Our evaluations included both simulated and real-world recordings, and confirmed its suitability for mask estimation. The WMM was shown to outperform the other generative model clustering algorithm used in this thesis, the GMM. Furthermore, the WMM also outperformed the wsFCM as proposed in Chapter 7 in simulated and real recordings when considering the SDR metric, but exhibited comparable performance with the SAR metric. However,
when we evaluated on the benchmark data of the SiSEC, the WMM failed to outperform the wsFCM. Given the reduced mask estimation ability of both the GMM and WMM in these conditions, it is recommended to run further evaluations to investigate the impact of background noise on generative clustering models.

Up until this point in the thesis, the number of clusters (and thus sources) are assumed to be known a priori. However, in a truly blind separation scheme, this information cannot be assumed to be readily available. As such, in the following chapter we introduce a novel approach to the source number estimation problem based on the FCM clustering algorithm.
Table 8.3: Source separation results with the SiSEC 2008 and 2010 data when the GMM, wsFCM and WMM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SDR (dB)</th>
<th>ISR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>wsFCM</td>
</tr>
<tr>
<td>male3_liverec_130ms_5cm</td>
<td>-0.54</td>
<td>2.68</td>
</tr>
<tr>
<td>female3_liverec_130ms_5cm</td>
<td>0.22</td>
<td>2.69</td>
</tr>
<tr>
<td>male4_liverec_130ms_5cm</td>
<td>0.55</td>
<td>2.16</td>
</tr>
<tr>
<td>female4_liverec_130ms_5cm</td>
<td>-0.55</td>
<td>2.01</td>
</tr>
<tr>
<td>male3_liverec_250ms_5cm</td>
<td>-0.17</td>
<td>2.65</td>
</tr>
<tr>
<td>female3_liverec_250ms_5cm</td>
<td>0.74</td>
<td>1.79</td>
</tr>
<tr>
<td>male4_liverec_250ms_5cm</td>
<td>-0.35</td>
<td>1.37</td>
</tr>
<tr>
<td>female4_liverec_250ms_5cm</td>
<td>-0.20</td>
<td>0.38</td>
</tr>
<tr>
<td>Average for RT&lt;sub&gt;60&lt;/sub&gt; = 130 ms</td>
<td>-0.08</td>
<td>2.39</td>
</tr>
<tr>
<td>Average for RT&lt;sub&gt;60&lt;/sub&gt; = 250 ms</td>
<td>0.01</td>
<td>1.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th>SDR (dB)</th>
<th>ISR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>wsFCM</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_A</td>
<td>0.15</td>
<td>3.32</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_B</td>
<td>-2.47</td>
<td>5.51</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_A</td>
<td>-1.22</td>
<td>3.92</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_B</td>
<td>0.12</td>
<td>3.06</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_A</td>
<td>-0.17</td>
<td>1.39</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_B</td>
<td>-1.23</td>
<td>1.32</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_A</td>
<td>-0.25</td>
<td>3.78</td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_B</td>
<td>-0.05</td>
<td>2.51</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td>-0.86</td>
<td>3.95</td>
</tr>
<tr>
<td>Average for Square</td>
<td>-0.43</td>
<td>2.25</td>
</tr>
</tbody>
</table>
Table 8.3: cont. Source separation results with the SiSEC 2008 and 2010 data when the GMM, wsFCM and WMM were used for mask estimation. Results given with respect to the SDR, ISR, SIR and SAR at the output, averaged over all sources in the mixture. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SIR (dB)</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>wsFCM</td>
</tr>
<tr>
<td>male3_liverec_130ms_5cm</td>
<td>-1.48</td>
<td>6.65</td>
</tr>
<tr>
<td>female3_liverec_130ms_5cm</td>
<td>-0.77</td>
<td><strong>7.67</strong></td>
</tr>
<tr>
<td>male4_liverec_130ms_5cm</td>
<td>-1.43</td>
<td><strong>4.95</strong></td>
</tr>
<tr>
<td>female4_liverec_130ms_5cm</td>
<td>2.05</td>
<td><strong>4.76</strong></td>
</tr>
<tr>
<td>male3_liverec_250ms_5cm</td>
<td>2.85</td>
<td>6.08</td>
</tr>
<tr>
<td>female3_liverec_250ms_5cm</td>
<td>0.85</td>
<td><strong>5.94</strong></td>
</tr>
<tr>
<td>male4_liverec_250ms_5cm</td>
<td>-2.03</td>
<td><strong>4.07</strong></td>
</tr>
<tr>
<td>female4_liverec_250ms_5cm</td>
<td>-1.99</td>
<td><strong>4.06</strong></td>
</tr>
<tr>
<td>Average for RT$_{60}$ = 130 ms</td>
<td>-0.41</td>
<td><strong>6.01</strong></td>
</tr>
<tr>
<td>Average for RT$_{60}$ = 250 ms</td>
<td>-0.08</td>
<td><strong>5.04</strong></td>
</tr>
<tr>
<td>SiSEC 2010</td>
<td>SIR (dB)</td>
<td>SAR (dB)</td>
</tr>
<tr>
<td></td>
<td>GMM</td>
<td>wsFCM</td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_A</td>
<td>-1.02</td>
<td><strong>8.29</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_B</td>
<td>0.89</td>
<td><strong>12.47</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_A</td>
<td>1.65</td>
<td><strong>8.76</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_B</td>
<td>0.13</td>
<td><strong>9.06</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_A</td>
<td>2.75</td>
<td><strong>13.43</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_B</td>
<td>4.45</td>
<td><strong>15.80</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_A</td>
<td>4.15</td>
<td><strong>17.69</strong></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_B</td>
<td>2.57</td>
<td><strong>13.61</strong></td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td>0.41</td>
<td><strong>9.65</strong></td>
</tr>
<tr>
<td>Average for Square</td>
<td>3.48</td>
<td><strong>15.13</strong></td>
</tr>
</tbody>
</table>
Chapter 9

Source number estimation via weighted adaptive FCM clustering

The time-frequency mask estimation approaches for BSS presented thus far have all required the number of sources to be known \textit{a priori}. In this chapter, we introduce a novel approach to the source number estimation problem based on an adaptive version of the standard FCM clustering algorithm. The rest of the chapter is organized as follows. We introduce the algorithm in Section 9.1 and discuss the algorithm in detail in Section 9.2. We present the results of our evaluations in Section 9.3.

9.1 Introduction

In the time-frequency clustering-based approach to BSS, most algorithms assume the number of sources, and thus clusters, are known \textit{a priori}. In this chapter, we introduce a novel approach to the source number estimation problem based on an adaptive version of the standard FCM clustering algorithm. The rest of the chapter is organized as follows. We introduce the algorithm in Section 9.1 and discuss the algorithm in detail in Section 9.2. We present the results of our evaluations in Section 9.3.

To this end, there have been a handful of algorithms devised for the purpose of source number estimation in audio applications \cite{12,20,21,31,124}, and whilst these have been successful, some required as many as five and sixteen microphones respectively \cite{20,21} which may not always be ideal especially in under-determined settings. There has been other work
with fewer microphones, for example, Araki et al. presented a method for simultaneous source number estimation and source separation with just three microphones [12]. This employed full-band GMM clustering of the DOA with a sparse prior probability on the distribution, however, this method was suited to conditions with little reverberation due to its assumption of anechoic propagation.

For the source number estimation problem in under-determined settings, the literature is limited. However, several clustering-based approaches to BSS demonstrate potential extensibility to the source number estimation problem. As discussed in Chapter 4, the MENUET algorithm and its modifications [22] utilize the HKM and FCM clustering algorithms respectively for the estimation of separation masks. In these schemes, the final cluster estimates were representative of the source signals, and the true number of sources was therefore required to be known.

An extension to the FCM has been proposed by Beringer and H"ullermeier where a local adaptive optimization scheme was applied for the autonomous determination of the optimal number of clusters [125]. This algorithm, termed the adaptive FCM (aFCM), required only an initial estimate of the number of clusters and employed a quality measure within a hill-climbing style procedure to autonomously determine the optimal number of clusters.

However, the work by Beringer and H"ullermeier was designed for the clustering of synthetic data streams rather than real-world multispeaker speech data as in our studies. As such, the quality measure used to evaluate the validity of the clusters did not provide the correct source count for audio data. In light of this, we investigated a number of alternative quality measures and composed a unique weighted sum in place of that used in the original aFCM.

Furthermore, real-world recordings of speech signals are often susceptible to outliers due to external sources of noise/interference, reverberation and non-ideal recording equipment. These outliers may compromise the quality of the clustering. Previous studies have proposed to weight the feature data in favor of reliable points in cluster centroid computation (cf. Section 7.1). In line with this notion, we introduce and customize a novel weighting of the feature data based on the relative amplitudes, and this is incorporated in the source number estimation algorithm.
Inspired by the promising work of the aFCM by Beringer and Hüllermeier and the data weighting by Kühne et al. we present an adaptive, hill-climbing algorithm for the source number estimation problem for audio sources with feature weighting. We modify the aFCM algorithm to incorporate a weighted sum of quality measures and we also scale the feature data in favor of the reliable data points. We evaluate our proposed method in real-world conditions and demonstrate that our proposed method is capable of estimating the number of sources in both even- and under-determined conditions without any \textit{a priori} knowledge.

9.2 Proposed source number estimation algorithm

Chapter 4 describes the general scheme for the clustering-based MENUET BSS algorithm. We base our proposed source number estimation system on this scheme: first, we generate features from the observed mixtures as in Section 4.1.3, we then weight these features for robustness and finally, use a modified version of the adaptive FCM clustering to determine the number of sources.

9.2.1 Calculation of weights

Given the assumption of the sparseness between the signals, it is reasonable to assume that not all of the time-frequency slots will contribute equally to the final source reconstructions. In the presence of reverberation the direct path will provide a higher initial response before the multipath reverberation effects become apparent. By favoring the time-frequency slots with higher level ratio amplitudes, we simultaneously preference this direct path and reduce the effect of random noise from the unused time-frequency slots.

To this end, we calculate a set of weights for the features at all time-frequency slots \(\{\delta_{\tau f}\}_{\Omega}\), where \(\delta_{\tau f} \in \mathbb{R}^+\), using the relative amplitude of the microphone observations in each time-frequency slot. The weights were designed such that the reliable features were given a higher weight without under-weighting the less reliable ones.

The weights are calculated as follows:

\[
\delta_{\tau f} = \nu_{\tau f}^{\log_{\max(x)}(\rho)},
\]

(9.1)
where \( v_{\tau f} \) encapsulates the relative amplitude as

\[
v_{\tau f} = \frac{||y_{\tau f}||}{\sum_{\Omega} ||y_{\tau f}||}
\]

where \( ||\cdot|| \) denotes the complex vector norm. The weights are set to lie within the range \((0, \rho]\), where \( \rho \) denotes the maximum amount that any time-frequency slot should be weighted above average. We do this by considering the upper bound on \( v_{\tau f} \), i.e. \( \max(v) \), and we calculate the weight by

\[
\delta_{\tau f} = v_{\tau f}^z,
\]

such that \( \max(v)^z = \rho \), to ensure \( \max(\delta_{\tau f}) = \rho \). This yields \( z = \log_{\max(v)}(\rho) \), and hence (9.1). For the application of this algorithm to source number estimation, the optimal value of \( \rho \) was empirically determined as \( \rho = 10 \).

9.2.2 Adaptive FCM clustering algorithm

We then cluster the features \( y_{\tau f} \) using the aFCM as proposed by Beringer and Hüllermeier. We modify the aFCM from the original to accommodate the weights in (9.1). In our weighted version of the aFCM (waFCM) we iteratively minimize the cost function:

\[
J_{\text{waFCM}} = \sum_{k=1}^{K} \sum_{\Omega} \delta_{\tau f} u_{\tau f}^q ||y_{\tau f} - v_{k,\text{waFCM}}||^2,
\]

where \( q \) defines the fuzziness of the membership (cf. Section 4.2.3), \( K \) is the number of clusters and \( v_{k,\text{waFCM}} \) is the centroid of the \( k \)th cluster. As with the FCM cost function (4.27), the minimization can be solved using Lagrange multipliers. We remark that the derivation of the update equations is akin to that of the wFCM, and we thus refer the interested reader to Appendix A.3 for the derivations. Beginning with a random partition matrix \( U_{\text{waFCM}} \), we alternate the following updates for the centroids and memberships

\[
v_{k,\text{waFCM}} = \frac{\sum_{\Omega} u_{\tau f}^q \delta_{\tau f} y_{\tau f}}{\sum_{\Omega} u_{\tau f}^q \delta_{\tau f}},
\]

\[
u_{k,\text{waFCM}}^z = \left[ \sum_{i=1}^{K} \left( \frac{||y_{\tau f} - v_{i,\text{waFCM}}||^2}{||y_{\tau f} - v_{k,\text{waFCM}}||^2} \right)^{\frac{1}{q-1}} \right]^{-1}.
\]

At each iteration of the waFCM, we test the quality of the solutions for the current cluster estimate \( K \) and its immediate neighbors, \( K - 1 \) and \( K + 1 \). This is done by computing the
corresponding partitions $U_{waFCM}$ for each of the three values $K$, and substituting it into the quality measure as described below in Section 9.2.3. The value of $K$ is then updated to that of $[K - 1, K, K + 1]$ with the highest quality, and the clustering continues until convergence is reached. Convergence is typically considered to be reached when the difference between successive partitions is sufficiently small [98].

