Audio Networks for Speech Enhancement and Indexing

Thorsten Kühnapfel

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University of Technology

December 2009
This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made.

____________________________________________________________________

Thorsten Kühnapfel

Date
CONTENTS

Abstract viii
Acknowledgements x
Published work xi
Abbreviations xi
Notations xiii
  0.1 Specific variables and functions xiv
  0.2 Indexing or counting variables xvii
  0.3 Constant xvii

1 Introduction 1
  1.1 Aims and approach 2
    1.1.1 Dynamic calibration of audio and video cues 3
    1.1.2 Distributed microphone network for acoustic surveillance 4
  1.2 Significance and contribution 5
  1.3 Thesis outline 6

2 Background 8
  2.1 Introduction 8
  2.2 Surveillance tasks and network architecture 8
    2.2.1 Audio surveillance 10
    2.2.2 Audio-Video surveillance 13
      2.2.2.1 Sensor calibration 14
      2.2.2.2 Data fusion 14
    2.2.3 Audio-Video systems 15
  2.3 Audio feature processing 17
  2.4 Audio classification 19
    2.4.1 Acoustic event detection 19
    2.4.2 Voice activity detection 21
  2.5 Sound source localisation 23
    2.5.1 Direction of arrival 24
    2.5.2 High-resolution spectral estimation 25
2.5.3 Time delay of arrival .................................................. 26
2.5.4 Energy based localisation ........................................... 26
2.6 Speech enhancement ................................................... 27
  2.6.1 Beamforming .......................................................... 28
  2.6.2 Blind source separation .......................................... 30
  2.6.3 Spectral subtraction ............................................... 31
2.7 Summary ..................................................................... 33

3 Audio-Video calibration .................................................. 35
  3.1 Introduction .............................................................. 35
  3.2 Methodology ............................................................. 36
    3.2.1 Angle of arrival ..................................................... 36
    3.2.2 Design parameter of the microphone array ................. 41
    3.2.3 Acoustic and vision calibration .................................. 43
    3.2.4 Audio-Video surveillance ....................................... 46
  3.3 Experiments ............................................................... 47
    3.3.1 Angle of arrival evaluation .................................... 48
    3.3.2 Calibration accuracy .............................................. 50
    3.3.3 Signal enhancement ............................................... 54
    3.3.4 Indexing audio-video events .................................... 59
  3.4 Conclusion ................................................................ 60

4 Distributed microphone network ....................................... 62
  4.1 Introduction .............................................................. 62
  4.2 Methodology ............................................................. 66
    4.2.1 Generating a “virtual” microphone ......................... 67
    4.2.2 Speech enhancement .............................................. 68
      4.2.2.1 Noise classification ......................................... 68
      4.2.2.2 Voice activity detection ................................... 70
      4.2.2.3 Spectral subtraction ......................................... 72
  4.3 Experiments ............................................................... 74
    4.3.1 “Virtual” microphone - Experimental setup ............... 74
      4.3.1.1 One moving target ........................................ 75
      4.3.1.2 Multiple stationary targets ............................. 78
    4.3.2 Speech enhancement - Experimental setup ............... 80
      4.3.2.1 Noise classification ......................................... 81
      4.3.2.2 Voice activity detection ................................... 83
      4.3.2.3 Noise changes during speech ............................ 86
      4.3.2.4 Qualitative evaluation of the enhanced speech ....... 91

iii
5 Improved speech enhancement

5.1 Introduction ................................................. 95
5.2 Methodology ............................................... 96
  5.2.1 Noise classification ................................... 97
  5.2.2 Voice activity detection ............................... 100
    5.2.2.1 VAD based on signal to noise ratio estimation .. 100
    5.2.2.2 VAD based on noise distance measure ............ 101
    5.2.2.3 Combined VAD .................................... 102
  5.2.3 Speech enhancement ................................. 103
5.3 Experiments ................................................ 104
  5.3.1 Noise classification ................................... 104
  5.3.2 Speech parameters estimation ......................... 109
  5.3.3 Voice activity detection .............................. 111
  5.3.4 Speech enhancement ................................ 111
5.4 Conclusion ................................................ 114

6 Conclusion .................................................. 116
  6.1 Future work .............................................. 117

Bibliography .................................................... 118

A Audio features ............................................... 130
  A.1 Pre-processing .......................................... 130
  A.2 Domain transformations .................................. 131
  A.3 Temporal features ...................................... 132
    A.3.1 Zero crossing rate .................................. 132
    A.3.2 Short time energy ................................... 132
    A.3.3 Silence ratio ....................................... 133
    A.3.4 Auto-correlation .................................... 133
  A.4 Spectral features ....................................... 134
    A.4.1 Energy sub-bands .................................... 134
    A.4.2 Mel filter bank ..................................... 135
    A.4.3 Spatial centroid .................................... 136
    A.4.4 Spectral spread ..................................... 137
    A.4.5 Mel Frequency Cepstral Coefficient ............... 137
LIST OF FIGURES

2.1 Illustration of reverberation effect. ................................. 11
2.2 Probability density function of the particle filter. ................. 15
2.3 Graphical model of audio-video fusion. .......................... 16
2.4 Sound wave propagation of a source in the near and far-field region. 28
2.5 Signal flow of conventional and adaptive beamformer. ............ 30
2.6 Signal flow of blind source separation for noise free environment. 31
2.7 Signal flow for general spectral subtraction algorithm. ............. 32

3.1 AOA estimation for a linear and uniformly distributed microphone array. 37
3.2 Smoothing of the steered power response for AOA estimation. ........ 41
3.3 Beam patterns for different microphone spacings. .................. 43
3.4 Beam patterns for different numbers of microphones K. ............. 44
3.5 Likelihood function ρ for face detection. ........................... 46
3.6 Source location for AOA evaluation. ................................ 48
3.7 AOA estimation for a single sound source with speech as the signal. 49
3.8 Error of the estimated AOA. ......................................... 49
3.9 Mean standard deviation of AOA estimation for different sound sources. 50
3.10 Illustration of the image partitioning into bins, object detection and face detection. ................................................. 52
3.11 Calibration functions for all walking patterns and Speech 1 as source signal. 52
3.12 Standard deviation of all three walking patterns with Speech 1 as source signal. .......................................................... 53
3.13 Experimental setup for signal enhancement evaluation. ............ 53
3.14 Normalised signal power of four different white noise sources for a scan between ±90°. .................................................. 55
3.15 Signal power of the unwanted source. .............................. 57
3.16 Theoretic beam pattern of seven linear aligned microphones with a uniformly distance of 4 cm. ............................................ 58
3.17 Illustration of two visually detected events. .......................... 59
3.18 Illustration of the second test sequence of a moving target. ........ 60

4.1 Inverse square law. ...................................................... 63
4.2 Sound intensity based on inverse square law. ......................... 64
4.3 Visualisation of the principle of a “virtual” microphone. .......... 65
4.4 System flow of the voice activity detection. .......................... 70
4.5 Layout of the microphone distribution and walking path patterns. . . . . . 75
4.6 Gaussian distance function and ground truth for one active target. . . . . . 76
4.7 Acoustics and distance importance weighting of walking pattern 1. . . . . . 77
4.8 Acoustics and distance importance weighting of walking pattern 2. . . . . . 77
4.9 Importance weighting $w$ of each microphone for multiple active target. . 79
4.10 Experiment setup for evaluation of the proposed system under realistic con-
ditions. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
4.11 Captured audio sequences of a café during lunch time and of a scooter. . . 81
4.12 Introduced noise pattern for classification. . . . . . . . . . . . . . . . . . . 82
4.13 Voice activity detection results for various SNR situations. . . . . . . . . . 84
4.14 Voice activity detection result of qualitative speech evaluation. . . . . . . 86
4.15 Speech masked by synthetic noise. . . . . . . . . . . . . . . . . . . . . . . . 87
4.16 Comparison of speech enhancement results for increasing numbers of mod-
elled noise sources. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88
4.17 Input signal of the speech enhancement approach for a male speaker. . . 89
4.18 Input signal of the speech enhancement approach for a female speaker. . 90
4.19 Estimated speech signal for a male speaker. . . . . . . . . . . . . . . . . . 90
4.20 Estimated speech signal for a female speaker. . . . . . . . . . . . . . . . . 91
4.21 Audio streams for evaluation of the enhanced speech signals. . . . . . . . 92

5.1 System flow of the speech enhancement for microphone 1. . . . . . . . . . . 97
5.2 Test signals that are used for noise classification. . . . . . . . . . . . . . . . 105
5.3 Eigenvalue of eigenvectors. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
5.4 Mahalanobis distance of noise classification. . . . . . . . . . . . . . . . . . . 107
5.5 Power spectral density of noise types. . . . . . . . . . . . . . . . . . . . . . 108
5.6 Mahalanobis distance of noise classification with one unknown noise source. 108
5.7 ROC graphs for parameter definition of $\nu_r$ and $\nu_P$. . . . . . . . . . . . 110
5.8 Test sequence for voice activity detection. . . . . . . . . . . . . . . . . . . . 112
5.9 Speech enhancement result. . . . . . . . . . . . . . . . . . . . . . . . . . . . 113

A.1 Mel scale filter bank. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136
A.2 Signal flow of MFCC extraction. . . . . . . . . . . . . . . . . . . . . . . . . 137
List of Tables

3.1 SNR (dB) for different signals and distances. ........................................ 48
3.2 Average estimation error of sound source location. .............................. 50
3.3 Average SNR for calibrating the sensor setup. ...................................... 51
3.4 Average error between ground truth AOA and the corresponding calibration value. ................................................................. 53
3.5 Average standard deviation for all walking patterns and source signals. .... 53
3.6 Centre frequency of bandlimited white noise sources for all test sequences. 55
3.7 Error between the angle that maximise the signal power and the angle that is computed based on the image location of the source. ......................... 56
3.8 Difference in signal power of the enhanced signal when steered based on the image location compared to the direction that maximises the signal. .... 56
3.9 Signal power reduction of the unwanted source in comparison to the spatial separation in angle between source locations. ................................. 58
3.10 Mean opinion score for the first test scenario between the original and the beam steered audio. ................................................................. 60
4.1 Noise classification results of the six test-sequences. .............................. 82
4.2 Noise classification result of the second data set. .................................. 83
4.3 Average false positive rate for voice activity detection. .......................... 85
4.4 Voice activity detection result for Speaker 1 and 2 in terms of true positive and false positive classification rate. ............................................ 86
4.5 Average error between the estimated speech signal to the original speech signal for synthetic noise. ......................................................... 89
4.6 Noise levels during speech sequence. .................................................... 91
4.7 Mean opinion score for the original and speech enhanced signal. .......... 92
5.1 Noise classification result for test sequences (a) to (e). .......................... 106
5.2 Speech detection result for the proposed VAD approach. ........................ 110
5.3 Comparison of voice activity detection methods. ................................... 111
5.4 Mean opinion score for ambient noise reduction. .................................. 114
5.5 Mean opinion score for speech enhanced signal. ................................... 114
A.1 Centre frequency and bandwidth of 1 and 1/3 octave spectrum. ............... 135
Audio Networks for Speech Enhancement and Indexing

by

Thorsten Kühnapfel

Submitted to the Department of Computing
in December, 2009 in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

For humans, hearing is the second most important sense, after sight. Therefore, acoustic
information greatly contributes to observing and analysing an area of interest. For this
reason combining audio and video cues for surveillance enhances scene understanding and
the observed events. However, when combining different sensors their measurements need
to be correlated, which is done by either knowing the exact relative sensor alignment
or learning a mapping function. Most deployed systems assume a known relative sensor
alignment, making them susceptible to sensor drifts. Additionally, audio recordings are
generally a mixture of several source signals and therefore need to be processed to extract
a desired sound source, such as speech of a target person.
In this thesis a generic framework is described that captures, indexes and extracts surveil-
lance events from coordinated audio and video cues. It presents a dynamic joint-sensor
calibration approach that uses audio-visual sensor measurements to dynamically and in-
crementally learn the calibration function, making the sensor calibration resilient to inde-
pendent drifts in the sensor suite. Experiments demonstrate the use of such a framework
for enhancing surveillance.
Furthermore, a speech enhancement approach is presented based on a distributed network
of microphones, increasing the effectiveness for acoustic surveillance of large areas. This
approach is able to detect and enhance speech in the presence of rapidly changing envi-
ronmental noise. Spectral subtraction, a single channel speech enhancement approach, is
modified to adapt quickly to rapid noise changes of two common noise sources by incorpo-
rating multiple noise models. The result of the cross correlation based noise classification
approach is also utilised to improve the voice activity detection by minimising false detec-
tion based on rapid noise changes. Experiments with real world noise consisting of scooter
and café noise have proven the advantage of multiple noise models especially when the
noise changes during speech.
The modified spectral subtraction approach is then extended to real world scenarios by
introducing more and highly non-stationary noise types. Thus, the focus is directed to
implement a more sophisticated noise classification approach by extracting a variety of
acoustic features and applying PCA transformation to compute the Mahalanobis distance
to each noise class. This distance measurement is also included in the voice activity
detection algorithm to reduce false detection for highly non-stationary noise types. How-
ever, using spectral subtraction in non-stationary noise environments, such as street noise,
reduces the performance of the speech enhancement. For that reason the speech enhance-
ment approach is further improved by using the sound information of the entire network
to update the noise model of the detected noise type during speech. This adjustment
considerably improved the speech enhancement performance in non-stationary noise en-
vironments. Experiments conducted under diverse real world conditions including rapid
noise changes and non-stationary noise sources demonstrate the effectiveness of the pre-
sented method.
This dissertation would not have been possible without the guidance, contributions and support of many people throughout every stage of this thesis and personal setbacks.

First, I would like to thank my supervisors, A/Prof. Tele Tan and Prof. Svetla Venkatesh, for giving me the opportunity to do a PhD and for providing expert guidance and motivation throughout the course. They taught me the difference between engineering a solution for a problem and researching the problem, whenever I wandered off and tried to do it the “Thorsten way”, resulting in the successful achievement of this thesis.

I would also thank the most important people in my life, my family who supported me in the decision to leave Germany and to live in Australia, a place where I did not know anyone and could barely speak the language at first. I promise you that I will come back home more regularly, now that I have finished my degree.

While on the subject of Germany, I would like to thank my former supervisor at the Fachhochschule Dortmund, Prof. Burkhard Igel, for introducing me to the idea of going to Australia and to do a postgraduate degree at Curtin University. I am very impressed by his continued interest and involvement in my development and his many flights to Australia.

Finally, I would like to thank all my friends here in Australia who welcomed me as they did and all their efforts into teaching me Australian English. That said, by writing these acknowledgements I can finally give all of you the answer to the question you asked me most over the last months, “Are you finished yet?”. I would like to especially thank the Costello family in that they “adopted” me into their family, making me feel like at home and for taking care of me after my surgery. I would not like Perth as I do, if I had not met Kellie Costello. I enjoyed every Saturday drinking coffee in Subiaco and I cannot thank her enough for proof reading this thesis, even though she did not have any technical background in this material.
The original work presented in Chapters 3, 4 and 5 of this thesis has previously been published in refereed conference papers.

The chapter dealing with speech enhancement for a linear microphone array in combination with a CCTV camera (Chapter 3):


The chapter dealing with recording audio of a freely moving target and enhancing speech based on a distributed microphone network (Chapter 4):


The chapter dealing with improved speech enhancement for non-stationary ambient noise types using a distributed microphone network (Chapter 5):

**ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMR</td>
<td>Adaptive multi-rate</td>
</tr>
<tr>
<td>AOA</td>
<td>Angle of arrival</td>
</tr>
<tr>
<td>BSS</td>
<td>Blind source separation</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier transformation</td>
</tr>
<tr>
<td>DOA</td>
<td>Direction of arrival</td>
</tr>
<tr>
<td>DSB</td>
<td>Delay and sum beamformer</td>
</tr>
<tr>
<td>ETSI</td>
<td>European Telecommunication Standards Institute</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier transformation</td>
</tr>
<tr>
<td>FP</td>
<td>False positive</td>
</tr>
<tr>
<td>GCC</td>
<td>Generalised cross correlation</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>GSC</td>
<td>Generalised sidelobe canceller</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov model</td>
</tr>
<tr>
<td>HOS</td>
<td>Higher order statistics</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent component analysis</td>
</tr>
<tr>
<td>IG</td>
<td>Information gain</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunication Union</td>
</tr>
<tr>
<td>LCMV</td>
<td>Linearly constraint minimum variance</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear discriminant analysis</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean opinion score</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>NB</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle component analysis</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability density function</td>
</tr>
<tr>
<td>PHAT</td>
<td>Phase transform</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial basis function</td>
</tr>
<tr>
<td>RMS</td>
<td>Root mean square</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
<tr>
<td>SCOT</td>
<td>Smoothed coherence transform</td>
</tr>
<tr>
<td>STFT</td>
<td>Short time Fourier transformation</td>
</tr>
<tr>
<td>SIL</td>
<td>Sound intensity level</td>
</tr>
<tr>
<td>SOS</td>
<td>Second order statistics</td>
</tr>
<tr>
<td>SR</td>
<td>Silence ratio</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>SRP</td>
<td>Steered response power</td>
</tr>
<tr>
<td>SSL</td>
<td>Sound source localisation</td>
</tr>
<tr>
<td>STE</td>
<td>Short time energy</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular value decomposition</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>TDOA</td>
<td>Time delay of arrival</td>
</tr>
<tr>
<td>TP</td>
<td>True positive</td>
</tr>
<tr>
<td>VAD</td>
<td>Voice activity detection</td>
</tr>
<tr>
<td>ZCR</td>
<td>Zero crossing rate</td>
</tr>
</tbody>
</table>
0.1 Specific variables and functions

\( A \)  
Audio feature set

\( A^n \)  
Audio features of noise type \( n \)

\( A' \)  
Reduced audio feature set

\( A'^n \)  
Reduced audio features of noise type \( n \)

\( \overline{B} \)  
Average background image

\( C \)  
Result of comparison with threshold

\( C_k^P \)  
VAD result of signal power based detection at microphone \( k \)

\( C_k^R \)  
VAD result of noise classification based detection at microphone \( k \)

\( C_k^N \)  
VAD result of Mahalanobis distance measurement based detection at microphone \( k \)

\( C \)  
Current image frame

\( D \)  
Feature set based on Mel scale filter bank

\( D_k^n \)  
Feature set of noise type \( n \) at microphone \( k \)

\( \overline{D} \)  
Noise model of average Mel scale features

\( \overline{D}_k^n \)  
Noise model of noise type \( n \) at microphone \( k \)

\( d \)  
Spatial distance

\( d_{kl} \)  
Spatial distance between microphone \( k \) and \( l \)

\( f \)  
Frequency

\( f_s \)  
Sampling frequency

\( G \)  
Gain function of spectral subtraction algorithm in the frequency domain

\( G_k \)  
Gain function of microphone \( k \)

\( g \)  
Gain function of spectral subtraction algorithm in the time domain

\( g_k \)  
Gain function of microphone \( k \)

\( H \)  
Filter function in frequency domain

\( H_k \)  
Filter function of microphone \( k \)

\( H_l \)  
Filter function of microphone \( l \)

\( H \)  
Filter function in frequency domain

\( h \)  
Filter function in frequency domain

\( h_k \)  
Filter function of microphone \( k \)