### 9.2.3 Cluster quality measurement

The quality measurement used in the original aFCM was applied to synthetic data streams and was not suited to the source number estimation application of this study. As such, we evaluated a range of quality measures, which we denote by $Q_{BH}$, $Q_{XB}$, $Q_B$ and $Q_{FS}$ as in Table 9.1, and deduced that a combination of the measures was the best. The subscripts BH, XB, B and FS denote the author initials: Beringer&Hüllermeier [125], Xie&Beni [126], Bezdek [127] and Fukuyama&Sugeno [128]. The given measures were designed for use in clustering algorithms to estimate the quality based on a balance between the intra-cluster spread and the inter-cluster distance. We modified the measures to include the weights $\delta_{rf}$ as in (9.1), and a summary of these can be found in Table 9.1.

The quality measure $Q_{BH}$ designed by the authors of the aFCM was inspired by the Xie-Beni quality index. The Xie-Beni quality index $Q_{XB}$ was initially designed for image segmentation, and is a widely-used measure which provides a ratio of the intra-cluster to inter-cluster variability [126]. The partition entropy $Q_B$ measure was developed by Bezdek for automatic classification and considers only the membership partition in its calculation [127]. Finally, the index proposed by Fukuyama and Sugeno denoted by $Q_{FS}$ incorporates the cluster centroid locations, feature data and partitions in computing the index [128].

We combine the different quality measures in Table 9.1 to utilize the advantages of each while minimizing the disadvantages. The quality measures were normalized and combined in a voting system; in this way, each of the quality measures provides a best estimate for each of the cluster numbers in $[K - 1, K, K + 1]$. We normalize the quality measures to the unit interval as

$$
\overline{Q}_* = \frac{Q_* - Q_{*, \text{worst}}}{Q_{*, \text{best}} - Q_{*, \text{worst}}},
$$

(9.7)
Table 9.1: The different quality measures included in this study in the estimation of the cluster quality (cf. Section 9.2.3). All measures are shown with our proposed weights \( \delta_{rf} \) included. The centroids, membership value and partition matrix are indicated by \( v_k, u_{k,rf} \) and \( U \) respectively. \( y_{rf} \) denotes the mean of the feature data set, and \( | \cdot | \) denotes the cardinality.

<table>
<thead>
<tr>
<th>Quality measure</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Q_{BH} ) [125]</td>
<td>[ \frac{1}{</td>
</tr>
<tr>
<td>( Q_{XB} ) [126]</td>
<td>[ \frac{1}{K} \sum_{k=1}^{K} \sum_{l \in \Omega} \delta_{rf} u_{k,rf}^q | y_{rf} - v_k |^2 / \min_{i,j} (v_i - v_j) ]</td>
</tr>
<tr>
<td>( Q_B ) [127]</td>
<td>[ -\frac{1}{</td>
</tr>
<tr>
<td>( Q_{FS} ) [128]</td>
<td>[ \sum_{k=1}^{K} \sum_{l \in \Omega} \delta_{rf} u_{k,rf}^q (| y_{rf} - v_k |^2 - | \bar{y}_{rf} - v_k |^2) ]</td>
</tr>
</tbody>
</table>

where \( * \in \{ BH, XB, B, FS \} \) denotes the algorithm used to generate the quality measure, and worst and best denote the worst and best quality measures respectively. The best and worst measures are defined according to the properties of the measures: for all four measures \( Q_{BH}, Q_{XB}, Q_B \) and \( Q_{FS} \) a smaller value was deemed better. Therefore, we selected the minimum of each as the \( Q_{*,\text{best}} \) and the maximum of each as the \( Q_{*,\text{worst}} \).

We then follow the statistics notion of a weighted sum and combine the four quality measures with appropriately selected weights as

\[
Q = w_{BH} Q_{BH} + w_{XB} Q_{XB} + w_B Q_B + w_{FS} Q_{FS},
\]

(9.8)

where \( w_{BH}, w_{XB}, w_B, w_{FS} \) denote the weights. The optimal values for the weights were empirically determined as \([w_{BH}, w_{XB}, w_B, w_{FS}] = [0.9, 1, 0.9, 1] \). The algorithm for source number estimation is summarized in Table 7.
Algorithm 7 Proposed weighted adaptive FCM (waFCM) clustering algorithm for source number estimation.

\textbf{input:} $\mathbf{Y}$, $\{\delta_{rf}\}_\Omega$, $K_{\text{init}}$, $q$

\textbf{output:} $K$

1: initialize partition matrix $\mathbf{U}^{(0)}_{\text{waFCM}}$
2: \textbf{repeat} for $r = 1, 2, \ldots$
3: update centroids $\mathbf{V}^{(r)}_{\text{waFCM}} = \{v_{k,\text{waFCM}}\}_{k=1}^K$ with $\mathbf{U}^{(r-1)}_{\text{waFCM}}$ using (9.5);
4: update partition matrix $\mathbf{U}^{(r)}_{\text{waFCM}}$ with $\mathbf{V}^{(r)}_{\text{waFCM}}$ using (9.6);
5: compute partition matrix for $K - 1$ and $K + 1$ and denote as $\mathbf{U}^{(r)}_{\text{waFCM},-1}$ and $\mathbf{U}^{(r)}_{\text{waFCM},+1}$ respectively;
6: find best partition among $\mathbf{U}^{(r)}_{\text{waFCM},-1}$, $\mathbf{U}^{(r)}_{\text{waFCM}}$, $\mathbf{U}^{(r)}_{\text{waFCM},+1}$ using $Q$ in (9.8);
7: update $K \leftarrow \arg \max_{K-1, K, K+1} \{Q_{K-1}, Q_K, Q_{K+1}\}$;
8: update iteration number $r = r + 1$;
9: \textbf{until} $||\mathbf{U}^{(r)}_{\text{waFCM}} - \mathbf{U}^{(r-1)}_{\text{waFCM}}||$; \textbf{return} $\mathbf{U}^{(r)}_{\text{wFCM}} \leftarrow \mathbf{U}^{(r)}_{\text{wFCM}}$ and $\mathbf{V}^{(r)}_{\text{wFCM}} \leftarrow \mathbf{V}^{(r)}_{\text{wFCM}}$

9.3 Evaluations

9.3.1 Experimental setup

To verify the performance of our proposed algorithm we ran several tests for a variety of microphone and source signal configurations in real conditions. The details of the collection of the recordings are described in Section 5.2, and we numbered the microphones and sources as in Figure 9.1. We considered both even- and under-determined conditions, and we included a setup similar to that of the source number estimation scheme of Araki et al. [12] for comparison. The source signals were obtained from the TIMIT database and were looped to a common length of 10 s. The distance between the sources and microphones was $R = 120$ cm. For the initialization of the waFCM clustering, the membership matrix was randomly initialized with values in the interval $[0, 1]$ and the fuzzification parameter was set to $q = 2$. For all evaluations, the initial estimate of the number of clusters was set as $K_{\text{init}} = 3$. Full details of the experimental conditions are in Table 9.2.

We tested 20 trials per experimental setup, and presented the results with respect to the source number estimation accuracy, in accordance with the results of Araki et al. [12].
## Table 9.2: Experimental conditions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of microphones</td>
<td>$M = 2, 3$</td>
</tr>
<tr>
<td>Number of sources</td>
<td>$N = 2, 3, 4$</td>
</tr>
<tr>
<td>Signal length</td>
<td>10 s</td>
</tr>
<tr>
<td>Reverberation time</td>
<td>$RT_{60} = 390 \text{ ms}$</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>8 kHz</td>
</tr>
<tr>
<td>STFT window</td>
<td>Hann</td>
</tr>
<tr>
<td>STFT frame size</td>
<td>1024 (128 ms)</td>
</tr>
<tr>
<td>STFT frame shift</td>
<td>256 (32 ms)</td>
</tr>
</tbody>
</table>

Figure 9.1: Configuration of microphones and sources with labeling convention shown for source number estimation experiments.

This is computed as

$$\% \text{ ACC} = \frac{C}{T} \times 100,$$

(9.9)

where $C$ denotes the number of correct trials and $T$ denotes the total number of trials.
Table 9.3: Source number estimation accuracy for various microphone and source configurations. The recordings were collected in an office environment with reverberation time $RT_{60} = 390\text{ ms}$. Results are given with respect to the accuracy measured over 20 trials. The standard deviation is shown in the brackets.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M = 3, N = 2$</td>
<td>90 (0.32)</td>
</tr>
<tr>
<td>$M = 2, N = 2$</td>
<td>100 (0.00)</td>
</tr>
<tr>
<td>$M = 3, N = 3$</td>
<td>90 (0.32)</td>
</tr>
<tr>
<td>$M = 2, N = 3$</td>
<td>90 (0.32)</td>
</tr>
<tr>
<td>$M = 3, N = 4$</td>
<td>80 (0.42)</td>
</tr>
<tr>
<td>$M = 2, N = 4$</td>
<td>70 (0.48)</td>
</tr>
</tbody>
</table>

### 9.3.2 Source number estimation results

Table 9.3 summarizes the results of our proposed method, where the standard deviation is included in the brackets. We present only the results of the proposed waFCM as the original aFCM failed to count sources as mentioned in Section 9.1. In the even- and under-determined cases of two or three sources we consistently achieve at least 90% accuracy. However, when the number of sources was increased to 4 the accuracy drops. The results compare well with Araki et al. whose experiments were performed in an environment with $RT_{60} = 130\text{ ms}$ compared with our 390 ms. Also worth noting is the use of fewer microphones than previous works [20, 21] while still obtaining good accuracy.

### 9.3.3 BSS results

Whilst the focus of this chapter remains on blind source number estimation, we also evaluated the effect of the weights $\delta_{rf}$ as defined in (9.1) on the FCM for mask estimation. We assumed a priori knowledge on the number of sources, and we denote the FCM with the $\delta_{rf}$ weights as $\text{FCM}_\delta$ to distinguish it from the wFCM with $w_{rf}$ weights in Chapter 7. Evaluations were conducted for a range of number of sources and the number of microphones was fixed at three. The fifth source was located at 180° (cf. Figure 9.1).

Table 9.4 shows the separation results with respect to the energy-based criteria of BSS
Table 9.4: Source separation results for varying number of sources, $R = 120$ cm and a microphone spacing of 4 cm with real data collected in an office environment, when the FCM and FCM$_\delta$ were used for mask estimation. Results are averaged over all sources and 20 trials.

<table>
<thead>
<tr>
<th>Number of sources</th>
<th>SDR (dB)</th>
<th>ISR (dB)</th>
<th>SIR (dB)</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FCM</td>
<td>FCM$_\delta$</td>
<td>FCM</td>
<td>FCM$_\delta$</td>
</tr>
<tr>
<td>2</td>
<td>2.87</td>
<td>2.98</td>
<td>4.71</td>
<td>4.87</td>
</tr>
<tr>
<td>3</td>
<td>2.28</td>
<td>2.42</td>
<td>3.73</td>
<td>4.05</td>
</tr>
<tr>
<td>4</td>
<td>1.74</td>
<td>2.14</td>
<td>3.22</td>
<td>3.43</td>
</tr>
<tr>
<td>5</td>
<td>1.32</td>
<td>1.63</td>
<td>2.28</td>
<td>2.84</td>
</tr>
</tbody>
</table>

EVAL as defined in Section 5.5 when compared against the standard FCM. We see from here that the inclusion of $\delta$ offers significant improvements over the FCM, especially when $N$ is increased to $N = 5$, where the FCM algorithm completely fails at separation. We also note that the weights of this chapter offer a substantial benefit over the weights that were introduced in Chapter 7, where $w_{rf}$ previously only offered slight improvement over the FCM. Therefore, the weights in $\delta_{rf}$ not only assist in the source number estimation problem, but also offer added robustness for mask estimation.

### 9.4 Summary

In this chapter we introduced a novel approach to the source number estimation problem in BSS. The estimation of the number of sources is of paramount importance within the field, however, there are only a handful of algorithms dedicated to this application. We proposed to use a variation on the adaptive FCM algorithm, previously used in cluster number detection for synthetic data streams, in conjunction with data weighting. We estimated weights based on the level ratio amplitudes where the features classified as reliable were given higher weights. We also investigated several measures to evaluate the quality of the clusters and combined them in a voting system manner. We evaluated our scheme in a real office environment with various even- and under-determined microphone and source configurations. Our scheme was proven to be robust with source number estimation accuracies between
80% and 100%. We also evaluated the effectiveness of the feature reliability weights for mask estimation and noted significant improvements in source separation performance.

The method presented in this chapter employed full-band clustering. However, as discussed in Chapter 3, an increasing number of BSS algorithms are focusing on frequency bin-wise clustering methods. In the following chapter, we propose a novel algorithm which uses the results of frequency bin-wise clustering to estimate the number of sources.
Chapter 10

Source number estimation via clustering of speech activity sequences

The research contained in this chapter was conducted at the Nippon Telephone and Telegraph Communication Science Laboratories in Kyoto, Japan in conjunction with co-authors Nobutaka Ito, Mehrez Souden, Shoko Araki and Tomohiro Nakatani. The previous chapter presented a novel source number estimation technique based on alterations to the standard FCM. In this chapter, we present another novel source number estimation technique. We explore the full-band clustering of speech activity sequences by the WMM, which was investigated as a method of mask estimation for source separation in Chapter 8. The rest of the chapter is organized as follows. Section 10.1 introduces the algorithm, with the details discussed in Section 10.2. The results of the evaluations are presented in Section 10.3.

10.1 Introduction

As discussed in the previous chapter, the estimation of the number of sources is a difficult problem with potentially significant repercussions. Previous sparseness-based approaches to source number estimation have used full-band clustering techniques with success, such as the GMM-based approach by Araki et al. [12], and the FCM-based approach just presented
in Chapter 9. On the other hand, Sawada et al. [18] and Tran Vu and Haeb-Umbach [81] have explored frequency bin-wise clustering approaches for source separation for a known number of sources. However, the extension of frequency bin-wise clustering schemes to an unknown number of sources is not as straightforward as in the instance of full-band clustering, because a sparse prior in bin-wise clustering would result in inconsistent source number estimation across frequencies.