\( \mathcal{I} \)  
Sound intensity

\( o \)  
Combination of VAD measurements

\( o_k \)  
Combination of VAD measurements at microphone \( k \)

\( \sigma \)  
Smoothed combination of VAD measurements

\( \sigma_k \)  
Combination of VAD measurements at microphone \( k \)

\( P \)  
Signal power
\( P_k \) Signal power of microphone \( k \)
\( P^b_k \) Signal power of microphone \( k \) at sub-band \( b \)
\( \overline{P} \) Estimated signal power
\( \overline{P}_k \) Estimated signal power of microphone \( k \)
\( \overline{P}^b_k \) Estimated signal power of microphone \( k \) at sub-band \( b \)
\( \mathcal{P} \) Magnitude spectrum
\( \mathcal{P}_k \) Magnitude spectrum of microphone \( k \)
\( \mathcal{P}_q \) Magnitude spectrum of microphone \( q \), where \( q \) is the closest microphone to microphone \( k \)
\( \mathcal{P}_{\eta_k} \) Magnitude spectrum of noise model \( \eta \) of microphone \( k \)
\( \overline{\mathcal{P}} \) Smoothed magnitude spectrum over time
\( \overline{\mathcal{P}}_{\eta_k} \) Smoothed magnitude spectrum of modelled noise of microphone \( k \)
\( \overline{\mathcal{P}}^*_{\eta_k} \) Smoothed noise magnitude spectrum of classified noise type \( n^* \) of microphone \( k \)
\( R \) Cross correlation value
\( R_{kl} \) Cross correlation value between microphone \( k \) and \( l \)
\( R_{kn} \) Cross correlation value between noise type \( n \) and the signal at microphone \( k \)
\( r \) Mahalanobis distance
\( r^n \) Mahalanobis distance of noise type \( n \)
\( r^{n*} \) Mahalanobis distance of classified noise type \( n^* \)
\( S \) Speech signal in frequency domain
\( \hat{S} \) Estimated speech signal in frequency domain
\( s \) Speech signal in time domain
\( T \) Threshold
\( T_B \) Threshold of background subtraction
\( T^p_k \) Threshold for VAD based on signal power at microphone \( k \)
\( T^C \) Threshold for VAD based on noise classification
\( T_v \) Threshold for final voice activity detection
\( T^b_k \) Threshold result for voice activity detection of microphone \( k \) at sub-band \( b \)
\( V \) Matrix of eigenvector
\( V' \) Reduced matrix of eigenvectors
\( V'^n \) Reduced matrix of eigenvectors of noise type \( n \)
\( v \) VAD result
\( v_k \) VAD result at microphone \( k \)
\( w \) Weighting factor
\( w_k \) Weighting factor for input signal at microphone \( k \)
\( X \) Audio signal in frequency domain
\( X_k \) \( X_l \) Audio signal at microphone \( k, l \)
\( \bar{X} \) Frequency vector over all \( k \) microphones
$x$ Audio signal in time domain
$x_k, x_l$ Audio signal at microphone $k, l$
$x'$ Windowed audio signal in time domain
$x'_k$ Windowed audio signal of microphone $k$
$Y$ Output of the delay and sum beamformer in the frequency domain
$y$ Output of the delay and sum beamformer in the time domain
$\alpha$ Angle of arrival
$\bar{\alpha}$ Mean angle of arrival
$\hat{\alpha}$ Estimated angle of arrival
$\beta$ Floor function for spectral subtraction algorithm
$\Gamma$ Cross power spectrum
$\gamma$ Exponential updating factor
$\gamma_s$ Updating factor for smoothing the steered response power distribution
$\gamma_\eta$ Updating factor for magnitude spectrum estimation
$\gamma_p$ Updating factor for signal power estimation
$\gamma_D$ Updating factor for feature set $D$
$\eta$ Noise signal in time domain
$\lambda$ Acoustic wavelength
$\mu$ Mean
$\mu_s$ Mean of threshold function for smoothed steered response function
$\mu_\rho$ Mean of likelihood function for detecting the face
$\mu^n$ Mean of test samples for Mahalanobis distance measure of noise type $n$
$\nu$ Scaling factor
$\nu_p$ Scaling factor for noise power threshold estimation $T^p_k$ and $T^b_k$
$\nu_{C_2}$ Subtraction factor
$\nu_b$ Scaling factor for combining VAD result over all sub-bands $b$
$\nu_r$ Scaling factor to control sensitivity of VAD
$\nu_w$ Scaling factor for microphone ranking
$\nu_s$ Scaling factor for smoothing the steered response power distribution
$\Theta$ Entropy
$\rho$ Likelihood function for detecting the face
$\Sigma$ Covariance
$\Sigma^n$ Covariance of training data for noise type $n$
$\sigma$ Standard derivation
$\sigma_s$ Standard derivation of threshold function for smoothed steered response function
$\sigma_\rho$ Standard derivation of likelihood function for detecting the face
$\tau$ Time delay
\[ \tau_{kl} \text{ Time delay between microphone } k \text{ and } l \]
\[ \tau_k, \tau_l \text{ Time delay for microphone } k, l \]
\[ \hat{\tau} \text{ Estimated time delay} \]
\[ \hat{\tau}_{kl} \text{ Estimated time delay between microphone } k \text{ and } l \]
\[ \omega \text{ Frequency component} \]

### 0.2 Indexing or counting variables

- \( a \) Number of attributes of features vector \( A \)
- \( a' \) Number of attributes of reduced feature vector \( A' \)
- \( b \) Sub-band index for sub-band signal power estimation
- \( c \) Detection count of noise type \( n \) over all \( K \) microphones
- \( i, I \) Time block index of audio frame and total number of time blocks
- \( j, J \) Frequency index and total number of frequency components
- \( K \) Total number of microphones
- \( k, l \) Microphone index
- \( k' \) Microphone index of the microphone with best directed speech path
- \( m, M \) Mel scale index and total number of Mel scale filters
- \( n, N \) Noise source index and total number of noise sources
- \( n' \) Classified noise type
- \( n'_k \) Classified noise type at microphone \( k \)
- \( n^* \) Classified noise type over all \( K \) microphones
- \( q \) Index of microphone that is closest to microphone with index \( k \)
- \( u, U \) Video bin index and total number of video bins
- \( t \) Audio sample index in time domain
- \( x, y \) Pixel coordinates in the video frame
- \( \delta \) Delta between index variables
- \( \epsilon \) Time index for VAD smoothing
- \( \psi \) VAD labels
- \( \psi_{no} \) No speech
- \( \psi_{low} \) Speech signal with low energy
- \( \psi_{high} \) Speech signal with high energy

### 0.3 Constant

- \( c \) The speed of sound \((\approx 340 \text{ m/s at } 15 \degree C)\)
- \( j \) \( \sqrt{-1} \)

xvii
Chapter 1

Introduction

Safety and security have become a critical concern in recent years. With the rapid progress of networking and storage technologies, surveillance systems have evolved from just observing small and well defined areas to systems that span over a large area like airports, mass transit systems, commercial hubs and shopping centres. The most commonly used sensors for infrastructure surveillance are undoubtedly CCTV cameras as they are able to provide quite a comprehensive visual communication of the monitored area. As the area of observation is increased more cameras are required, therefore, new ways of accessing and processing this vastly increasing amount of data has been the focus of current research (Greenhill and Venkatesh, 2006). Lately there is a growing trend to complement visual surveillance with non-visual sensors to provide better situation awareness of the monitored area. This enables the system to compensate for sensor limitations or to increase the observed information content for detecting multi-modal events. For example, to overcome the limitation of CCTV cameras that are unable to record meaningful data at night because of the lighting condition, a passive infrared sensor can be incorporated to trigger a floodlight and illuminate the scene when activity is detected (Bryant and Braun, 2003). In Zhu and Huang (2007) an experimental study based on a pan tilt and zoom camera, an infrared camera and laser Doppler vibrometer is presented for long range voice detection in the context of multi-modal surveillance systems. A human based tracking system consisting of multiple cameras and passive infrared sensors is investigated by Prati et al. (2005). This system was able to separate between background motion and moving people and detect direction changes of the target even when it is occluded in the video recordings. Increasing the information content for surveillance related events, the combination of audio and video sensors are the obvious choice as humans dominantly rely on sight and hearing for observing an environment. Hence, the authors in Pradeep et al. (2006); Cristani et al. (2007) investigate a system for multi-modal event detection and work by Asano et al. (2004); Maganti et al. (2007) investigates the speech enhancement abilities of such systems.

Audio can greatly contribute to enhance the information content of multi-modal surveillance events, such as the identification of the active speaker or enhancing a particular
CHAPTER 1. INTRODUCTION

source. However, extracting useful information from audio for surveillance purposes has a few considerable challenges. The main reason is that the recording of a desired source signal is usually mixed with ambient noise. Therefore, research has been conducted over the years to process audio for enhancing a particular source, detecting voice activity and classifying sound samples. Based on this research audio can be used to detect events such as shooting (Ledeczi et al., 2005), screaming (Ntalampiras et al., 2009), scratching surfaces (Tan et al., 2007), speech (ITU-T, 1996; ETSI, 1999; Gazor and Zhang, 2003) etc. For enhancing a desired signal out of the mixture with ambient noises several techniques have been developed that utilise either multiple microphones in a known or unknown alignment (McCowan, 2001; Benesty and Huang, 2003; Christensen and Hald, 2004; Brandstein and Ward, 2001; Hild, 2003; Pedersen et al., 2007) or a single microphone (Boll, 1979; Benesty et al., 2005). Another challenge when using audio is that the signal power decreases exponentially when the distance between the microphone and the location of the sound source is increasing.