In this chapter, we propose a novel approach to source number estimation that is robust to reverberation consisting of two main stages: firstly, the bin-wise clustering of time-frequency slots, and secondly, the full-band clustering of the speech activity sequences. The bin-wise nature of the first clustering stage enables robustness to reverberation while the full-band nature of the second clustering enables full-band estimation of the source number. In the first stage, the reverberation-robust bin-wise clustering is employed to obtain speech activity sequences represented by posterior probabilities [18]. In the second stage, these posterior probabilities from all frequency bins are clustered using a WMM with a sparse prior to count sources. The underlying assumption of this approach is that the posterior probability sequences corresponding to the same source exhibit similar patterns along different frequency bins [18]. We evaluate our proposed scheme on various over-, even- and under-determined configurations in real room conditions over a range of reverberation times.

10.2 Proposed source number estimation algorithm

10.2.1 Estimation of speech activity sequences

Figure 10.1 depicts the overall processing flow of the proposed method. We consider the same observation model as defined in Section 4.1.1, and employ the same STFT analysis as in Section 4.1.2. Using the STFT representations of the microphone observations $x_m(\tau, f)$ we denote the observation vector $x(\tau, f) = [x_1(\tau, f), \ldots, x_M(\tau, f)]^T$, where $x(\tau, f) \in \mathbb{C}^M$, and calculate speech activity sequences [18]. We normalize $x(\tau, f)$ and cluster in a frequency bin-wise manner using a complex Gaussian density mixture model into $K$ clusters, where each cluster is associated with a steering vector of a source. The outcome of the clustering is a set
of posterior probability sequences, \( \{P(C_i|x(\tau, f))\}_{i\tau f} \) for all \( f \) and \( i \), where \( P(C_i|x(\tau, f)) \) is a time-series at a fixed value of \( f \) and \( i \). This posterior probability represents the likelihood that the observation \( x(\tau, f) \) belongs to the \( i^{th} \) cluster class. These posterior probabilities are representative of the speech activity present in the \( K \) clusters. In this algorithm, the sequences are used in the subsequent source number estimation stage. Details of the calculation of the speech activity sequences are included in Appendix C.1.

In the BSS algorithm to which the speech activity sequences were originally intended, the number of clusters \( K \) was set to equate the number of sources, of which \textit{a priori} knowledge was assumed. However, in this application, we set \( K \) to an overspecified number of clusters. The second stage speech activity clustering is then used to reduce \( K \) to \( N \), where \( N \) represents the true number of sources, given only the speech activity sequences \( \{P(C_i|x(\tau, f))\}_{i\tau f} \).

### 10.2.2 WMM for clustering sequences

At the conclusion of the bin-wise time-frequency clustering, we have the speech activity sequences at each frequency bin: a set of real-valued vectors \( \{P(C_i|x(\tau, f))\}_{i\tau f} \). For the purpose of source number estimation, we denote the set of these sequences as \( \{v(l)\}_{l=1}^{L} \), where \( v \) denotes the speech activity sequence and \( L \) the total number of available sequences. Since we have \( K \) clusters at each frequency bin and \( F \) total frequency bins, the number of sequences are \( L = K \cdot F \). Each sequence \( v(l) \) is \( T \)-dimensional, where \( T \) is the total number of time frames. We denote \( v(l) \) by \( v_l \).
CHAPTER 10. SOURCE NUMBER ESTIMATION VIA CLUSTERING OF SPEECH
ACTIVITY SEQUENCES

After unit normalization of the data in \( \{ v_l \}^L_{l=1} \), i.e. \( v_l \leftarrow v_l/\|v_l\| \), we can form an approximate model for the sequences by a real-valued WMM defined as follows:

\[
p(v_l|\Theta) = \sum_{k=1}^{K} \alpha_k p(v_l|a_k, \kappa_k).
\] (10.1)

This model is a mixture of \( K \) Watson distributions as

\[
p(v_l|w_k, \kappa_k) = \frac{\left( \frac{T}{2} - 1 \right)!}{2\pi^{T/2} M \left( \frac{1}{2}, \frac{T}{2}, \kappa_k \right)} \exp \left( \kappa_k (a_k^T v_l)^2 \right)
\] (10.2)

where \( M(\cdot, \cdot, \cdot) \) is Kummer’s confluent hypergeometric function [121, 123], \( \alpha_k \) is the \( k \)th mixture weight, and \( a_k \) and \( \kappa_k \) denote the mean orientation vector and concentration parameter of the \( k \)th mixture component respectively. Note that the index \( k \) denotes the cluster index of the speech activity clustering (i.e. WMM cluster index), in contrast to the bin-wise clustering, where \( i \) is the bin-wise cluster index. The same number of clusters in the WMM is assumed as in the bin-wise, i.e. the deliberate overspecification.

The parameter set of the mixture model is given as

\[
\Theta = \{ \alpha_k, a_k, \kappa_k \}^K_{k=1},
\] (10.3)

where the parameters \( \alpha_k \) and \( a_k \) satisfy the following constraints:

\[
\sum_{k=1}^{K} \alpha_k = 1, \|a_k\| = 1.
\] (10.4)

To reduce the number of non-empty clusters and to promote sparseness, we impose a Dirichlet prior on the mixture weight \( \alpha_k \) as [12]

\[
p(\{ \alpha_k \}^K_{k=1}) = \frac{\Gamma(K \cdot \phi)}{\Gamma(\phi)^K} \prod_{k=1}^{K} \alpha_k^{\phi-1},
\] (10.5)

where \( \Gamma \) denotes the gamma function and \( \phi \) is a hyperparameter controlling the sparseness of the Dirichlet distribution. The other parameters in the mixture model are assumed to have uniform priors, and as such \( p(\Theta) = p(\{ \alpha_k \}^K_{k=1}) \).
10.2.3 EM algorithm for MAP parameter estimation

Assuming the data in \( \{v_l\}_{l=1}^L \) are independent, we estimate the model parameters by the MAP estimation of \( \log p(\Theta|\{v_l\}_{l=1}^L) \) defined by

\[
\max_{\Theta} \log p(\Theta|\{v_l\}_{l=1}^L) = \max_{\Theta} \sum_{l=1}^L \log p(v_l|\Theta) + \log p(\Theta). \tag{10.6}
\]

This optimization can be efficiently executed via the EM algorithm. The E-step computes the posterior probability, \( \gamma_{kl} \), as

\[
\gamma_{kl} = \frac{p(v_l|\Theta')}{\sum_{k=1}^K p(v_l|\Theta')} = \frac{\alpha'_k \exp \left( \kappa'_k (a_k^T v_l)^2 \right)}{\sum_{k=1}^K \alpha'_k \exp \left( \kappa'_k (a_k^T v_l)^2 \right)} \tag{10.7}
\]

with the current parameter set estimate

\( \Theta = \{\alpha'_k, a'_k, \kappa'_k\}_{k=1}^K \). \tag{10.8} \]

The M-step maximizes the following \( Q \) function, using the posterior \( \gamma_{kl} \) in (10.7), with respect to each parameter in \( \Theta \). The \( Q \) function is defined as

\[
Q(\Theta, \Theta') = \sum_{k=1}^K \sum_{l=1}^L p(k|v_l, \Theta') \log p(v_l|\Theta) + \log p(\{\alpha_k\}_{k=1}^K) \\
= \sum_{k=1}^K \sum_{l=1}^L \gamma_{kl} \log \alpha_k p(v_l|k, a_k, \kappa_k) + (\phi - 1) \sum_{k=1}^K \log \alpha_k \\
= \sum_{k=1}^K \left( \sum_{l=1}^L \gamma_{kl} + (\phi - 1) \right) \log \alpha_k - \sum_{k=1}^K \left( \sum_{k=1}^L \gamma_{kl} \right) \log \mathfrak{M} \left( \frac{1}{2}, \frac{T}{2}, \kappa_k \right) \tag{10.9}
\]

\[
+ \sum_{k=1}^K \left( \sum_{l=1}^L \gamma_{kl} \right) \kappa_k a_k^T R_k a_k,
\]

where \( R_k \) is defined by

\[
R_k = \frac{\sum_{l=1}^L \gamma_{kl} v_l v_l^T}{\sum_{l=1}^L \gamma_{kl}}. \tag{10.10}
\]

The update rules for the parameters in \( \Theta \) are derived through the differentiation of (10.9) with respect to each parameter; the details of the derivation are omitted for brevity here, but included in Appendix C.2.
The update for the mean orientation vector in $a_k$ is given by the unit eigenvector corresponding to the maximum eigenvalue of $R_k$ in (10.10). For the mixture weights, the update equation is as follows:

$$\alpha_k = \frac{\sum_{l=1}^{L} \gamma_{kl} + (\phi - 1)}{L + K(\phi - 1)}.$$  \hspace{1cm} (10.11)

For the update of the concentration parameter, we observe the following equation:

$$\frac{\partial \ln \mathcal{M}}{\partial \kappa_k} \left( \frac{1}{2}, T, \kappa_k \right) = a_k^T R_k a_k = r_k,$$  \hspace{1cm} (10.12)

where $r_k$ is the principal eigenvalue of $R_k$. We approximately solve (10.12) as [123]:

$$\kappa_k \approx \frac{T}{2} r_k - 1 \left( 1 + \sqrt{1 + \frac{4}{3} \frac{r_k(1 - r_k)}{\frac{1}{2} \left( \frac{T}{2} - r_k \right)}} \right).$$  \hspace{1cm} (10.13)

Beginning with an appropriate initialization of the parameter set $\Theta$, we iterate between the E-step (10.7) and M-step (10.9), (10.11) and (10.13) until the change between successive parameter set estimates is negligible.

### 10.2.4 Adaptive mixture weight for source number estimation

The Dirichlet prior on $\alpha_k$ promotes sparseness in the mixture model, and thus assists in the reduction of $K$ towards $N$. However, it was determined in experiments that the standard Dirichlet prior was not always suitable, as an ideal global value of the hyperparameter $\phi$ could not be calculated for all source and microphone configurations. In light of this, a modification to this hyperparameter was required.

An adaptive prior for Gaussian mixture decomposition was introduced by Medasani and Krishnapuram [129, 130]. In this technique, a penalty term was integrated in the objective function to facilitate the calculation of the optimal number of mixture components. This calculation of the optimal number was systematically executed by the emphasis of strong components and the suppression of weak components. This has since been applied to other probabilistic distributions, for example to the Dirichlet distribution [131].

Inspired by this notion, we propose the introduction of an adaptive hyperparameter term $\psi$ in lieu of the static hyperparameter $\phi$. In a similar direction as the work proposed
CHAPTER 10. SOURCE NUMBER ESTIMATION VIA CLUSTERING OF SPEECH ACTIVITY SEQUENCES

Figure 10.2: Depiction of clusters in the proposed speech activity sequence clustering. The speech activity sequences are clustered with the WMM into $K$ speech activity clusters, with $N$ classified as active and $K - N$ classified as 'garbage'.

by Medasani and Krishnapuram, we modify the objective function in (10.9) as

$$Q(\Theta, \Theta') = \sum_{k=1}^{K} \sum_{l=1}^{L} \gamma_{kl} \log \alpha_k p(v_l | a_k, \kappa_k) + \psi \sum_{k=1}^{K} \log \alpha_k,$$

(10.14)

where $\psi$ is updated at each iteration of the EM algorithm as

$$\psi = \log \sum_{k=1}^{K} \frac{\sum_{l=1}^{L} \alpha_k p(v_l | a_k, \kappa_k)}{\sum_{k=1}^{K} \log \alpha_k}.$$

(10.15)

The update equation for the mixture weights in $\{\alpha_k\}_k$ then changes accordingly:

$$\alpha_k = \frac{\sum_{l=1}^{L} \gamma_{kl} + \psi}{L + K \psi}.$$

(10.16)

10.2.5 Decision of source number

Figure 10.2 depicts the resulting clusters from the WMM clustering described in Section 10.2.2, and Figure 10.3 depicts the corresponding binary indicator masks. Each mask depicts the binary probability (0 or 1) whether the posterior probability sequence belongs to that cluster. These resulting clusters are then processed to determine the number of active sources. At the correct convergence of the algorithm, there will be $N$ clusters corresponding to the speech sources, and $K - N$ clusters corresponding to the non-active clusters (i.e. the 'garbage' clusters). This is visible in Figure 10.3 where there are two empty clusters. We aim to correctly group these clusters into these two classifications, and then to determine the number of clusters in the non-garbage class as the true number of sources.
To assist in this, we form quantitative measures of the speech activity for each of the final cluster sets resulting from the speech activity clustering. We define these as:

$$\rho_k = \sum_{(\tau, f) \in C_k} P(C_k | x(\tau, f)), \quad (10.17)$$

where \(\{C_k\}_{k=1}^{K}\) denotes the \(K\) final WMM cluster sets. The \((\tau, f)\) indices are mapped to the speech activity sequence index \(l\) in the set \(\{v_l\}_{l=1}^{L}\), and thus to the speech activity cluster set \(C_k\), accordingly.

Since the values in \(\{\rho_k\}_{k=1}^{K}\) provide a measure of the speech activity in the \(k^{th}\) cluster, these values can be classified into active and non-active (‘garbage’) clusters. This is easily achieved using a standard classification algorithm, for example, the \(k\)-means algorithm.
Table 10.1: Experimental conditions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of microphones</td>
<td>$M = 2, 3$</td>
</tr>
<tr>
<td>Number of sources</td>
<td>$N = 2, 3, 4$</td>
</tr>
<tr>
<td>Signal length</td>
<td>8 s</td>
</tr>
<tr>
<td>Reverberation time</td>
<td>$RT_{60} = 130 \sim 445$ ms</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>16 kHz</td>
</tr>
<tr>
<td>STFT window</td>
<td>Hann</td>
</tr>
<tr>
<td>STFT frame size</td>
<td>2048 (128 ms)</td>
</tr>
<tr>
<td>STFT frame shift</td>
<td>512 (32 ms)</td>
</tr>
</tbody>
</table>

10.3 Evaluations

10.3.1 Experimental setup

The proposed source number estimation algorithm was evaluated on a variety of source and microphone configurations to encompass over-, even- and under-determined settings. The room configuration is depicted in Figure 10.4, with experimental conditions summarized in Table 10.1. The mixtures were generated by convolving clean English-speaking sources with the measured room impulse responses, in an enclosure of dimensions $4.45 \times 3.55 \times 2.50$ m. The loudspeakers and microphones were at an elevation of 120 cm. A total of 8 speech combinations per acoustic condition and configuration was created, in line with previous related studies [12, 13, 18].

Our preliminary experiments demonstrated that the behavior of the posterior probability speech activity sequences was not necessarily consistent over all frequency ranges. We therefore limited the frequency bin range for the second WMM clustering to within the mid-frequency range of approximately 4 kHz to 5.5 kHz (200 frequency bins). This resulted in the number of data points $L = K \cdot 200$, assuming the experimental conditions as in Table 10.1.

The parameter set for the bin-wise clustering was initialized as specified by Sawada et al. [18], and the EM algorithm for the speech activity clustering was initialized as follows,
similar to other works that employ the WMM [13, 81]: the mean orientation vectors \( a_k \) were selected randomly from \( \{ v_i \} \), \( \kappa_k = 20 \) and \( \alpha_k = \frac{1}{K} \). Convergence was considered to be obtained when the difference between the old and new mean orientation vectors was less than a predetermined threshold \( \epsilon \), or when 100 iterations was reached.

10.3.2 Results

Table 10.2 denote the results of the proposed algorithm, where accuracy is defined as the correct estimation of the exact number of sources (cf. Section 9.3.1, (9.9)). The number of clusters, \( K \), was varied in order to evaluate its effect on the algorithm accuracy. For over-determined and even-determined scenarios of up to three simultaneously active speech sources, perfect source counting was achieved for all values of \( K \); even in the higher room reverberation times of almost 450 ms.

In the more challenging under-determined setting, there is still notable source number estimation ability. For the case of three sources and two microphones, when \( K = 12 \) we note perfect estimation for all reverberation times. However, when four sources were present
three microphones, we note a reduction in accuracy, with the optimal number of clusters at $K = 10$ or $K = 12$ depending on the reverberation time.

As previously mentioned in Chapter 9, Araki et al. have proposed source number estimation for under-determined conditions, but the evaluations were restricted to a low reverberation time of 130 ms. In contrast, this method has demonstrated remarkable source number estimation ability even for high reverberation times up to 445 ms.

10.3.3 Comparison of source number estimation methods

Whilst the waFCM in Chapter 9 and the WMM clustering in this chapter present two novel approaches to source number estimation, each has its own advantages and disadvantages. For example, the waFCM is computationally expensive, since at each iteration of the algorithm the partition matrices for the adjacent $K$ values need to be computed. Furthermore, we were not able to evaluate the waFCM on as many reverberation times as the WMM clustering in this chapter due to a lack of resources. However, the waFCM method integrates well within the MENUET BSS system, and as shown in Section 9.3.3 the proposed weights $\delta_{rf}$ have the potential to improve BSS performance over the standard FCM.

The WMM based source number estimation in this chapter has the advantage of greater robustness against reverberation due to the bin-wise clustering in the estimation of speech activity sequences. Furthermore, the method in this chapter has the potential to be extended to a BSS system by adding a permutation alignment stage of the active speech activity sequences once the number of sources has been determined. However, this is at the expense of the inclusion of an additional stage, which may introduce more complexities.
CHAPTER 10. SOURCE NUMBER ESTIMATION VIA CLUSTERING OF SPEECH ACTIVITY SEQUENCES

Table 10.2: Source number estimation accuracy for various source and microphone configurations. Six different room reverberations were tested for the cluster numbers as indicated in the leftmost column. Results are given with respect to the accuracy measured over 8 trials.

<table>
<thead>
<tr>
<th>RT₆₀₀</th>
<th>130 ms</th>
<th>200 ms</th>
<th>245 ms</th>
<th>300 ms</th>
<th>375 ms</th>
<th>445 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition: M = 3, N = 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 8</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 12</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>Condition: M = 2, N = 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 8</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 10</td>
<td>100</td>
<td>88</td>
<td>100</td>
<td>100</td>
<td>88</td>
<td>100</td>
</tr>
<tr>
<td>K = 12</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>Condition: M = 2, N = 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>K = 8</td>
<td>88</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>88</td>
</tr>
<tr>
<td>K = 10</td>
<td>100</td>
<td>88</td>
<td>100</td>
<td>88</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 12</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>Condition: M = 3, N = 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 8</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>K = 12</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>Condition: M = 3, N = 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 6</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>K = 8</td>
<td>88</td>
<td>100</td>
<td>88</td>
<td>75</td>
<td>75</td>
<td>88</td>
</tr>
<tr>
<td>K = 10</td>
<td>88</td>
<td>100</td>
<td>100</td>
<td>88</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>K = 12</td>
<td>100</td>
<td>100</td>
<td>88</td>
<td>88</td>
<td>75</td>
<td>75</td>
</tr>
</tbody>
</table>
10.4 Summary

In this chapter we proposed a novel source number estimation algorithm based on the clustering of speech activity sequences with a real-valued WMM. We estimated such speech activity sequences by clustering the normalized observation vectors in a frequency bin-wise manner using a variant on the GMM based on the line orientation idea, where we estimated the number of clusters with an upper bound. These sequences were then modeled with a WMM and an adaptive variation on the Dirichlet distribution was implemented as a sparse prior of the mixture weights to promote the formation of empty ‘garbage’ clusters. We conducted experimental evaluations in real conditions with varying reverberation times for over-, even- and under-determined source and microphone configurations. Our proposed algorithm demonstrated successful source number estimation ability even in high reverberation times.

In the next chapter, we review the work presented in this thesis and comment on its limitations. We also discuss possible directions for future work in the field of clustering-based time-frequency mask estimation and source number estimation.
Chapter 11

Conclusions and future work

11.1 Conclusions

In this thesis we investigated and compared several time-frequency clustering techniques for
mask estimation in multichannel BSS. We implemented our algorithms within the MENUET
framework in a reverberant environment in simulated and real-world conditions. We also
considered the source number estimation problem and proposed two novel solutions. To
conclude we summarize our contributions and limitations of the work and we end the dis-
cussion with several directions for future research.

The thesis commenced with an introduction to methods of multichannel BSS. We formal-
ized the BSS problem and presented the three associated mixing models. Early techniques
of BSS based on ICA were introduced with the subsequent related techniques of NMF and
SCA. This led to the introduction of sparseness-based techniques for BSS, as well as those
based on the human auditory system and techniques which employ a distributed micro-
phone array. As the focus of the thesis is in time-frequency clustering approaches to BSS,
we also presented a comprehensive review of such BSS approaches from its origins to the
current state-of-the-art algorithms.

The contributions to the thesis began with the modification of the MENUET BSS
scheme. We proposed to improve upon the mask estimation stage, and we began by in-
vestigating the use of the FCM for a basic stereo setup. We compared this to published
work by Araki et al. with improvements in performance noted. Motivated by these im-
provenments we then applied it to the MENUET scheme. For fair comparison against other soft masking approaches we also implemented the GMM clustering, which is often used in clustering-based mask estimation methods in the BSS field. We conducted evaluations over simulated, real room and international benchmark data, and for all conditions the FCM proved a superior choice for mask estimation.

However, the FCM suffered from reduced performance at higher reverberation times and also when the source speakers were moved to a greater distance from the microphones. As such, we investigated advancements on the FCM in an effort to improve robustness in such conditions. Due to the fact that location features based on the level/phase ratios may not correspond to the correct sources in reverberant conditions, we incorporated a reliability weighting for the feature data and integrated this into the FCM cost function to yield the wFCM. However, given that recent research suggests the homogenous nature of speech signals in the time-frequency domain, we also sought to exploit this by including a contextual term, which was originally used in the field of image segmentation. This combination of weights and contextual term integrated into the FCM, which we termed wsFCM, resulted in substantial improvements in the separation performance, especially in the conditions where the FCM had previously failed to effectively separate.

Furthermore, we investigated recent advances in time-frequency clustering based approaches to BSS and were conscious of the increasing use of the WMM in this framework. Previous works which involved the WMM had not focused on full-band clustering with under-determined, nonlinear microphone arrangements. As such, we evaluated the WMM in the context of the MENUET, and compared its performance against the other generative model of the thesis, the GMM, and the proposed wsFCM of Chapter 7. Our results concluded the superior mask estimation ability of the WMM with higher performance ratios noted for most conditions. However, when we evaluated the WMM on the international benchmark data of the SiSEC, the WMM suffered reduced separation ability compared to the wsFCM. As the GMM also suffered from lower performance on this benchmark data set, it was recommended to conduct further evaluations in similar noise-corrupted environments to the SiSEC.

Up until this point in the thesis, all algorithms had assumed that the number of sources
were known to the system. However, to adopt a truly blind source separation scheme, such knowledge cannot be assumed to be known. As such, we developed a novel approach to source number estimation that can also potentially be applied for source separation. We modified an existing adaptive FCM scheme, which was originally designed to count the optimal number of clusters for synthetic data streams, to suit the audio application at hand. We introduced a novel weighting of the feature data to reduce the effects of outliers and also presented a unique quality measure which incorporated a range of existing quality measures. Our proposed source number estimation scheme was evaluated for a range of configurations in real room conditions with success. Furthermore, our proposed scheme also reported improvements in BSS compared to the standard FCM.

We also introduced another novel source number estimation scheme which employed a frequency bin-wise clustering in combination with full-band clustering. Using the speech activity sequences for an overspecified number of clusters as estimated from the frequency bin-wise clustering of the microphone observations, we used a WMM to estimate the number of active clusters. Such active clusters were deemed to contain speech activity, and were thus representative of a speaker. We conducted evaluations over a range of configurations in real room conditions with highly encouraging results.

\subsection{Limitations and future work}

As with any research, the work in this thesis has its limitations and we must consider these. Firstly, the experimental evaluations in this thesis considered two or three microphone arrangements located in the same $z$-plane. Whilst this was sufficient in demonstrating the effectiveness of our proposed techniques, a three-dimensional array with the microphones located at different elevations could potentially separate sources located at different elevations.

Secondly, our source separation work does not account for moving sources, which in real life are a natural occurrence. As such, the application of our work is limited to situations where the speakers do not move. This restriction could potentially be removed by considering a recursive clustering approach, such as the recursive WMM-based EM algorithm by
Souden et al. [89].

There are a number of possible future directions for research. The FCM clustering could be further improved by replacing the Euclidean $l_2$-norm distance with a more robust measure, for example the $l_1$ or kernel-based measures [11,132,133]. Furthermore, in the wFCM and wsFCM we assumed a neighborhood of fixed size, irrespective of the characteristics of the feature data in question. However, to truly encapsulate variations in the feature space, a neighborhood of varying dimensions is suggested [134]. This could be effected by defining a suitable criterion for which the neighborhood size can be made to vary.

Furthermore, given the success of the modifications on the FCM in the mask estimation, we propose modifications to the WMM to suit the application at hand. The work we presented in Chapter 8 used the standard WMM, however, the integration of contextual information in a similar spirit to the work in Chapter 7 may potentially result in better mask estimation. Alternatively, a sparse prior such as the Dirichlet distribution could be investigated [12,13,18].

The work in this thesis was largely focused on the mask estimation stage of the MENUET algorithm to improve BSS ability. However, another stage in the algorithm equally as significant in affecting the source separation ability is the feature extraction stage. It is known that in reverberant conditions, the performance of full-band clustering schemes is compromised when the features are based on differences between phases as measured by the microphones [18,85]. It is therefore proposed to replace or augment the current feature vectors with features that are potentially robust against such degrading effects, such as the CASA features of pitch or harmonicity.

With respect to the adaptive FCM source number estimation scheme of Chapter 9, future work should consider a mathematically-motivated derivation of the quality measure weights, as well as evaluations on international benchmark data, should suitable data become available. Furthermore, given the promising results of the adaptive FCM scheme with the inclusion of the amplitude-based weights, the inclusion of contextual information as with the wsFCM may improve the source number estimation accuracy. We can also consider the combination of weighted adaptive FCM for simultaneous source number estimation and mask estimation for a truly autonomous blind source separation scheme.
The reliability of the WMM-based source number estimation scheme could also be improved by considering a more sophisticated method of source activity detection. At present, we did a simple sum of the speech activity sequences classified to each WMM cluster and used the $k$-means to detect the active and garbage clusters. However, such a simple measure may not be truly reflective of the activity in each cluster, and as such, modeling the speech activity in the WMM clusters with a more suitable method could yield greater accuracy.

Finally, while the work of this thesis has been for the application of speech separation and number estimation, we must recognize that BSS systems are often applied across a range of situations, for example, as a front-end to automatic speech recognition. It is therefore recommended to evaluate the performance of the proposed algorithms in an alternative context, such as speech recognition, and to investigate its integrability. There are several works that explore the link between time-frequency masking and automatic speech recognition with promising potential [135, 136].

### 11.3 Summary

In summary, this thesis investigated methods of full-band time-frequency mask estimation within the MENUET framework. We devised and evaluated a range of suitable clustering algorithms: the GMM, FCM, wFCM, wsFCM and WMM. The improvements of our proposed mask estimation approaches confirm that there is potential within the full-band clustering approach to mask estimation, and that we should continue exploring this domain and keep it as an active field of research within the time-frequency clustering BSS community.