This thesis explores the use of audio in the surveillance area. First, a combined system of video and audio is developed for capturing and indexing surveillance events with multimodal labels. For example, the system detects visually the event “gathering” when two or more people get together and uses enhanced audio to provide better event understanding. The audio component of this system is based on a uniform and linear microphone alignment that is known in advance. The motivation for this investigation is the development and evaluation of a generic framework that dynamically calibrates the observation of the different sensor types in a common space. We further investigate the audio component for an arbitrarily distributed microphone network that is scalable for large scale surveillance tasks. This distributed microphone network is able to enhance speech by detecting voice activity and classifying the ambient noise for modelling the noise type and removing it from the recorded mixture of speech and noise.

1.1 Aims and approach

The focus of this thesis is the use of acoustic information, essentially detected speech sequences, for surveillance purposes. Therefore, the combination of audio and video cues and a distributed network of microphones are investigated.
1.1.1 Dynamic calibration of audio and video cues

The first general aim of the thesis is the development and evaluation of a generic framework that uses coordinated audio and video cues for extracting surveillance related events. The following are the detailed aims of this framework presented in Chapter 3:

- A novel smoothing approach that uses the estimated accuracy of the sound source location to minimise the estimation error caused by reverberations or other ambient noise sources.
- Dynamic learning of a non-linear function that maps the one dimensional acoustic source direction to the two dimensional location of a target object in the video image; enabling the system to record audio discriminately based on the selected image location.
- Detection of surveillance events with multi-modal labels. These events can be used to inform a user of a situation of interest and provide visual and spatially filtered audio to assess the situation.
- Evaluation of the acoustic source localisation in terms of accuracy and distance from the microphones. The final assessment of the proposed framework is conducted by a qualitative evaluation of the extracted surveillance events.

These aims are achieved by developing a generic framework for capturing and synchronising the sensor information that are recorded with different sampling rates. Audio is recorded from a linear and uniformly distributed microphone array and video from a single CCTV camera. Initially, each media is processed separately to extract high level information. Such high level information consists of foreground objects of the video image and direction of the sound source based on the centre of the microphone array. Audio is further processed to extract characteristic features of speech for detecting instances with voice activity. The obtained high level information is subsequently used to dynamically calibrate the modalities into a common space. After the automatic calibration of the system, the image location of a “target” can be utilised to steer an audio beam towards the particular location for recording spatially filtered audio.
1.1.2 Distributed microphone network for acoustic surveillance

Based on the investigation of the acoustic source localisation and speech enhancement of a linear microphone array, the focus of the remainder of the thesis is directed towards acoustic surveillance based on a distributed microphone network, presented in Chapter 4 and 5. In particular, the aims of such a network are as follows:

- Generation of a “virtual” microphone that is able to record the speech signal with the best possible signal to noise ratio based on the unprocessed audio data;
- Reliable noise classification of multiple, non-stationary sound sources;
- Voice activity detection in non-stationary and rapidly changing noise situations;
- Speech enhancement based on a recording of a single microphone in real world conditions.

The generation of a “virtual” microphone is achieved by distributing the microphones over the entire area of interest and dynamically selecting the microphone within the best direct speech path of the target. However, it was discovered that such a technique does not enhance the desired speech signal when masked by environmental noise. Therefore, an improved spectral subtraction approach is applied to subtract the model ambient noise to enhance the speech signal. The proposed approach uses multiple modelled noise types to instantly adjust for rapid changes of the ambient noise type. These modelled noise types are also updated during detected speech sequences by utilising the entire network for dealing with non-stationary noise situations. When using multiple noise models, it is implied that the current ambient noise type must be classified first. Therefore, a variety of acoustic features are extracted in order to learn the characteristic sub-space for each noise type via principle component analysis (PCA). These noise characteristics are used to determine the closest match of a test sample when projected into each sub-space and computing the Mahalanobis distance metric. The last requirement for the proposed speech enhancement approach is the detection of speech sequences. Therefore, the distance metric of the noise classification approach is combined with the signal power estimation to detect voice activity. This combination significantly reduces false detection in non-stationary noise backgrounds.
1.2 Significance and contribution

This work makes two main contributions to the field of surveillance - (1) the dynamic calibration of audio and video cues by a generic framework and (2) a large scale speech enhancement approach based on a distributed acoustic surveillance network. Following is the significance of the proposed approaches.

For combining sensors it is essential that the sensor informations are fused into a common space to extract events based on all available information. Such a common space is created by calibrating the sensors. In general, there are two ways of calibrating the sensors: Statically by assuming a known sensor alignment; or dynamically by using training data for estimating a function that correlates the sensor information. In real environments the sensor alignment can change over time through vibration, wear and improper installation, so only a dynamic calibration is considered suitable for such situations. Therefore, the presented approach automatically identifies a moving and talking person to estimate a mapping function that correlates the one dimensional audio and the two dimensional video data dynamically. The proposed framework uses the correlated data to assess the situation and highlight particular situations of interest by generating a surveillance event. Such an event is composed of video and enhanced audio of the area of interest to provide an operator with multiple media to assess the situation.

Based on the results of the audio-video surveillance system, a distributed network of microphones is further investigated. Such a network provides a cost effective solution for covering large areas of interest for acoustic surveillance. By distributing the microphones around the area of interest the distance between a target source and the closest microphone is minimised, compared to an array based solution where the microphones are generally aligned closely together. The speech enhancement of the microphone network is realised by a modified spectral subtraction approach. The modification enables the system to adjust to complex ambient noise situations where the ambient noise can rapidly change and is generally non-stationary. In such a situation it is challenging to detect speech activity, therefore the distance metric of the noise classification is utilised to considerably reduce false detection rates during rapid noise type changes or significant intensity changes of the ambient noise.
1.3 Thesis outline

The thesis is organised as follows: The context and motivation for the work is introduced in Chapter 2. It presents an overview of surveillance systems for different kinds of sensors and system architectures. The requirements and advantages of the most commonly used sensor, a CCTV camera, are highlighted and the challenges of incorporating other sensors to a video surveillance domain are introduced. Issues of current technologies are examined and it is shown how audio is used for enhancing the capabilities of a surveillance system. Therefore, the focus is directed to research areas containing concepts of acoustic features, voice activity detection, sound classification and speech enhancement to utilise them for acoustic surveillance purposes.

Chapter 3 presents an initial investigation into the benefits of the coordinated use of video and audio cues to capture and index surveillance events with multi-modal labels. The focus of this chapter is the development of a joint-sensor calibration technique. The calibration is dynamically estimated and therefore does not require prior knowledge of the relative sensor alignment and is resilient to drifts. Experiments are conducted that examine the accuracy of the calibration and higher level audio information. Finally, the proposed surveillance system is evaluated for different scenarios.

To improve the acoustic surveillance for a large area of interest, Chapter 4 examines a distributed microphone network for capturing the audio of a target person with the best signal to noise ratio and enhancing the speech signal. Techniques are investigated to detect speech sequences and identify the ambient noise type. Speech enhancement is achieved by improving the performance of the general spectral subtraction approach (Boll, 1979) for real world noise situations. The significant contribution is that this proposed approach can adjust to rapid changes in ambient noise during speech instances. Experiments demonstrate the advantages of the improved speech enhancement for synthetic and real noise situations.

Chapter 5 examines further the findings of the distributed microphone network by developing a speech enhancement and noise classification approach that accounts for real world situations. The noise classification is developed further using a variety of acoustic features that are used to learn the characteristics of the noise type via PCA. This improves the reliability of classifying multiple, non-stationary noise sources. A significant improvement to enhance a speech signal in non-stationary noise situations is achieved by updating the modelled noise types during speech sequences, this time utilising the entire network. Experiments are conducted to evaluate the noise classification, speech detection
and enhancement approach for real world scenarios.

Finally, in Chapter 6 a summary of the thesis is given along with possible future directions for acoustic surveillance.
CHAPTER 2

BACKGROUND

2.1 Introduction

The scope of this thesis addresses the issues associated with the use of multiple sensors with specific application to surveillance. It includes a variety of elements from the acoustic and image processing areas such as object or target localisation, event detection and signal enhancement. This chapter reviews the literature for all applied sub-fields in this thesis.

The chapter is organised as follows: An overview of surveillance related tasks is given in section 2.2 to provide the reader with a background on issues and implementations regarding audio and audio-video based surveillance systems. This is followed by issues, solutions and approaches for acoustic data processing for surveillance related tasks. Section 2.3 explains the need and techniques for acoustic feature selection and dimensionality reduction. Section 2.4 provides approaches for sound classification, including voice activity detection. An essential task for analysing an area of interest and gathering surveillance events is the determination of the location of a sound source. Thus, section 2.5 presents the state of the art localisation techniques for cross sensor alignment. Section 2.6 provides an overview of speech enhancement techniques, which is a key part of acoustic surveillance. The chapter concludes with a summary in section 2.7.

2.2 Surveillance tasks and network architecture

Surveillance systems are widely accepted these days and viewed as a part of protecting public spaces rather than as an intrusion. This acceptance of surveillance systems in public allows effective monitoring of areas of interest. Most of the currently deployed systems basically record the video with the intention of then getting monitored by operators. However, as the number of sensors increases, the focus is directed towards autonomous systems that can alert an operator to particular situations of interest. Principally, if a
surveillance system with any kind of sensors is to analyse the observed scene automatically, it needs to address the following tasks:

- **Target localisation**: The fundamental task is to detect a target of interest, generally a person. However, this target may not be observed by all sensors or sensor noise prevents accurate localisation. For CCTV based systems, certain localisation tasks can be achieved by analysing the video stream and detecting foreground objects (Comaniciu et al., 2003; Mittal and Davis, 2003). In acoustic systems, localisation can be realised by detecting speech and computing its source (Cha et al., 2008; Ma et al., 2006).