Furthermore, we also introduced two novel approaches to blind source number estimation. The first was based upon an adaptive FCM algorithm and the second based upon the full-band WMM clustering of speech activity sequences. Our promising results in a variety of conditions have significant repercussions and have opened up numerous pathways for future research in this near-dormant field.
Part I

Bibliography
Bibliography


Appendices
Appendix A

Derivation of GMM, FCM and WMM clustering update equations for mask estimation

A.1 Derivation of GMM update equations

In this section we present the derivation of the update equations for the GMM clustering as used in Section 4.2.2.

The clustering of the feature vectors $y_{rf} \in Y$ by the GMM is equivalent to the optimization of the parameter set $\Theta$. We consider the MAP estimation of the log-likelihood function, as defined in (4.21), and repeated below for reference:

$$\max_\Theta \left[ \sum_\Omega \log p(y_{rf}|\Theta) + \log p(\Theta) \right]$$ (A.1)

assuming uniform prior distributions of the mean and covariance $\mu_k$ and $\Sigma_k$, and subject to the constraint $\sum_{k=1}^K \alpha_k = 1$.

The optimization of (A.1) is efficiently achieved by the EM algorithm. We seek to maximize the auxiliary function $Q(\Theta, \Theta')$ in 4.23, and we expand it as follows [137]:
APPENDIX A. DERIVATION OF GMM, FCM AND WMM CLUSTERING UPDATE EQUATIONS FOR MASK ESTIMATION

\[
Q(\Theta, \Theta') = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{k \tau f} \log \alpha_k + \sum_{k=1}^{K} \sum_{\Omega} \gamma_{k \tau f} \log \rho(y_{\tau f}|\theta_k)
\]

\[
= \sum_{k=1}^{K} \sum_{\Omega} \gamma_{k \tau f} \left( \log \alpha_k - \frac{1}{2} \log(|\Sigma_k|) - \frac{1}{2} (y_{\tau f} - \mu_k)^T \Sigma_k^{-1} (y_{\tau f} - \mu_k) \right)
\]

To find the update equations for each of the parameters in \(\Theta\), we find the partial derivative of (A.2) with respect to each parameter.

To derive the update equation for \(\mu_k\), we consider terms of \(Q(\Theta, \Theta')\) that concern \(\mu_k\):

\[
\sum_{k=1}^{K} \sum_{\Omega} \gamma_{k \tau f} \left( -\frac{1}{2} \log(|\Sigma_k|) - \frac{1}{2} (y_{\tau f} - \mu_k)^T \Sigma_k^{-1} (y_{\tau f} - \mu_k) \right).
\]

We then compute \(\frac{\partial Q}{\partial \mu_k}\) and set to 0:

\[
\sum_{\Omega} \gamma_{k \tau f} \Sigma_k^{-1} (y_{\tau f} - \mu_k) = 0,
\]

and we obtain the update equation for \(\mu_k\) as

\[
\mu_k = \frac{\sum_{\Omega} \gamma_{k \tau f} y_{\tau f}}{\sum_{\Omega} \gamma_{k \tau f}}.
\]

For the covariance matrix \(\Sigma_k\) we rewrite the terms from \(Q(\Theta, \Theta')\) that concern \(\Sigma_k\), and by recalling some results from matrix algebra [137]:

\[
\sum_{k=1}^{K} \left( \frac{1}{2} \sum_{\Omega} \gamma_{k \tau f} \log(|\Sigma_k|^{-1}) - \frac{1}{2} \sum_{\Omega} \gamma_{k \tau f} \text{tr}(\Sigma_k^{-1} (y_{\tau f} - \mu_k)(y_{\tau f} - \mu_k)^T) \right)
\]

where \(\text{tr}(\cdot)\) denotes the trace of a matrix. We denote the matrix \((y_{\tau f} - \mu_k)(y_{\tau f} - \mu_k)^T\) by \(C\) for conciseness. At this point we recall some more matrix algebra calculus:

\[
\frac{\partial \log |A|}{\partial A} = 2A^{-1} - \text{diag}(A^{-1})
\]

\[
\frac{\partial \text{tr}(AB)}{\partial A} = B + T^T - \text{diag}(B)
\]

where \(A\) and \(B\) are matrices, \(|\cdot|\) denotes the determinant of a matrix and \(\text{diag}(\cdot)\) denotes the diagonal of a matrix.
We differentiate (A.6) with respect to $\Sigma_k$, recalling the relations in (A.7):

$$= \frac{1}{2} \sum_{\Omega} \gamma_{k\tau f} (2\Sigma_k - \text{diag}(\Sigma_k)) - \frac{1}{2} \sum_{\Omega} \gamma_{k\tau f} (2C - \text{diag}(C))$$

$$= \sum_{\Omega} \gamma_{k\tau f} (2(\Sigma_k - C)) - \frac{1}{2} \sum_{\Omega} \gamma_{k\tau f} \text{diag}(\Sigma_k - C).$$

Setting the derivative in (A.8) to 0, it implies that $\sum_{\Omega} \gamma_{k\tau f} (2(\Sigma_k - C)) = 0$. Therefore,

$$\sum_{\Omega} \gamma_{k\tau f} - \sum_{\Omega} \gamma_{k\tau f} C = 0,$$

and we thus obtain the update equation for $\Sigma_k$ as

$$\Sigma_k = \frac{\sum_{\Omega} \gamma_{k\tau f} (y_{\tau f} - \mu_k)(y_{\tau f} - \mu_k)^T}{\sum_{\Omega} \gamma_{k\tau f}}.$$ (A.10)

For the update of the mixture weights, we recall the constraint $\sum_{k=1}^{K} \alpha_k = 1$ and consider the terms of $Q(\Theta, \Theta')$ that concern $\alpha_k$. We introduce the Lagrangian multiplier $\lambda$ as:

$$\mathcal{L}(\alpha, \lambda) = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{k\tau f} \log \alpha_k + \lambda \left( \sum_{k=1}^{K} \alpha_k - 1 \right).$$ (A.11)

We differentiate (A.11) with respect to $\alpha_k$ and set to zero:

$$\alpha_k = -\frac{1}{\lambda} \left( \sum_{\Omega} \gamma_{k\tau f} \right).$$ (A.12)

We substitute this into the constraint on $\alpha_k$ to yield:

$$\lambda = -(|Y|),$$ (A.13)

and we thus obtain the update for $\alpha_k$ as

$$\alpha_k = \frac{\sum_{\Omega} \gamma_{k\tau f}}{|Y|}.$$ (A.14)

A.2 Derivation of FCM update equations

This section presents the derivation of the update rules for the fuzzy $c$-means clustering algorithm (cf. Section 4.2.3). The optimization is subject to the constraints


\[ u_{nrf} \in [0, 1], \sum_{\Omega} u_{nrf} = 1 \quad (A.15) \]

and we assume \( q > 1 \). By introducing a Lagrange multiplier \( \lambda \), we formulate the equation

\[
\mathcal{L}(U_{\text{FCM}}, \lambda) = \sum_{n=1}^{N} \sum_{\Omega} u_{nrf}^q \| y_{rf} - v_{n,\text{FCM}} \|^2 + \sum_{\Omega} \lambda \left( 1 - \sum_{n=1}^{N} u_{nrf} \right). \quad (A.16)
\]

To obtain the update equation for \( v_{n,\text{FCM}} \) we take the derivative of \( \mathcal{L} \) with respect to \( v_{n,\text{FCM}} \), i.e. \( \partial \mathcal{L} / \partial v_{n,\text{FCM}} \), and set to 0. This follows as

\[
-2 \sum_{\Omega} u_{nrf}^q (y_{rf} - v_{n,\text{FCM}}) = 0. \quad (A.17)
\]

The final update equation for the centroid is then expressed as

\[
v_{n,\text{FCM}} = \frac{\sum_{\Omega} u_{nrf}^q y_{rf}}{\sum_{\Omega} u_{nrf}^q}. \quad (A.18)
\]

To derive the update equation for the membership values, we compute \( \partial \mathcal{L} / \partial u_{nrf} \) and set to 0:

\[
q u_{nrf}^{q-1} \| y_{rf} - v_{n,\text{FCM}} \|^2 - \lambda = 0. \quad (A.19)
\]

Solving for \( u_n \) we obtain

\[
u_{nrf} = \left( \frac{\lambda}{q \| y_{rf} - v_{n,\text{FCM}} \|^2} \right)^{\frac{1}{q-1}}. \quad (A.20)
\]

Recall that \( \sum_{n=1}^{N} u_{nrf} = 1 \). We substitute (4.29) into this constraint to yield

\[
\sum_{n=1}^{N} \left( \frac{q \| y_{rf} - v_{n,\text{FCM}} \|^2}{\lambda} \right)^{-\frac{1}{q-1}} = 1, \quad (A.21)
\]

and we therefore obtain the final form of the update equation for \( u_n \):

\[
u_{nrf} = \left( \frac{\| y_{rf} - v_{n,\text{FCM}} \|^2}{\sum_{n=1}^{N} \| y_{rf} - v_{n,\text{FCM}} \|^2} \right)^{-\frac{1}{q-1}}. \quad (A.22)
\]
A.3 Derivation of wFCM update equations

This section presents the derivation of the update rules for the variance-weighted fuzzy c-means clustering algorithm (wFCM, cf. Section 7.2.2). We solve the minimization problem with Lagrange multipliers, and we form the Lagrangian function as

$$\mathcal{L}(U_{\text{wFCM}}, \lambda) = \sum_{n=1}^{N} \sum_{\Omega} u_{nrf}^q w_{rf} \|y_{rf} - v_{n,\text{wFCM}}\|^2 + \sum_{\Omega} \lambda \left(1 - \sum_{n=1}^{N} u_{nrf}\right)$$  \hspace{1cm} (A.23)

To obtain the update equation for $v_{n,\text{wFCM}}$, we differentiate $\mathcal{L}$ with respect to $v_{n,\text{wFCM}}$ and set it to 0. This yields:

$$-2 \sum_{\Omega} u_{nrf}^q w_{rf} \|y_{rf} - v_{n,\text{wFCM}}\| = 0. \hspace{1cm} (A.24)$$

The final update equation for the centroid is then expressed as

$$v_{n,\text{wFCM}} = \frac{\sum_{\Omega} u_{nrf}^q w_{rf} y_{rf}}{\sum_{\Omega} u_{nrf}^q w_{rf}}. \hspace{1cm} (A.25)$$

The membership update equation remains the same as for the standard FCM.

A.4 Derivation of WMM update equations

We present the derivation of the update equations used in Section 8.2.2. For reference, we repeat the expanded auxiliary function $Q(\Theta, \Theta')$:

$$Q(\Theta, \Theta') = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} (\log \alpha_k - \log(2\pi^M \mathcal{M}(1, M, \kappa_k)) + \kappa_k |a_k^H y_{rf}|^2)$$

$$= \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \log \alpha_k - \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \log(2\pi^M \mathcal{M}(1, M, \kappa_k)) + \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \kappa_k a_k^H R_k a_k$$  \hspace{1cm} (A.26)

where $R_k$ is defined as

$$R_k = \frac{\sum_{\Omega} \gamma_{krf} y_{rf} y_{rf}^H}{\sum_{\Omega} \gamma_{krf}}. \hspace{1cm} (A.27)$$
To derive the update equation for the mean orientation vector, we recall the constraint \( \| \mathbf{a}_k \| = 1 \) and we consider the terms of \( Q(\Theta, \Theta') \) in (A.26) that concern \( \mathbf{a}_k \). We introduce the Lagrangian multiplier \( \lambda \) as:

\[
\mathcal{L}(\mathbf{a}_k, \lambda) = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{k,\tau,f} \kappa_k \mathbf{a}_k^H \mathbf{R}_k \mathbf{a}_k + \lambda (\| \mathbf{a}_k \|^2 - 1).
\] (A.28)

We differentiate with respect to \( \mathbf{a}_k \) and set to 0:

\[
\mathbf{R}_k \mathbf{a}_k = - \frac{\lambda \mathbf{a}_k}{\kappa_k}.
\] (A.29)

Therefore, \( \mathbf{a}_k \) should be an eigenvector of the matrix \( \mathbf{R}_k \). Since \( \kappa_k > 0 \), the maximum value of \( \mathbf{a}_k \) is the unit eigenvector corresponding to the maximum eigenvalue of \( \mathbf{R}_k \) that gives the maximum of (A.28).