- **Event detection**: Detecting pre-defined or unknown events is essential for informing an operator of a situation of interest or generating a protocol of the observed area. Such events can be unusual target tracks or locations, gathering of multiple targets, single or sequences of sounds. Commonly, this is realised by training the system with anticipated target behaviour and sound signals. Widely used classification approaches are Gaussian Mixture Model (GMM) (Radhakrishnan et al., 2005), Hidden Markov Models (HMMs) (Rabiner, 1989) or Support Vector Machine (SVM) (Vapnik, 2000).

- **Enhancement**: If an event of interest is detected, further signal processing techniques can be applied to enhance the recorded data. In particular when the data is noisy, it is important to enhance the data and increase the information content of the scene. For example, audio recordings are always masked by environmental noise that can be removed to some extent by single or multi-channel techniques (Brandstein and Ward, 2001; Benesty and Huang, 2003; Benesty et al., 2005). For video data the resolution is problematic when monitoring a larger area, making detailed observations difficult. The only solutions for this problem are the installation of a pan tilt and zoom camera or implementing a super-resolution technique (Borman and Stevenson, 1998).

Most surveillance systems are deployed for long period of time at fixed locations. The sensors are therefore connected by cables and powered by the grid, making the system more reliable and reducing maintenance. Nevertheless, sometimes a surveillance system must be deployed quickly and only for a short period of time, in which case wireless sensors are often used. Such sensors include wireless cameras and microphones as well as multi-purpose sensor nodes with several sensors (passive infrared motion sensor, magnetometer, microphone or radar motion sensor) and a microprocessor. However, such sensors are
often restricted in terms of power consumption, processing capability, sensor resolution, bandwidth and transmission reliability (Yick et al., 2008). Another disadvantage is that such sensors are more expensive and more complex to synchronise. Wireless networks of multi-sensor nodes are often developed for military purposes. For example, work by Oh et al. (2007) investigated the use of such networks in pursuit and invasion scenarios, Ledeczi et al. (2005) for detection and localisation of snipers and Sheng and Hu (2005) to detect and track an amphibious assault vehicle on a road. However, the scope of this thesis focuses on fix surveillance infrastructure, thus we chose a wired solution.

2.2.1 Audio surveillance

Acoustic information can significantly contribute to analysing a scene of interest. The reason is that for humans, hearing is the second most important sense after sight. Acoustic information provides vital information about the scene content and alerts us to an unusual situation. Another advantage of utilising sound information is that general microphones are inexpensive and are able to capture the entire surrounding area in comparison to a CCTV camera that only records a certain view. Therefore, acoustic based surveillance systems are developed to provide scene analysis for detecting unusual events.

Acoustic scene analysis faces a few challenges because of the characteristics of sound in terms of propagation and its interference with other sound signals. Following are the major challenges that have to be addressed when designing an audio surveillance system:

- **Background noise**: In real world situations, the sound recording is usually a mixture of sound signals from various sources, that can be correlated (the reflection of a signal or a signal that is broadcasted by multiple loudspeakers) or uncorrelated (different active sources at the same time). For example, a speech signal in outdoor environments is often masked by wind or the noise of passing cars. In indoor situations, machines or air conditioning vents introduce noise that masks the desired signal.

- **Reverberations**: This acoustic phenomenon is usually restricted to indoor environments and produces correlated and time shifted signals of the source signal. The reason being sound propagates in all directions from its origin and is reflected by walls or other surfaces. Thus, a microphone not only records the direct path of the signal but also a time delayed version, as shown in figure 2.1. The delay is referred to as the reverberation time and is defined by the time (in seconds) when the sound
signal decreases by 60 dB after the source stops emitting the signal. Sabine (1993) presents a formula to compute the reverberation time based on the room volume and surface absorption.

- **Signal power**: The signal power of the desired source is an important factor as it determines how well it can be detected and recognised. There are techniques that are able to enhance the source signal in some conditions, such as beamforming (McCowan, 2001; Benesty and Huang, 2003; Christensen and Hald, 2004), spectral subtraction (Gustafsson et al., 2001; Benesty et al., 2005; Wójcicki et al., 2006; Breithaupt et al., 2007) or blind source separation (Brandstein and Ward, 2001; Hild, 2003; Pedersen et al., 2007). However, if the microphone is too far from the source or no directed speech path can be recorded it is hard to enhance such signals. In these situations the only solution is the deployment of more microphones or microphone arrays that are spread around the area of interest. An indication on how fast the signal power decreases is given by the inverse square law, which states that by doubling the distance between two measurement points, the signal intensity level drops by about 6 dB (Howard and Angus, 2006). Thus, this thesis investigates the use of a distributed microphone network for recording and enhancing a speech signal.

- **Direct speech path and speech pauses**: For localisation of a speech source, it is essential that the microphones record the direct speech path, otherwise the source location is estimated based on the reflected signal by other surfaces. Should the source be tracked over a series of measurements it is important to account for speech breaks, especially for longer pauses. Therefore, tracking of a target person is much harder with acoustic measurements than with video measurements.

![Diagram](https://example.com/diagram.png)

**Figure 2.1**: Illustration of reverberation effect for an enclosed environment. The recording microphone is denoted by $\bigcirc$.

Acoustic scene analysis can be divided in four sub-tasks: 1) extraction of suitable acoustic features that characterise the desired sound, 2) the detection of an acoustic event, 3) the
localisation of the acoustic event for a single time instance or a time series and 4) the enhancement of the acoustic source when the environment is noisy.

- **Feature extraction**: For detection of acoustic events, it is essential to describe the characteristics of the sound by extracting suitable features. The event detection rate is directly related to how well the features describe the sound and how discriminative they are among different sound events. Appendix A gives a brief overview of commonly used features and in section 2.3 techniques for selection relevant features are described.

- **Event detection**: Detecting situations of interest is an important component of any surveillance system. Such situations are generally called events and consist of one or more acoustic sources. For surveillance purposes, it is desirable to detect events that are connected to hazardous or unusual situations. In general, supervised learning techniques are applied to detect such events by defining the events of interest and providing labelled data of the events. State of the art techniques and approaches for detecting events are surveyed in section 2.4.

- **Source localisation**: For many applications it is not only important to detect the event but also locate where it occurs. Therefore, sound source localisation (SSL) algorithms are developed utilising the spatial separation of two or more microphones to compute the location of a sound source of interest. Different techniques are proposed using a known linear or circular array geometry (Nordholm et al., 2004; Dmochowski et al., 2007b) or a distributed network (Meesookho et al., 2008). These techniques are investigated in detail in section 2.5.

- **Source tracking**: Tracking a sound source over time is difficult because the source localisation algorithms generally use small time frames to estimate the source location and sound might not be emitted from the source every time frame. For example, if speech is the desired target signal, the target path must be predicted during speech pauses. The most commonly used tracking algorithms in literature are the particle filter, also known as sequential Monte Carlo method, (Doucet et al., 2000; Lehmann, 2004) or the Kalman filter (Bar-Shalom and Li, 1993; Klee et al., 2006). Both filters provide a statistical technique to model recursively the state of a system from a series of noisy measurements. The Kalman filter assumes linear and the particle filter non-linear dynamics respectively.

- **Enhancement**: The recorded audio signal from a detected and localised event is usually mixed with other undesired sound signals, making perception difficult. If the target signal is speech, it is beneficial to enhance the signal by separating it,
thus achieving better event understanding or improved results in automatic speech recognition. Therefore, several techniques utilising multiple or single microphones are examined in section 2.6.

2.2.2 Audio-Video surveillance

Combining audio and video information increases the observed information content of an area of interest significantly. A video stream can provide a general view of the scene whilst acoustic information can indicate who is speaking, whether there is an unusual sound event, or where the location of the sound source is. Additionally, each sensor strength can compensate for a weakness of the other sensor when the noise is independent and disparate. Challenges in audio based scene analysis was highlighted in section 2.2.1. The fundamental tasks in video based surveillance systems is object segmentation, which faces the following problems:

- **Occlusion**: Segmenting foreground objects for recognition or location estimation is essential. However, occlusions of the target object by other objects occur frequently, making accurate localisation and recognition difficult (Hu et al., 2004). One solution is to employ multi-camera systems, observing the object from various view points (Mittal and Davis, 2003).

- **Low light situations**: Most CCTV cameras are not able to deal with low light situations, making object appearance and colour perception difficult (Haritaoglu et al., 2000).

- **Object or target identification**: The appearance of objects can change considerably depending on the viewing angle, making it difficult to generate a consistent appearance model (Foresti, 1999).

- **Processing time**: Current automatic video based algorithms are computationally complex and extremely difficult to scale when number of cameras increases (Wakin et al., 2006). Additional computational resources are usually required to deal with automatic tracking or recognition over multiple cameras (Hu et al., 2004).
2.2.2.1 Sensor calibration

Importantly, when using audio and video sensors it is essential to map the measurements into a common space. This is achieved by calibration of the sensors by either knowing their relative sensor geometry or by estimating a mapping function using training data. Assuming a fixed and therefore known sensor geometry is certainly the easiest approach. Systems with known sensor alignment are presented in (Vermaak et al., 2001; Chen and Rui, 2004) for the task of identification and tracking of single or multiple speakers and in (Smeaton and McHugh, 2005; Pradeep et al., 2006) for event detection. For small systems such as video conferencing this assumption of a fixed sensor geometry may be reasonable, but for larger systems such as surveillance systems this assumption is no longer valid. The reason is that the sensor alignment can shift over time because of external causes or the time shift of a sound event between two or more spatially separated microphones, depending on the speed of sound in the air and that again depends on temperature. Thus, learning a calibration function that maps the sensor measurement into a common space makes the systems more resilient. For computing the calibration, training data from a single object that produces location information in both modalities is generally collected (Asoh et al., 2004; Gatica-Perez et al., 2007; Maganti et al., 2007). This process requires more time than a known sensor alignment, but when implemented dynamically, the system adapts to changes and can therefore be deployed more reliably.

2.2.2.2 Data fusion

The fusion of audio and video measurements is generally used for improving the accuracy of target tracking and allows the detection of more diverse events. For tracking, there are two common techniques for combining the different modalities: Particle filter (sequential Monte Carlo method) (Doucet et al., 2001) and probabilistic generative models (graphical models) (Bishop, 2006). The particle filter converts the location information of the video and audio measurements into a probability density function (see figure 2.2) that is used to update a state of a target which fundamentally contains the location and velocity of the target. Cevher et al. (2007) applied this technique for fusion to track a vehicle even when it is occluded in either modality. The authors presented an adaptive time-synchronisation technique for multi-modal data to enhance the tracking accuracy. Chen and Rui (2004) proposed a particle filter based fusion algorithm that utilises the results of an individual tracker from a single modality. The results of the combined location estimation is then recursively integrated into the individual tracking algorithm. In probabilistic generative models, a graph denotes the conditionally independent structure of the two observed
modalities. This graph models the data association in terms of internal variables, as shown in figure 2.3. Hospedales and Vijayakumar (2008) and Beal et al. (2003) presented a similar model that associates the audio and video data during training by estimating all parameters of the model. These frameworks are then tested on data recorded in indoor environments where the targets are occluded or other non-speaking targets are in the background. Experiments in an outdoor environment are presented by Kushal et al. (2006) using a modified model of Beal et al. (2003). What all these models have in common is that they use a small sensor set with only one camera and two microphones. This is because the entire system needs to be modelled and when the number of sensors increases the models become more complex, making them impractical for larger systems as compared to particle filters.

![PDF of Particle Filter](image)

Figure 2.2: Probability density function (PDF) of the particle filter. Particles are represented as circles and are weighted based on the sensor measurements. Weighted particles are represented by the filled circles where the size indicates the importance. The source localisation is indicated by the maximum of the PDF that is modelled by the weighted particles.

2.2.2.3 Audio-Video systems

In surveillance, knowing where a target object is located is important but detecting events that indicate unusual situations or providing events that describe a situation of interest in more detail is as important, if not more important. Therefore, using audio and video to capture and index surveillance events with multi-modal labels significantly improves the ability to analyse a situation compared to using a single modality. The authors in Pradeep et al. (2006) propose a multimedia surveillance system for surveillance events based on a hierarchical probabilistic framework. They investigate the questions *when*, *what* and *how* to assimilate the data of different modalities. Each sensor is first used to extract individual events that are later combined into compound events based on two or more modalities. Fusion is achieved by learning the likelihood of detecting events of a single modality between different and spatially separated sensors at the same time. The system is evalu-
CHAPTER 2. BACKGROUND

Figure 2.3: Graphical model from Hospedales and Vijayakumar (2008) for audio-video association. Observed measurements are illustrated as filled nodes. The modalities are connected by associating the estimated source location \( l \) and the time delay \( \tau \) of the acoustic signal between the spatial separated microphones. © 2008 IEEE

rated on events such as standing, walking, running, standing/talking, standing/shouting, standing/door knocking, walking/talking etc. Results show that the false rejection rate decreases on average by 7.65% and that the mean accuracy increases by 20.0% when using multi-sensor informations in comparison to individual sensor information.

Another audio-video system in the field of surveillance is presented by Cristani et al. (2007). Their method integrates modalities from one camera and one microphone for automatic scene analysis. The algorithm first extracts foreground patterns in each modality and fuses the patterns based on the degree of synchronisation between the modalities via an audio-video concurrent matrix. This audio-video association is achieved without training by learning the association online. However, it is assumed that both sensors are able to observe the cause of the event. During evaluation, events such as making a call, receiving a call, first at work, etc. are observed and discriminated.

For improving the recorded audio quality of a speech event, Maganti et al. (2007) propose an integrated framework based on audio and video measurements. Both modalities are used for tracking multiple people and their speech activity by fusing the data via a particle filter. The combined speaker location is then used to direct a beamforming algorithm to enhance the speech signal. Their approach is evaluated for a multi-person meeting scenario in terms of signal to noise ratio and automatic speech recognition, comparing it to a single microphone recording. Results show an enhancement when both sensor informations are
used to steer the beamformer compared to enhancing the speech based on acoustic source localisation alone. However, the system needs to be calibrated by using a checker board for camera calibration and an undefined acoustic reference for computing the audio-video calibration.

Another speech event detection, enhancement and recognition approach is presented by Asano et al. (2004). Their system consists of a circular microphone array and a stationary stereo camera mounted above the array. The calibration of the sensor data is achieved by recording a single person who is stationary at defined positions and talking in an environment with little noise. Both modalities are then used for tracking multiple persons using a Bayesian network. The location of the persons are used to direct a maximum likelihood adaptive beamformer for enhancing the speech signal and performing speech recognition.

2.3 Audio feature processing

Appendix A provides a brief overview on the most commonly used acoustic features. In principle, no acoustic feature on its own is capable of differentiating between many different sounds. The main reasons are:

- Sounds, such as speech or music, have significant different signal characteristics;
- The characteristics of a particular sound type, such as speech, varies even during short periods of time;
- Different sound sources can be mixed together, making discrimination difficult.

Hence, usually several different features are extracted and combined into a feature vector describing the audio data, representing as much of the sound characteristics as possible. However, this approach can be misleading, as by just increasing the number of features the classification accuracy may not improve (Guyon and Elisseeff, 2003). The reason is that by increasing the number of features more and more features can have similar properties among several sound types, making them similar during classification (Fodor, 2002; Chizi and Maimon, 2005). Therefore, an essential step after feature extraction is to select a good set of features using classification accuracy as the validation criteria. The selection of features can be realised by either evaluation of each feature and ranking them or reducing the dimensionality of the feature vector by projecting them into a sub-space.
Ranking approaches usually do not select a sub-set of features, rather they indicate the importance of each feature based on a variety of methods (Koller and Sahami, 1996; Dash and Liu, 1997). The approaches are commonly divided into two categories: Filter and wrapping approaches (Chizi and Maimon, 2005). Filter approaches are characterised by selecting a sub-set of features via a pre-processing step independent of the classification approach applied. Well known supervised techniques are Relief (Kira and Rendell, 1992) and information gain (IG) (Lee and Lee, 2006). These approaches rank the features based on information theoretic criteria, the mutual information between the individual features for a single class label. However, these approaches are ineffective in removing highly correlated and therefore redundant features (Koller and Sahami, 1996; Dash and Liu, 1997). On the other hand, wrapping approaches are defined by incorporating a classification algorithm that evaluates a feature sub-set by adding or removing a single feature (Das, 2001; Peng et al., 2005). The usefulness of this feature is then evaluated using a cross-validation of the new sub-set in a one versus all class problem involving the use of training data. In principle, any classification approach can be used for evaluation, i.e. Naive Bayes (NB), support vector machine or linear discriminant analysis (LDA) (Peng et al., 2005; Witten and Frank, 2005). The advantage of wrapper approaches is that the feature selection is based on an induction algorithm (Das, 2001). Hence, the selection is already based on maximising the classification accuracy between the test classes. However, the use of an induction algorithm in combination with a cross-validation makes such algorithms computationally complex compared to filter algorithms (Das, 2001).

An alternative to ranking approaches are the family of dimensionality reduction techniques. These algorithms try to reduce the feature space by mapping the original features into a sub-space of lower dimensionality. The advantage of this approach is its computational efficiency in comparison to feature ranking approaches (Hyvärinen and Oja, 2000). Widely used dimensional reduction techniques are principle component analysis (Fodor, 2002), also known as singular value decomposition (SVD) or Karhunen-Loève transformation, and independent component analysis (ICA) (Hyvärinen and Oja, 2000). PCA is a second-order statistic method that does orthogonal linear dimension reduction and performs best in terms of mean square error for Gaussian distributed data (Fodor, 2002). ICA is a higher-order statistic model that uses a linear transformation of non-Gaussian and independent data for feature reduction (Hyvärinen and Oja, 2000). The main difference between both algorithms is that PCA looks for uncorrelated features and ICA for independent features (Fodor, 2002).
CHAPTER 2. BACKGROUND

2.4 Audio classification

This section gives an overview of commonly applied classification algorithms that are used in acoustic surveillance systems. The focus is directed to two particular tasks: General multi-class event detection such as explosions, screaming, footsteps etc. and two-class problem of speech versus non-speech classification.

2.4.1 Acoustic event detection

Detecting situations of interest (events) is an important component of any surveillance system. For surveillance purposes it is desired to detect events that are connected to hazardous or unusual situations. In general, supervised learning techniques are applied to detect such events by defining the events of interest and providing the classification approach with test samples of such events. To enhance the classification accuracy and reduce the computational load for such a multi-class problem, events are often grouped into a two class problem and classified in stages (Lin et al., 2005; Atrey et al., 2006; Ntalampiras et al., 2009). The following are some commonly used classification techniques that are applied in recent work for surveillance related sound analysis.

Radhakrishnan et al. (2005) applies a Gaussian Mixture Model approach, that is commonly used in video processing to model the background scene (Stauffer et al., 2000), for estimating the ambient noise conditions of an elevator. The GMM is trained on 12 Mel Frequency Cepstral Coefficients to determine events or situations that do not match the modelled background noise. Outliers are manually labelled into two groups: Type 1 (false alarm or non suspicious events) and Type 2 (classes of suspicious events). Because the ambient noise generally changes over time the GMM is updated when the test sample lies within a threshold of the modelled background noise. The authors reported an accuracy of about 82% for detecting four defined events (banging, footsteps, normal speech and non-neutral speech) involving 61 test sequences using one hour of data. However, no results are reported on the performance for long term observations and how the background model would adapt to long term changes of the ambient noise.