For the concentration parameter, we consider the terms in the \( Q \)-function that contain \( \gamma \):

\[
\sum_{k=1}^{K} \sum_{\Omega} \gamma_k (\tau, f) \left( - \log(2\pi^M \mathbb{M}(1, M, \gamma_k)) + \kappa_k |\mathbf{a}_k^H \mathbf{y}(\tau, f)|^2 \right)
\] (A.30)

We differentiate this with respect to \( \gamma_k \) and set to 0 to give the following relation

\[
\frac{\partial \mathbb{M}'(1, M, \gamma_k)}{\partial \gamma_k} = \mathbf{a}_k^H \mathbf{R}_k \mathbf{a}_k.
\] (A.31)

where \( \mathbb{M}'(1, M, \gamma_k) \) denotes the partial derivative \( \frac{\partial}{\partial \gamma_k} \mathbb{M}(1, M, \gamma_k) \). By substituting the optimal value for \( \mathbf{a}_k \) in we have

\[
\frac{\mathbb{M}'(1, M, \gamma_k)}{\mathbb{M}(1, M, \gamma_k)} = \frac{\mathbf{a}_k^H \mathbf{Y}_k \mathbf{a}_k}{\sum_{\Omega} \gamma_k (\tau, f)} = r_k,
\] (A.32)

where \( r_k \) is the principal eigenvalue of \( \mathbf{R}_k \). This equation has no closed-form solution, however, a closed-form approximation can be calculated following the method by Sra and Karp [123]. We approximate \( \gamma_k \) by using a bound on the solution as

\[
\gamma_k \approx \frac{M r_k - 1}{2 r_k (1 - r_k)} \left( 1 + \sqrt{1 + \frac{4(M+1)r_k(1-r_k)}{M-1}} \right).
\] (A.33)

For the mixture weight update we recall the constraint \( \sum_{k=1}^{K} \alpha_k = 1 \) and consider the terms of \( Q(\Theta, \Theta') \) that concern \( \alpha_k \). We introduce the Lagrangian multiplier as
\[ \mathcal{L}(\alpha_k, \lambda) = \sum_{k=1}^{K} \sum_{\Omega} \gamma_{krf} \log \alpha_k + \lambda \left( \sum_{k=1}^{K} \alpha_k - 1 \right). \]  
(A.34)

We differentiate (A.34) with respect to \( \alpha_k \) and set to 0 to obtain

\[ \alpha_k = -\frac{1}{\lambda} \left( \sum_{\Omega} \gamma_{krf} \right). \]  
(A.35)

We substitute this into the constraint on \( \alpha_k \) (8.5) to yield:

\[ \lambda = -|Y| \]  
(A.36)

where \(| \cdot |\) denotes the cardinality. We then obtain the update equation for \( \alpha_k \) as

\[ \alpha_k = \frac{\sum_{k=1}^{K} \gamma_k(\tau, f)}{|Y|}. \]  
(A.37)
Appendix B

Determination of optimal number of mixture model components

B.1 Determination of optimal number of Gaussian components

In Chapter 6 we presented the BSS results for $K = N$ when the GMM was used for mask estimation. However, due to the nature of the GMM where the optimal $K$ value is not necessarily $N$, we also ran evaluations for different numbers of mixture components, ranging from $K = N$ to $K = N + 8$. We found that the optimal number of clusters was highly dependent on the environment as well as the performance ratio in question. Figure B.1 shows the BSS results in simulated conditions for different $K$ values, where the number of sources $N$ is four, and the number of microphones $M$ is three.

It is evident that there is no optimal value of $K$ across the four performance ratios, for example, the best SDR values are achieved when $K = 10$, however, the best SIR values are achieved when $K$ is either 8 or 10, depending on the reverberation time. Furthermore, for the ISR favorable results are generally attained when $K = N$. This trend is repeated in the recorded data results in Figures B.2 and B.3.

The selection of clusters for the SiSEC 2008 and 2010 benchmark data is discussed in Section 6.4 and the details on the configuration of the sources and microphones is outlined...
in Section 5.3. For these data sets, \( K \) was again dependent on the condition. This suggests that the GMM is not very robust in real-world environments, and that the selection of an appropriate number of clusters should be done with caution.

As with the case of simulated and the recorded office data, the optimal number of Gaussian mixtures for the SiSEC data varied for each condition, and we could not find an overall optimal number of clusters. We evaluated the GMM on each mixture of the SiSEC databases we included in this thesis, and the results are shown in Tables B.1–B.4. We denoted the maximum achieved average ratios in the boldface in the tables. It is evident that the optimal number of clusters varies according to the conditions and performance ratio, even varying within the data set.
Figure B.1: Source separation results for $R = 50 \text{ cm}$ and a microphone spacing of 4 cm with simulated data when different $K$ values are used with the GMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Figure B.2: Source separation results for $R = 50$ and $120$ cm and a microphone spacing of $4$ cm with real data collected in an office environment when different $K$ values are used with the GMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and $20$ trials. The error bars show the standard deviation.
Figure B.3: Source separation results for $R = 50$ and $120$ cm and a microphone spacing of $8$ cm with real data collected in an office environment when different $K$ values are used with the GMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Table B.1: Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output SDR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SDR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM$N$</td>
</tr>
<tr>
<td>dev1_male3_liverec_130ms_5cm</td>
<td>-0.54</td>
</tr>
<tr>
<td>dev1_female3_liverec_130ms_5cm</td>
<td>0.22</td>
</tr>
<tr>
<td>dev1_male4_liverec_130ms_5cm</td>
<td>0.55</td>
</tr>
<tr>
<td>dev1_female4_liverec_130ms_5cm</td>
<td>-0.55</td>
</tr>
<tr>
<td>dev1_male3_liverec_250ms_5cm</td>
<td>-0.17</td>
</tr>
<tr>
<td>dev1_female3_liverec_250ms_5cm</td>
<td>0.74</td>
</tr>
<tr>
<td>dev1_male4_liverec_250ms_5cm</td>
<td>-0.35</td>
</tr>
<tr>
<td>dev1_female4_liverec_250ms_5cm</td>
<td>-0.19</td>
</tr>
<tr>
<td><strong>Average (130 ms reverberation)</strong></td>
<td>-0.08</td>
</tr>
<tr>
<td><strong>Average (250 ms reverberation)</strong></td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th>SDR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Ce_A</td>
<td>0.15</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Ce_B</td>
<td>-2.47</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_A</td>
<td>-1.22</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_B</td>
<td>0.12</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_A</td>
<td>-0.17</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_B</td>
<td>-1.23</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_A</td>
<td>-0.25</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_B</td>
<td>-0.05</td>
</tr>
<tr>
<td><strong>Average for Cafeteria</strong></td>
<td>-0.85</td>
</tr>
<tr>
<td><strong>Average for Square</strong></td>
<td>-0.43</td>
</tr>
</tbody>
</table>
Table B.2: Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output ISR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>ISR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMMN</td>
</tr>
<tr>
<td>dev1_male3_liverec_130ms_5cm</td>
<td>6.21</td>
</tr>
<tr>
<td>dev1_female3_liverec_130ms_5cm</td>
<td>5.95</td>
</tr>
<tr>
<td>dev1_male4_liverec_130ms_5cm</td>
<td>5.56</td>
</tr>
<tr>
<td>dev1_female4_liverec_130ms_5cm</td>
<td>4.19</td>
</tr>
<tr>
<td>dev1_male3_liverec_250ms_5cm</td>
<td>6.48</td>
</tr>
<tr>
<td>dev1_female3_liverec_250ms_5cm</td>
<td>4.81</td>
</tr>
<tr>
<td>dev1_male4_liverec_250ms_5cm</td>
<td>3.83</td>
</tr>
<tr>
<td>dev1_female4_liverec_250ms_5cm</td>
<td>3.80</td>
</tr>
<tr>
<td>Average (130 ms reverberation)</td>
<td>5.48</td>
</tr>
<tr>
<td>Average (250 ms reverberation)</td>
<td>4.73</td>
</tr>
<tr>
<td>SiSEC 2010</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Ce_A</td>
<td>4.24</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Ce_B</td>
<td>3.37</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_A</td>
<td>3.79</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_B</td>
<td>4.33</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_A</td>
<td>6.16</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_B</td>
<td>5.86</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_A</td>
<td>6.49</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_B</td>
<td>6.03</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td>3.93</td>
</tr>
<tr>
<td>Average for Square</td>
<td>6.14</td>
</tr>
</tbody>
</table>
## APPENDIX B. DETERMINATION OF OPTIMAL NUMBER OF MIXTURE MODEL COMPONENTS

Table B.3: Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output SIR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SIR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMMₙ</td>
</tr>
<tr>
<td>dev1_male3_liverec_130ms_5cm</td>
<td>-1.48</td>
</tr>
<tr>
<td>dev1_female3_liverec_130ms_5cm</td>
<td>-0.77</td>
</tr>
<tr>
<td>dev1_male4_liverec_130ms_5cm</td>
<td>-1.42</td>
</tr>
<tr>
<td>dev1_female4_liverec_130ms_5cm</td>
<td><strong>2.05</strong></td>
</tr>
<tr>
<td>dev1_male3_liverec_250ms_5cm</td>
<td><strong>2.85</strong></td>
</tr>
<tr>
<td>dev1_female3_liverec_250ms_5cm</td>
<td><strong>0.85</strong></td>
</tr>
<tr>
<td>dev1_male4_liverec_250ms_5cm</td>
<td>-2.03</td>
</tr>
<tr>
<td>dev1_female4_liverec_250ms_5cm</td>
<td>-1.99</td>
</tr>
<tr>
<td><strong>Average (130 ms reverberation)</strong></td>
<td>-0.41</td>
</tr>
<tr>
<td><strong>Average (250 ms reverberation)</strong></td>
<td><strong>-0.08</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dev2ch_3src_CaCeA</td>
<td>-1.02</td>
</tr>
<tr>
<td>dev2ch_3src_CaCeB</td>
<td>0.89</td>
</tr>
<tr>
<td>dev2ch_3src_CaCoA</td>
<td>1.65</td>
</tr>
<tr>
<td>dev2ch_3src_CaCoB</td>
<td>0.13</td>
</tr>
<tr>
<td>dev2ch_3src_SqCeA</td>
<td>2.75</td>
</tr>
<tr>
<td>dev2ch_3src_SqCeB</td>
<td><strong>4.45</strong></td>
</tr>
<tr>
<td>dev2ch_3src_SqCoA</td>
<td>4.15</td>
</tr>
<tr>
<td>dev2ch_3src_SqCoB</td>
<td>2.57</td>
</tr>
<tr>
<td><strong>Average for Cafeteria</strong></td>
<td>0.41</td>
</tr>
<tr>
<td><strong>Average for Square</strong></td>
<td>3.48</td>
</tr>
</tbody>
</table>
Table B.4: Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components in the GMM algorithm with respect to the output SAR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM(_N)</td>
</tr>
<tr>
<td>dev1_male3_liver3_130ms_5cm</td>
<td>6.68</td>
</tr>
<tr>
<td>dev1_female3_liver3_130ms_5cm</td>
<td>7.40</td>
</tr>
<tr>
<td>dev1_male4_liver3_130ms_5cm</td>
<td>5.05</td>
</tr>
<tr>
<td>dev1_female4_liver3_130ms_5cm</td>
<td>6.34</td>
</tr>
<tr>
<td>dev1_male3_liver3_250ms_5cm</td>
<td>7.74</td>
</tr>
<tr>
<td>dev1_female3_liver3_250ms_5cm</td>
<td>8.65</td>
</tr>
<tr>
<td>dev1_male4_liver3_250ms_5cm</td>
<td>4.85</td>
</tr>
<tr>
<td>dev1_female4_liver3_250ms_5cm</td>
<td>5.95</td>
</tr>
<tr>
<td>Average (130 ms reverberation)</td>
<td>6.37</td>
</tr>
<tr>
<td>Average (250 ms reverberation)</td>
<td>6.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dev2_ch3src_Ca_Ce_A</td>
<td>3.34</td>
</tr>
<tr>
<td>dev2_ch3src_Ca_Ce_B</td>
<td>3.83</td>
</tr>
<tr>
<td>dev2_ch3src_Ca_Co_A</td>
<td>4.16</td>
</tr>
<tr>
<td>dev2_ch3src_Ca_Co_B</td>
<td>2.95</td>
</tr>
<tr>
<td>dev2_ch3src_Sq_Ce_A</td>
<td>-1.88</td>
</tr>
<tr>
<td>dev2_ch3src_Sq_Ce_B</td>
<td>-0.37</td>
</tr>
<tr>
<td>dev2_ch3src_Sq_Co_A</td>
<td>2.06</td>
</tr>
<tr>
<td>dev2_ch3src_Sq_Co_B</td>
<td>-0.01</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td><strong>3.57</strong></td>
</tr>
<tr>
<td>Average for Square</td>
<td><strong>-0.05</strong></td>
</tr>
</tbody>
</table>
B.2 Determination of optimal number of Watson components

We also evaluated over different $K$ values as just presented in Section B.1 for the WMM. For the simulated conditions shown in Figure B.4, we noticed much less variation in the optimal cluster number than for the GMM. For example, for the ratios of SDR, ISR and SAR, in general $K = N$ was the best value over all reverberation times. We did notice variation in the SIR, with a higher value of $K$ generally resulting in a higher SIR value. The results were identical for the recorded conditions in Figures B.5 and B.6, where we had stable performance for the SDR, ISR and SAR but varied for the SIR.

Despite this discrepancy, in general the difference between the minimum and maximum achieved SIR values were relatively small (i.e. less than 2 dB), so we can safely conclude that for the WMM, $K = N$ results in optimal performance for simulated and recorded office environments.