Atrey et al. (2006) further investigated the GMM for acoustic event detection by comparing a hierarchical top down approach to a single-level approach. In the hierarchical approach the audio is first classified into general background noise and events of interest. Then the events of interest are further separated into vocal and non-vocal events. Finally, non-vocal events are classified into the low-level events Knocking and Footsteps and vocal
events into Talking and Shouting. The approach is evaluated on a two hour audio stream that contains about 10 minutes of each low-level event. The classification results are compared between the single-level and the hierarchical approach as well as the numbers of Gaussian distributions and for varying time and frequency domain features. For the hierarchical approach a 90% accuracy is reported for vocal/non-vocal event detection using log-frequency Cepstral coefficients. The low-level events Talking, Shouting, Knocking and Footsteps are detected with an accuracy of about 70%. In comparison, the single-level approach was not able to achieve the same accuracy across all low-level event.

Hidden Markov Models are another common statistical approach to extract events from audio data. In Ntalampiras et al. (2009) a two stage schema based on HMMs is investigated to detect hazardous situation in a metro station. The authors used fully connected HMMs for detecting first typical and atypical situations, where atypical situations are further classified into Gunshots, Screaming and Explosions. This approach is evaluated on audio of professional sound effects databases and therefore effects like reverberations or wind noises are not considered. Experiments are conducted for events with a signal to noise ratio between -5 dB to 15 dB, where 0 dB is considered as the target signal strength for general operation of the framework. At this level the framework achieved a 6% false positive rate with an equal error rate between 8.54% and 24.5% for the three trained events.

Even though HMM are very common in automatic speech recognition (Rabiner, 1989; Gales and Young, 2007; O’Shaughnessy, 2008), for segmentation of acoustic events the support vector machine is used more widely. In Lin et al. (2005) the authors proposed a multi-class SVM method to classify audio clips. The audio is described by features based on the Fourier and wavelet transform. As a typical SVM is a two-class classifier (1-vs-1), the authors proposed a bottom up binary tree scheme that recursively compares two classes. Evaluation of the approach is conducted on the Muscle Fish database\(^1\), containing 410 sounds in 16 classes of high quality audio. The method uses an exponential radial basis function (RBF) kernel and is able to classify the data with an error of 3%.

In Rabaoui et al., 2008 a one-class SVM is proposed for sound-based surveillance applications. One-class SVMs are generally used for outlier detection by learning a function that describes a class or event by minimising the volume of a set of training samples in a multi-dimensional space. The authors applied such a classifier for a multi-class problem by estimating for each known acoustic event a separate classifier by learning the parameter for a Gaussian RBF kernel that describes the event. A dissimilarity measurement is then

\(^1\)http://www.musclefish.com
computed over all classifiers that either indicates a detected event, or when it is larger than a threshold, an unknown event or situation. The approach is evaluated for a nine class problem and compared across multi-class SVMs and a HMM classifier. Additionally various combinations of audio features are evaluated. Results on noise free data and data masked by stationary noise are reported with the highest accuracy for the one-class SVM. However, no results are reported for non-stationary noise situations and adaptation strategies are left for further research.

2.4.2 Voice activity detection

Detecting speech or the absence of speech is an essential requirement for a variety of applications, and commonly referred to as voice activity detection (VAD). For example, VAD is used in mobile communication to reduce co-channel interference and power consumption (Chena et al., 2007), reducing the data rate by silence compression via discontinuous transmission (Beritelli et al., 2002), enhancing the speech signal in noisy environments by estimating the noise characteristics (Benesty et al., 2005) and triggering automatic speech recognition systems (ETSI, 2007). Two telecommunication organisations, the International Telecommunication Union (ITU) and the European Telecommunication Standards Institute (ETSI), have proposed a number of standards for VAD that are frequently used for performance evaluations. ITU presented a speech coding algorithm named G.729 (ITU-T, 1996) that uses VAD to reduce the bit rate. This VAD technique is based on the following acoustic features: Linear predictive coefficients, full-band and low-band energy and zero crossing rate. ETSI proposed two VAD algorithms: ES 202 050 - Advanced front-end feature extraction algorithm (ETSI, 2007) for distributed speech recognition and ES 301 708 - Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels (ETSI, 1999). The advanced front-end algorithm uses two stages for VAD detection: First a frame by frame decision is made and the final VAD result is based on hangover filtering. The frame based decision is based on energy measurements of the full and sub-regions of the spectrum, as well as the acceleration of the energy measurements. In comparison, the AMR detector is proposed in two versions: The first uses the signal to noise ratio in combination with a pitch detector of different frequency bands. The second version is a further development of the first using parameters from the speech encoding algorithm (Ramirez et al., 2004).

In the open literature several other techniques have been investigated for speech detection. A fast and straightforward technique is the use of the energy measurement. Martin (1993) proposed an algorithm to estimate the noise spectrum over time and uses it to determine
CHAPTER 2. BACKGROUND

the signal to noise ratio for voice activity detection. A similar approach is investigated by Chena et al. (2007), in which they decompose the audio spectrum first into critical sub-bands using the perceptual wavelet-packet transform. These sub-bands are then used to compute the Teager energy operator (Kaiser, 1990) to determine speech content. However, energy measurement based approaches usually apply some kind of threshold function to determine speech content, making them inaccurate in non-stationary noise or for low signal to noise situations (Tanyer and Ozer, 2000).

Alternatively, speech sequences can be detected by statistically modelling the speech and noise characteristics (Sohn and Sung, 1998; Gazor and Zhang, 2003; Chang et al., 2006). The classification into either active speech or noise is determined by computing the likelihood ratio of a test signal to these models. Generally, all these models assume a certain distribution of the signal characteristics such as Gaussian (Sohn and Sung, 1998), Laplacian (Gazor and Zhang, 2003) or Gamma distribution (Chang et al., 2006). The evaluation in Chang et al. (2006) conducted on the parametric representation of the noisy speech spectra distribution suggested that the Laplacian and the Gamma distribution are more suitable than a Gaussian distribution. However, in real environments the noise characteristics are non-stationary and hence the model needs to be updated recursively. Depending on the update rate, this takes some time and therefore only slowly varying noise scenarios are modelled well (Gazor and Zhang, 2003). Additionally, such approaches are more computationally complex (Chang et al., 2006) than VAD using energy based features.

Supervised learning techniques are another common approach to detect or classify speech by using features that describe the characteristics of the audio. A frequently applied algorithm is a two-class support vector machine (Vapnik, 2000). In Enqing et al. (2002) the SVM is compared to VAD of G.729 Annex B (ITU-T, 1996) using spectral distortion, full-band energy, low-band energy and zero-crossing rate features. Kinnunen et al. (2007) applied Mel Frequency Cepstral Coefficients (MFCC) and Wang et al. (2008) uses ICA transformed MFCCs, total and sub-band power, brightness, bandwidth and pitch features as input data for SVM based VAD. However, using supervised learning techniques implies that representative training data are needed that describe the typical sound environment. Additionally, when using a number of different features, their numerical range must be normalised to achieve a high classification rate. Finding the optimal parameters of the classifier that maximise the classification accuracy of the training data is a time consuming task.

In contrast to the presented frame by frame VAD decision, the assumption that a speech sequence consists of several consecutive time frames can be used to improve the classifi-
CHAPTER 2. BACKGROUND

cation rate. The easiest approach is to apply a smoothing filter on the frame by frame decision such as the proposed hangover time in ES 202 050 (ETSI, 2007). Alternatively, a prediction algorithms can be applied such as a HMM (Gazor and Zhang, 2003) that uses a transition function to update the internal state of a system (detected speech/non speech).

2.5 Sound source localisation

In sound source localisation, the general aim is to localise single or multiple sound sources with high spatial precision. Historically, many SSL strategies are originated from array based radar or sonar localisation applications where the target signal is narrowband. However, for acoustic surveillance systems, the target signal is generally a broadband signal such as speech and the environment is usually noisy and reverberant. Therefore, filter algorithms like phase transform (PHAT), smoothed coherence transform (SCOT) or Roth processor (Knapp and Carter, 1976) are developed to reduce estimation inaccuracy due to signal characteristics and environmental conditions. In general, the accuracy of the source localisation is influenced by (Brandstein and Ward, 2001):

- The number and quality of the microphones;
- The microphone alignment;
- The distance of the source to the microphones;
- The environment that is characterised by the ambient noise and reverberation time;
- The number of active sound sources.

There are varied approaches proposed for localising a single or multiple sound sources. These strategies can be divided into four general categories:

1. Direction of arrival (DOA), where sources are located by using the phase difference between two or more microphones.

2. High-resolution spectral estimation that is derived from the spatio-spectral correlation matrix of the recorded array signals.

3. Time delay of arrival (TDOA) where the time delay of the recorded signal between two microphones is exploited to estimate the source location.
4. Energy based localisation where the fact is used that the sound energy decreases as the distance to the source increases.

2.5.1 Direction of arrival

The most commonly used DOA techniques are based on the maximisation of the steered response power (SRP) of a beamforming algorithm. These methods compute the source location based on the filtered, weighted and summed signal over all microphones. In the simplest case, a delay and sum beamformer (DSB) is used to compute a spatial spectrum or “energy map” of the area of interest by adjusting the steering delays to localise the acoustic source defined as peak of the spatial spectrum. An advantage of such methods is that the location is computed based on all microphones in one step by maximising the estimated signal power and that multiple sound sources can be identified (Nordholm et al., 2004; Zotkin and Duraiswami, 2004). However, it is computationally intensive especially when the numbers of microphones increases and the search area is multidimensional (Zotkin and Duraiswami, 2004), making real-time implementation difficult.