For the SiSEC results in Tables B.5–B.8, we noticed much more stability when compared against the GMM in Tables B.1–B.4. For the majority of the cases, $K = N$ resulted in the optimal performance.
Figure B.4: Source separation results for $R = 50$ cm and a microphone spacing of 4 cm with simulated data when different $K$ values are used with the WMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
APPENDIX B. DETERMINATION OF OPTIMAL NUMBER OF MIXTURE MODEL COMPONENTS

Figure B.5: Source separation results for \( R = 50 \) and 120 cm and a microphone spacing of 4 cm with real data collected in an office environment when different \( K \) values are used with the WMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
Figure B.6: Source separation results for $R = 50$ and 120 cm and a microphone spacing of 8 cm with real data collected in an office environment when different $K$ values are used with the WMM for mask estimation. Results given with respect to the absolute (a) SDR, (b) ISR, (c) SIR and (d) SAR measured at the output. Results are averaged over all sources and 20 trials. The error bars show the standard deviation.
### SiSEC 2008

<table>
<thead>
<tr>
<th>Sample Description</th>
<th>SDR (dB)</th>
<th>WMM_N</th>
<th>WMM_{N+2}</th>
<th>WMM_{N+4}</th>
<th>WMM_{N+6}</th>
<th>WMM_{N+8}</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev1_male3_liverec_130ms_5cm</td>
<td>2.82</td>
<td>2.73</td>
<td>2.28</td>
<td>2.17</td>
<td>1.93</td>
<td></td>
</tr>
<tr>
<td>dev1_female3_liverec_130ms_5cm</td>
<td>0.40</td>
<td>2.29</td>
<td>0.93</td>
<td>1.55</td>
<td>1.53</td>
<td></td>
</tr>
<tr>
<td>dev1_male4_liverec_130ms_5cm</td>
<td>1.29</td>
<td>1.83</td>
<td>1.52</td>
<td>1.80</td>
<td>1.73</td>
<td></td>
</tr>
<tr>
<td>dev1_female4_liverec_130ms_5cm</td>
<td>0.81</td>
<td>0.95</td>
<td>1.00</td>
<td>0.96</td>
<td><strong>1.22</strong></td>
<td></td>
</tr>
<tr>
<td>dev1_male3_liverec_250ms_5cm</td>
<td>2.96</td>
<td>2.90</td>
<td><strong>2.97</strong></td>
<td>2.84</td>
<td>1.53</td>
<td></td>
</tr>
<tr>
<td>dev1_female3_liverec_250ms_5cm</td>
<td>1.38</td>
<td>1.99</td>
<td><strong>2.72</strong></td>
<td>1.17</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td>dev1_male4_liverec_250ms_5cm</td>
<td>0.94</td>
<td><strong>2.34</strong></td>
<td>2.08</td>
<td>2.24</td>
<td>2.17</td>
<td></td>
</tr>
<tr>
<td>dev1_female4_liverec_250ms_5cm</td>
<td>0.88</td>
<td><strong>2.00</strong></td>
<td>1.04</td>
<td>1.22</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td><strong>Average (130 ms reverberation)</strong></td>
<td></td>
<td><strong>1.95</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average (250 ms reverberation)</strong></td>
<td></td>
<td><strong>2.31</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### SiSEC 2010

<table>
<thead>
<tr>
<th>Sample Description</th>
<th>SDR (dB)</th>
<th>WMM_N</th>
<th>WMM_{N+2}</th>
<th>WMM_{N+4}</th>
<th>WMM_{N+6}</th>
<th>WMM_{N+8}</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev2ch_3src_Ca_Ce_A</td>
<td><strong>2.54</strong></td>
<td>2.15</td>
<td>0.55</td>
<td>0.21</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Ce_B</td>
<td><strong>3.69</strong></td>
<td>0.76</td>
<td>1.39</td>
<td>1.07</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_A</td>
<td>2.00</td>
<td>1.38</td>
<td><strong>2.72</strong></td>
<td>1.33</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_B</td>
<td><strong>1.61</strong></td>
<td>0.69</td>
<td>0.16</td>
<td>0.07</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_A</td>
<td>0.70</td>
<td>0.68</td>
<td><strong>0.97</strong></td>
<td>0.72</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_B</td>
<td>0.68</td>
<td>1.63</td>
<td>1.71</td>
<td>1.54</td>
<td><strong>1.79</strong></td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_A</td>
<td><strong>3.69</strong></td>
<td>3.19</td>
<td>3.54</td>
<td>3.48</td>
<td>3.43</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_B</td>
<td><strong>1.91</strong></td>
<td>1.60</td>
<td>1.75</td>
<td>1.67</td>
<td>1.56</td>
<td></td>
</tr>
<tr>
<td><strong>Average for Cafeteria</strong></td>
<td></td>
<td><strong>2.46</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average for Square</strong></td>
<td></td>
<td><strong>1.74</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table B.6: Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components $K$ in the WMM algorithm with respect to the output ISR. SISEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>ISR (dB)</th>
<th>WMM$_N$</th>
<th>WMM$_{N+2}$</th>
<th>WMM$_{N+4}$</th>
<th>WMM$_{N+6}$</th>
<th>WMM$_{N+8}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev1_male3_liverec_130ms_5cm</td>
<td>8.16</td>
<td>7.92</td>
<td>4.99</td>
<td>4.62</td>
<td>3.25</td>
<td></td>
</tr>
<tr>
<td>dev1_female3_liverec_130ms_5cm</td>
<td>5.44</td>
<td>5.31</td>
<td>3.03</td>
<td>4.54</td>
<td>4.40</td>
<td></td>
</tr>
<tr>
<td>dev1_male4_liverec_130ms_5cm</td>
<td>5.78</td>
<td>5.61</td>
<td>4.42</td>
<td>5.57</td>
<td>5.02</td>
<td></td>
</tr>
<tr>
<td>dev1_female4_liverec_130ms_5cm</td>
<td>4.91</td>
<td>3.95</td>
<td>2.97</td>
<td>2.03</td>
<td>4.27</td>
<td></td>
</tr>
<tr>
<td>dev1_male3_liverec_250ms_5cm</td>
<td>7.95</td>
<td>7.68</td>
<td>5.79</td>
<td>5.34</td>
<td>3.98</td>
<td></td>
</tr>
<tr>
<td>dev1_female3_liverec_250ms_5cm</td>
<td>6.13</td>
<td>4.83</td>
<td>5.10</td>
<td>2.39</td>
<td>3.75</td>
<td></td>
</tr>
<tr>
<td>dev1_male4_liverec_250ms_5cm</td>
<td>5.21</td>
<td>6.10</td>
<td>4.99</td>
<td>5.65</td>
<td>5.07</td>
<td></td>
</tr>
<tr>
<td>dev1_female4_liverec_250ms_5cm</td>
<td>4.86</td>
<td>5.30</td>
<td>3.39</td>
<td>3.27</td>
<td>3.58</td>
<td></td>
</tr>
<tr>
<td><strong>Average (130 ms reverberation)</strong></td>
<td><strong>6.07</strong></td>
<td>5.70</td>
<td>3.85</td>
<td>4.19</td>
<td>4.24</td>
<td></td>
</tr>
<tr>
<td><strong>Average (250 ms reverberation)</strong></td>
<td><strong>6.04</strong></td>
<td>5.98</td>
<td>4.81</td>
<td>4.17</td>
<td>4.09</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dev2ch_3src_Ca_Ce_A</td>
<td>9.42</td>
<td>3.82</td>
<td>3.82</td>
<td>0.54</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Ce_B</td>
<td>10.91</td>
<td>1.72</td>
<td>2.20</td>
<td>1.80</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_A</td>
<td>7.37</td>
<td>2.37</td>
<td>5.55</td>
<td>1.99</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_B</td>
<td>7.47</td>
<td>1.71</td>
<td>0.41</td>
<td>0.26</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_A</td>
<td>12.31</td>
<td>2.89</td>
<td>2.80</td>
<td>1.85</td>
<td>1.76</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_B</td>
<td>12.97</td>
<td>5.01</td>
<td>4.70</td>
<td>4.62</td>
<td>4.25</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_A</td>
<td>15.13</td>
<td>7.09</td>
<td>6.72</td>
<td>6.39</td>
<td>5.88</td>
<td></td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_B</td>
<td>12.40</td>
<td>4.07</td>
<td>3.68</td>
<td>3.25</td>
<td>2.77</td>
<td></td>
</tr>
<tr>
<td><strong>Average for Cafeteria</strong></td>
<td><strong>8.79</strong></td>
<td>2.41</td>
<td>3.00</td>
<td>1.15</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td><strong>Average for Square</strong></td>
<td><strong>13.20</strong></td>
<td>4.76</td>
<td>4.47</td>
<td>4.03</td>
<td>3.66</td>
<td></td>
</tr>
</tbody>
</table>
Table B.7: Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components $K$ in the WMM algorithm with respect to the output SIR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SIR (dB)</th>
<th>WMM$_N$</th>
<th>WMM$_{N+2}$</th>
<th>WMM$_{N+4}$</th>
<th>WMM$_{N+6}$</th>
<th>WMM$_{N+8}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev1_male3_liverec_130ms_5cm</td>
<td>6.84</td>
<td>8.50</td>
<td>8.03</td>
<td>8.24</td>
<td><strong>8.74</strong></td>
<td></td>
</tr>
<tr>
<td>dev1_female3_liverec_130ms_5cm</td>
<td>2.03</td>
<td><strong>6.25</strong></td>
<td>0.40</td>
<td>-0.47</td>
<td>2.32</td>
<td></td>
</tr>
<tr>
<td>dev1_male4_liverec_130ms_5cm</td>
<td>3.46</td>
<td>4.65</td>
<td>4.77</td>
<td>5.55</td>
<td><strong>5.70</strong></td>
<td></td>
</tr>
<tr>
<td>dev1_female4_liverec_130ms_5cm</td>
<td><strong>3.33</strong></td>
<td>0.57</td>
<td>-0.24</td>
<td>1.95</td>
<td>1.90</td>
<td></td>
</tr>
<tr>
<td>dev1_male3_liverec_250ms_5cm</td>
<td>6.47</td>
<td>7.81</td>
<td>8.48</td>
<td><strong>8.55</strong></td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>dev1_female3_liverec_250ms_5cm</td>
<td>1.94</td>
<td>1.47</td>
<td>5.55</td>
<td>2.16</td>
<td><strong>5.83</strong></td>
<td></td>
</tr>
<tr>
<td>dev1_male4_liverec_250ms_5cm</td>
<td>2.09</td>
<td>5.70</td>
<td>5.59</td>
<td>5.92</td>
<td><strong>6.49</strong></td>
<td></td>
</tr>
<tr>
<td>dev1_female4_liverec_250ms_5cm</td>
<td>1.59</td>
<td><strong>3.38</strong></td>
<td>0.36</td>
<td>0.69</td>
<td>1.73</td>
<td></td>
</tr>
<tr>
<td><strong>Average (130 ms reverberation)</strong></td>
<td>3.92</td>
<td><strong>4.99</strong></td>
<td>3.24</td>
<td>3.82</td>
<td>4.67</td>
<td></td>
</tr>
<tr>
<td><strong>Average (250 ms reverberation)</strong></td>
<td>3.02</td>
<td>4.59</td>
<td><strong>4.99</strong></td>
<td>4.33</td>
<td>3.53</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th>SIR (dB)</th>
<th>WMM$_N$</th>
<th>WMM$_{N+2}$</th>
<th>WMM$_{N+4}$</th>
<th>WMM$_{N+6}$</th>
<th>WMM$_{N+8}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev_2ch_3src_Ca_Ce_A</td>
<td><strong>4.51</strong></td>
<td>2.06</td>
<td>-0.51</td>
<td>-0.50</td>
<td>-1.68</td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Ce_B</td>
<td><strong>8.45</strong></td>
<td>-1.00</td>
<td>3.17</td>
<td>1.99</td>
<td>3.40</td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_A</td>
<td><strong>5.10</strong></td>
<td>3.01</td>
<td>2.94</td>
<td>4.81</td>
<td>1.19</td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Ca_Co_B</td>
<td><strong>4.63</strong></td>
<td>-0.31</td>
<td>1.58</td>
<td>0.47</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_A</td>
<td><strong>8.94</strong></td>
<td>2.36</td>
<td>1.56</td>
<td>3.22</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Ce_B</td>
<td><strong>11.86</strong></td>
<td>5.32</td>
<td>5.65</td>
<td>6.22</td>
<td>6.65</td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_A</td>
<td><strong>15.42</strong></td>
<td>5.56</td>
<td>4.44</td>
<td>4.14</td>
<td>4.39</td>
<td></td>
</tr>
<tr>
<td>dev_2ch_3src_Sq_Co_B</td>
<td><strong>11.25</strong></td>
<td>3.86</td>
<td>3.11</td>
<td>3.52</td>
<td>3.21</td>
<td></td>
</tr>
<tr>
<td><strong>Average for Cafeteria</strong></td>
<td><strong>5.67</strong></td>
<td>0.94</td>
<td>1.79</td>
<td>1.69</td>
<td>1.18</td>
<td></td>
</tr>
<tr>
<td><strong>Average for Square</strong></td>
<td><strong>11.87</strong></td>
<td>4.28</td>
<td>3.69</td>
<td>4.27</td>
<td>3.97</td>
<td></td>
</tr>
</tbody>
</table>
Table B.8: Separation results for the SiSEC 2008 and 2010 data sets for different number of mixture components $K$ in the WMM algorithm with respect to the output SAR. SiSEC 2010 results are averages over the A and B recordings. The highest achieved ratio per condition is denoted in the boldface.

<table>
<thead>
<tr>
<th>SiSEC 2008</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WMM$_N$</td>
</tr>
<tr>
<td>dev1_male3_liverec_130ms_5cm</td>
<td>8.07</td>
</tr>
<tr>
<td>dev1_female3_liverec_130ms_5cm</td>
<td>6.64</td>
</tr>
<tr>
<td>dev1_male4_liverec_130ms_5cm</td>
<td>5.84</td>
</tr>
<tr>
<td>dev1_female4_liverec_130ms_5cm</td>
<td>5.55</td>
</tr>
<tr>
<td>dev1_male3_liverec_250ms_5cm</td>
<td>6.67</td>
</tr>
<tr>
<td>dev1_female3_liverec_250ms_5cm</td>
<td>6.39</td>
</tr>
<tr>
<td>dev1_male4_liverec_250ms_5cm</td>
<td>4.49</td>
</tr>
<tr>
<td>dev1_female4_liverec_250ms_5cm</td>
<td>4.50</td>
</tr>
<tr>
<td>Average (130 ms reverberation)</td>
<td>6.53</td>
</tr>
<tr>
<td>Average (250 ms reverberation)</td>
<td>5.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SiSEC 2010</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev2ch_3src_Ca_Ce_A</td>
<td>4.87</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Ce_B</td>
<td>4.89</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_A</td>
<td>3.80</td>
</tr>
<tr>
<td>dev2ch_3src_Ca_Co_B</td>
<td>3.86</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_A</td>
<td>-0.83</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Ce_B</td>
<td>0.44</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_A</td>
<td>3.16</td>
</tr>
<tr>
<td>dev2ch_3src_Sq_Co_B</td>
<td>1.17</td>
</tr>
<tr>
<td>Average for Cafeteria</td>
<td>4.35</td>
</tr>
<tr>
<td>Average for Square</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Appendix C

Calculation of speech activity sequences and WMM update equations for source number estimation

C.1 Calculation of speech activity sequences

We explain the calculation of the speech activity sequences used in the source number estimation algorithm of Chapter 10. We remark that the following explanations and derivations are similar to that of Sawada et al. [18] but included for reference and completeness.

C.1.1 Model

We have the observation vectors $\mathbf{x}(\tau, f) = [x_1(\tau, f), \ldots, x_M(\tau, f)]$ in the STFT domain. As the clustering is performed in a frequency bin-wise manner, we omit the frequency dependence for simplicity to yield $\mathbf{x}(\tau)$. To rid any effects of the source amplitude, we normalize the observation vectors to unit norm as:

$$
\mathbf{x}(\tau) \leftarrow \frac{\mathbf{x}(\tau)}{\|\mathbf{x}(\tau)\|}.
$$

(C.1)
We also pre-whiten [3] the observation vectors for robust execution of the clustering as follows:

\[ x(\tau) \leftarrow Vx(\tau) \]  

(C.2)

where \( V \) is the whitening matrix calculated by \( V = D^{-\frac{1}{2}}E^H \) with an eigenvalue decomposition \( E\{x(\tau)x(\tau)^H\} = ED^HE^H \). The normalization in (C.1) is then applied again.

To model the observation vectors \( x(\tau) \) we follow the line orientation idea in [79,80] and employ a complex Gaussian density function:

\[
p(x(\tau)|a_i, \sigma_i) = \frac{1}{(\pi \sigma_i^2)^{M/2}} \exp \left( -\frac{\|x(\tau) - (a_i^H x(\tau)) \cdot a_i\|^2}{\sigma_i^2} \right) \]  

(C.3)

where \( a_i \) is the centroid of the \( i \)-th component and \( \sigma_i^2 \) the variance. The distance measure in \( \|x(\tau) - (a_i^H x(\tau)) \cdot a_i\| \) represents the minimum distance between the data point \( x(\tau) \) and the subspace spanned by \( a_i \), which indicates how likely the vector in \( x(\tau) \) belongs to the \( i \)-th class [18].

The density function is described by a mixture model as

\[
p(x(\tau)|\Theta) = \sum_{i=1}^{K} \alpha_i p(x(\tau)|a_i, \sigma_i), \]  

(C.4)

where \( \Theta = \{a_1, \sigma_1, \alpha_1, \ldots, a_K, \sigma_K, \alpha_K\}^{K}_{i=1} \) is the parameter set for the mixture model, \( \alpha_i \) the \( i \)-th mixture weight and \( K \) denotes the number of mixture components (clusters). The parameter set is subject to the constraints:

\[
\sum_{i=1}^{K} \alpha_i = 1, \|a_i\| = 1. \]  

(C.5)

The mixture weights \( \alpha_i \) are modeled by a Dirichlet distribution as

\[
p(\alpha_i) = \frac{\Gamma(K \cdot \phi)}{\Gamma(\phi)^K} \prod_{i=1}^{K} \alpha_i^{(\phi-1)}, \]  

(C.6)

where \( \phi \) is a hyperparameter to influence the strength of the Dirichlet distribution.

### C.1.2 EM algorithm

We employ the EM algorithm to estimate the parameters in \( \Theta \) and the final posterior probabilities (speech activity sequences) \( P(C_i|x(\tau)) \), where \( C_i \) denotes the \( i \)-th cluster class.
The E-step computes the posterior probabilities as
\[ P(C_i|x(\tau), \Theta') = \frac{\alpha'_i p(x(\tau)|a'_i, \sigma'_i)}{\sum_{i=1}^{K} \alpha'_i p(x(\tau)|a'_i, \sigma'_i)} \] (C.7)

using the current parameter set estimate \( \Theta' = \{a'_1, \sigma'_1, \alpha'_1, \ldots, a'_K, \sigma'_K, \alpha'_K\}_{i=1}^{K} \).

The parameter set \( \Theta \) is updated in the M-step by maximizing the following auxiliary function of the data likelihood
\[ Q(\Theta, \Theta') = \sum_{i=1}^{K} \sum_{\tau} P(C_i|x(\tau), \Theta') \log \alpha_i p(x(\tau)|a_i, \sigma_i) + \log p(\Theta) \] (C.8)

where \( T \) is the number of time frames and \( p(\Theta) \) is any prior distribution on the parameters \( \Theta \). Due to the Dirichlet prior on the mixture ratios we have
\[ \log p(\Theta) = (\phi - 1) \sum_{i=1}^{K} \log \alpha_i + \text{const.} \] (C.9)

To solve for this, we employ a Lagrange multiplier \( \lambda \) for the update of each parameter. For the centroid \( a_i \), we consider the unit-norm constraint \( \|a_i\|^2 = 1 \) and consider the function
\[ L(a_i, \lambda) = Q(\Theta, \Theta') + \lambda(\|a_i\|^2). \] (C.10)

We derive (C.10) with respect to \( a_i \) and obtain
\[ Ra_i = \frac{\lambda}{\sigma_i^2} a_i, \] (C.11)

where \( R \) is defined as
\[ R = \sum_{\tau} P(C_i|x(\tau), \Theta') \cdot x(\tau)x(\tau)^H. \] (C.12)

It is clear from (C.11) that \( a_i \) will be an eigenvector of \( R \) at stationary points. The update of \( a_i \) is therefore the eigenvector corresponding to the maximum eigenvalue of \( R \).

For the update of the variance \( \sigma_i^2 \) we derive \( Q(\Theta, \Theta') \) with respect to \( \sigma_i^2 \) as
\[ \frac{\partial Q(\Theta, \Theta')}{\partial \sigma_i^2} = \sum_{\tau} P(C_i|x(\tau), \Theta') \left( -\frac{(M - 1)}{\sigma_i^2} + \frac{\|x(\tau) - (a_i^H x(\tau) \cdot a_i)\|^2}{(\sigma_i^2)^2} \right) \] (C.13)
To optimize for $\sigma_i^2$ we set this to 0 and solve as:

$$\sum_{\tau} P(C_i|x(\tau), \Theta')(M - 1) = \sum_{\tau} P(C_i|x(\tau), \Theta') \frac{\|x(\tau) - (a_i^H x(\tau) \cdot a_i)\|^2}{\sigma_i^2}. \quad (C.14)$$

Solving for $\sigma_i^2$, we have the update equation:

$$\sigma_i^2 = \frac{\sum_{\tau} P(C_i|x(\tau), \Theta') \cdot \|x(\tau) - (a_i^H x(\tau) \cdot a_i)\|^2}{(M - 1) \cdot \sum_{\tau} P(C_i|x(\tau), \Theta')} \quad (C.15)$$

For the mixture update $\alpha_i$, we observe the constraint $\sum_{i=1}^{K} \alpha_i = 1$ and introduce the Lagrange multiplier to obtain the following function

$$\mathcal{L}(\alpha_i, \lambda) = Q(\Theta, \Theta') + \lambda \left( \sum_{i=1}^{K} \alpha_i - 1 \right) \quad (C.16)$$

We derive (C.16) with respect to $\alpha_i$ and set to zero and obtain

$$\sum_{\tau} P(C_i|x(\tau), \Theta') + (\phi - 1) + \alpha_i \lambda = 0 \quad (C.17)$$

for $i = 1, \ldots, K$. We sum these up with $i = 1, \ldots, K$ and obtain

$$\lambda = -(T + N \cdot (\phi - 1)) \quad (C.18)$$

Substituting this into (C.17) yields the mixture weight update:

$$\alpha_i = \frac{\sum_{\tau} P(C_i|x(\tau), \Theta') + (\phi - 1)}{T + N \cdot (\phi - 1)} \quad (C.19)$$

Upon convergence of the EM algorithm, the bin-wise clustering results are represented by the final estimates of the posterior probability sequences in (C.7), $P(C_i|x(\tau), \Theta)$.

### C.2 Derivation of WMM update equations

We remark that the derivation is similar to that in Appendix A.4, but we include it for completeness and reference. We aim to cluster the speech activity sequences in $\{v_i\}_{i=1}^{L}$ into $K$ clusters via the WMM. We note that the index $k$ is used to denote the WMM clusters whilst in Appendix C.1 the index $i$ denoted the bin-wise clusters. We repeat the auxiliary function $Q(\Theta, \Theta')$ for completeness:
APPENDIX C. CALCULATION OF SPEECH ACTIVITY SEQUENCES AND WMM UPDATE EQUATIONS FOR SOURCE NUMBER ESTIMATION

\[ Q(\Theta', \Theta) = \sum_{k=1}^{K} \sum_{l=1}^{L} \gamma_{kl} \log \alpha_{kl} p(\mathbf{v}_l | \mathbf{a}_k, \kappa_k) + (\phi - 1) \sum_{k=1}^{K} \log \alpha_k. \]  

(C.20)

We expand this as

\[ Q(\Theta', \Theta) = \sum_{k=1}^{K} \left( \sum_{l=1}^{L} \gamma_{kl} + (\phi - 1) \right) \log \alpha_k - \sum_{k=1}^{K} \sum_{l=1}^{L} \gamma_{kl} 2\pi^{-\frac{T}{2}} \log \mathcal{N} \left( \frac{1}{2}, \frac{T}{2}, \kappa_k \right) \]

\[ + \sum_{k=1}^{K} \sum_{l=1}^{L} \gamma_{kl} \kappa_k \mathbf{a}_k^T \mathbf{R}_k \mathbf{a}_k \]  

(C.21)

where \( \mathbf{R}_k \) is the covariance matrix defined as

\[ \mathbf{R}_k = \frac{\sum_{l=1}^{L} \gamma_{kl} \mathbf{v}_l \mathbf{v}_l^T}{\sum_{l=1}^{L} \gamma_{kl}}. \]  

(C.22)

We employ the Lagrange multiplier method to maximize the auxiliary function subject to the constraints in (10.4). For the update of the mean orientation we recall the constraint \( \| \mathbf{a}_k \|^2 = 1 \) and consider the terms of the Q-function relevant to \( \mathbf{a}_k \):

\[ \mathcal{L}(\mathbf{a}_k, \lambda) = \sum_{k=1}^{K} \left( \sum_{l=1}^{L} \gamma_{kl} \right) \kappa_k \mathbf{a}_k^T \mathbf{R}_k \mathbf{a}_k + \lambda(\| \mathbf{a}_k \|^2 - 1). \]  

(C.23)

We derive this with respect to \( \mathbf{a}_k \) and set to 0 to obtain:

\[ \mathbf{R}_k \mathbf{a}_k = \frac{\lambda \mathbf{a}_k}{\kappa_k}. \]  

(C.24)

As \( \kappa_k > 0 \), the maximizer is the unit eigenvector corresponding to the maximum eigenvalue of \( \mathbf{R}_k \).

For the concentration parameter \( \kappa_k \) we consider the terms in the Q-function that concern \( \kappa_k \), i.e.

\[ - \sum_{k=1}^{K} \left( \sum_{l=1}^{L} \gamma_{kl} \right) \log 2\pi^{-\frac{T}{2}} \mathcal{N} \left( \frac{1}{2}, \frac{T}{2}, \kappa_k \right) + \sum_{k=1}^{K} \left( \sum_{l=1}^{L} \gamma_{kl} \right) \kappa_k \mathbf{a}_k^T \mathbf{R}_k \mathbf{a}_k. \]  

(C.25)

We derive this with respect to \( \kappa_k \) and set to 0 as follows
**APPENDIX C. CALCULATION OF SPEECH ACTIVITY SEQUENCES AND WMM UPDATE EQUATIONS FOR SOURCE NUMBER ESTIMATION**

\[
\frac{\partial M(\frac{1}{2}, \frac{T}{2}, \kappa_k)}{\partial \kappa_k} + \mathbf{a}_k^T \mathbf{R}_k \mathbf{a}_k = \mathbf{0},
\]

\[
\frac{\partial M(\frac{1}{2}, \frac{T}{2}, \kappa_k)}{\partial \kappa_k} = \mathbf{a}_k^T \mathbf{R}_k \mathbf{a}_k = r_k,
\]

where \( M(\cdot, \cdot, x) \) denotes \( \frac{\partial}{\partial x} M(\cdot, \cdot, x) \), and \( r_k \) is the principal eigenvalue of \( \mathbf{R}_k \). The above equation has no closed-form solution, however it can be approximated by [123]

\[
\kappa_k \approx \frac{T r_k - 1}{2 r_k (1 - r_k)} \left( 1 + \sqrt{1 + \frac{4(T + 1) r_k (1 - r_k)}{2(T - 1)}} \right).
\]

For the mixture weight update, we recall the constraint \( \sum_{k=1}^K \alpha_k = 1 \) and consider the terms of the auxiliary function that concern \( \alpha_k \). We introduce the Lagrangian multiplier to obtain

\[
\mathcal{L}(\alpha_k, \lambda) = \sum_{k=1}^K \sum_{l=1}^L \gamma_{kl} \log \alpha_k + \sum_{k=1}^K (\phi - 1) \log \alpha_k + \lambda(\sum_{k=1}^K -1).
\]

We derive this with respect to \( \alpha_k \) and set to 0:

\[
\alpha_k = -\frac{1}{\lambda} \left( \sum_{l=1}^L \gamma_{kl} + (\phi - 1) \right).
\]

Substituting this into the constraint on \( \alpha_k \) yields

\[
\lambda = -(T + K \cdot (\phi - 1)),
\]

and we thus obtain the update for the mixture weight parameter as

\[
\alpha_k = \frac{\sum_{l=1}^L \gamma_{kl} + (\phi - 1)}{T + K \cdot (\phi - 1)}.
\]