Intensive research has shown that the combination of PHAT filtering and SRP (SRP-PHAT) is well suited for speech source localisation in reverberant environments (Cha et al., 2008). The reason being that PHAT weighting reduces local maxima caused by reverberation and sharpens the peak of the true source location in the SSL function. Nordholm et al. (2004) evaluate the performance of three SSL techniques for real environments in terms of computational load and accuracy. The evaluation was conducted for coherent wideband Root-MUSIC and near-field and far-field SRP-PHAT for varying signal to noise situations. The stationary sound source was a female voice at a distance of 1.5 metres of an eight element microphone array in a real room and a synthesised environment generated by a free-field model. Results show that the two SRP implementations are most suited for noisy and reverberant environments where far-field SRP-PHAT performs as well as near-field SRP-PHAT, but is 10 times as fast.

The main restriction of SRP algorithms is the high computational demand by computing the sound energy map and searching for peaks. Hence, research has focused on developing SRP based techniques that decrease the complexity by maintaining accuracy. Zotkin and Duraiswami (2004) propose an hierarchical search algorithm that uses a coarse-to-fine strategy in space and frequency. First, the algorithm searches in a coarse grid for peaks by computing the sound energy from only low frequencies. This is only a reasonable assumption because the target signal is speech and the frequency structure of speech...
is mainly in the lower frequency ranges (Silverman, 1987; Baken and Daniloff, 1991). However, for low frequency ranges SRP algorithms can only estimate an approximate source location. Therefore, the grid is recursively subdivided and the frequency range is increased until the source location is computed within a defined accuracy. Such an approach reduces the computational load significantly by maintaining identical accuracy but assumes that for multiple sources the distance between them is large enough for separate peaks at the lowest grid size.

An alternative approach for reducing the computational complexity is suggested by Dmochowski et al. (2007b). The authors proposed a generalised SRP algorithm that reformulates the standard SRP approach by introducing an inverse mapping function that maps relative steering delays to candidate locations via a lookup table. The build up spatial spectrum is further optimised by evaluating only a defined range of delays that correspond to significant peaks in the cross-correlation function. Results suggest a comparable accuracy for SSL by less than 10% of the computational demand compared to the conventional search of the SRP approach.

### 2.5.2 High-resolution spectral estimation

The second category of sound source localisation algorithms are based on high-resolution spectral analysis techniques including: autoregressive modelling, minimum variance spectral estimation and eigenanalysis of the correlation matrix (e.g. MUSIC) (Brandstein and Ward, 2001; Dmochowski et al., 2007a). In regard to SSL for surveillance purposes, where speech is most commonly the target signal, these techniques are less suitable. The primary reason is that all approaches use a spectral correlation matrix which is not known beforehand and must be estimated by averaging the recorded data over a period of time where source and noise is stationary. But a speech signal can only be considered stationary for very short periods of time and therefore cannot be averaged. Another drawback is that high-resolution approaches are designed for narrowband source signals in the far-field region and linear array configuration. However, further research has shown that minimum variance and MUSIC techniques can be extended to near-field and general array geometries (Brandstein and Ward, 2001). For broadband source localisation these algorithms can be serialised (Asano et al., 2001) which, however, make these algorithms more computationally expensive.


2.5.3 Time delay of arrival

TDOA based strategies are also widely used. This technique uses a single or multiple pairs of spatially separated microphones to compute the time delay between the received signals. The delay is commonly estimated by maximising the cross-correlation function between a microphone pair for the possible time range defined by the spatial microphone separation (Knapp and Carter, 1976). For multiple microphone pairs, the source location can then be estimated in a second step. Based on the known microphone locations, hyperbolic curves are generated and their intersections mark the source location. However, the true delay cannot be computed and is only estimated. Hence, the intersection of the curves must be estimated in some optimal sense. Because of the early decision making when estimating the delay based on two microphones, this approach is less accurate in noisy and reverberant environments (Bechler and Kroschel, 2005). However, in environments with low reverberation effects and a high signal to noise ratio the source location can be computed faster and with less required microphones than SRP based techniques.

In work by Ma et al. (2006) a framework for localisation and tracking of unknown and varying numbers of speaker is proposed. Localisation is based on TDOA estimation between multiple microphone pairs. Experiments show that the localisation accuracy decreases as the room reverberation time increases, causing peaks in the cross-correlation function for wrong time delays. However, the authors proposed a Bayesian RFS (random finite set) filter to tackle the problem of multiple active speakers and hence all peaks of the cross-correlation function are evaluated. Therefore, this approach can adjust to false peaks introduced by reverberation effects. Experiments confirm that the approach was able to locate an active source by assuming the source is moving and reverberation effects are not stationary for long periods of time.

2.5.4 Energy based localisation

The last localisation category analyses the signal power measurement over a set of spatially separated microphones. The general idea is that the recorded source signal from a microphone that is closer to the source is stronger than that from a microphone further away. This characteristic can be mathematically described by the inverse square law (Howard and Angus, 2006) for free field environments, showing an exponential decay in signal power with distance. The advantage of this technique is that it is computationally inexpensive compared to all other methods and the microphone network can be arbitrarily distributed. Therefore, such approaches are often used in wireless networks. However, for
accurate source localisation generally more microphones are needed than for other SSL strategies. The reason is that the microphone coverage and hence the distance between the source and microphone determines the accuracy, especially for moving or closely separated targets. Additionally, wind noise or highly varying microphone gains can effect the localisation accuracy.

In Sheng and Hu (2005) a maximum likelihood (ML) estimation technique is presented for SSL using acoustic energy measurements from a wireless ad-hoc sensor network. The authors proposed a multi resolution search and an expectation-maximisation algorithm to solve the maximum likelihood problem in a computationally efficient way. Their approach is based on single time instance energy readings of different sensors evaluated by an energy attenuation model as a function of source to sensor distance. The approach is compared to non-linear least square, least square energy ratio approaches and nearest neighbour localisation, where the microphone with the highest energy reading is associated with the source location. Experiments on synthetic data with zero-mean Gaussian random noise and on real data of a moving military vehicle showed that the approach outperforms other compared methods.

The disadvantage of the maximum likelihood estimator is that it requires an iterative solution, increasing the computational complexity. Therefore, Meesookho et al. (2008) proposed a weighted least mean squares method for SSL that achieves similar results as ML with a reduction in complexity and increase in performance. This algorithm was tested on two measurements: Energy ratio and the direct measured energy, with similar results. Utilising the measured energy directly offers computational advantages compared to the energy ratio and therefore making it suitable for real-time applications. The limitation of this approach is that the presented least squares method only accounts for single source problems.

2.6 Speech enhancement

Speech enhancement techniques try to improve the perception or recognition of a degraded speech signal caused by noise. In real world situations, the major challenge is that the ambient noise is constantly changing in time and frequency. This section gives an overview of the most commonly applied techniques to improve a degraded speech signal. Multi and single-channel approaches are discussed. Widely used multi-channel techniques are beamforming algorithms that create an acoustic beam to enhance the sound of a spatially defined location and blind source separation that instantly separate several source sig-
nals by estimation of its mixing matrix. Spectral subtraction, a single-channel approach, estimates the noise conditions and removes it from the recorded mixture of speech and uncorrelated noise.

### 2.6.1 Beamforming

The term beamforming characterised all techniques that use a number of spatially separated microphones to generate a directed audio pattern to enhance a sound signal of a given direction. These algorithms use the physical principle that sound propagates through air with a known and constant speed, assuming the temperature is constant. Thus, spatially separated microphones record the audio of a sound source with a certain delay, defined by the relative source location and the distance between the microphones. In principle, the sound of a point source propagates to all directions as spherical waves. However, to simplify the beamforming algorithm, a common assumption is made that the source is in the far-field region, as shown in figure 2.4(a). Hence, the acoustic wave can be described as a planar wave and the steering direction can be described by a single steering delay that can be multiplied over the microphones depending on a reference microphone and the spatial distance.

![Figure 2.4: Sound wave propagation of a source in the near and far-field region.](image)

The general idea of beamforming can be best described by the most simplest beamforming algorithm, the delay and sum beamformer. In principle, the algorithm aligns the signals of all \( K \) microphones to a given direction before adding them. Hence, the aligned signal of the desired directions is amplified but signals from other direction are misaligned, and therefore cancel each other out. The algorithm is commonly performed in the frequency domain as (Benesty and Huang, 2003):

28
where $w_k$ is a weighting factor, $X$ the input signal and $\omega$ the frequency component. Hence, the output $Y$ of the DSB is determined by the direction that is given by the steering delays $\tau_k$ of microphone $k$.

Over the years many different beamforming techniques are proposed and surveyed in McCowan (2001); Benesty and Huang (2003); Christensen and Hald (2004). These algorithms can be categorised into fixed or conventional and adaptive approaches. Fixed beamforming algorithms primarily use information about the sensor locations and the source location by a set of fixed filter weights to enhance the sound of a desired direction, such as the DSB. In comparison, adaptive beamforming algorithms use additionally the signal characteristics of a desired source to improve the reduction of unwanted sources. Therefore, these algorithms use optimisation techniques such as minimum mean squared error or linearly constraint minimum variance (LCMV). A well known adaptive LCMV beamformer is the generalised sidelobe canceller (GSC) (Benesty and Huang, 2003). Figure 2.5 illustrates the different signal flow of a DSB and GSC beamformer.

The performance of the beamformer is not only determined by the chosen beamforming algorithm, but also by the array geometry. In work by Christensen and Hald (2004) an evaluation is conducted into regular array designs, such as grid or cross arrays, and irregular array designs, such as Archimedian spiral or optimised wheel array. The experiments were performed on large array sizes with about 65 microphones. The evaluation shows the major limitation of regular arrays is in terms of sidelobes (aliasing problem) for broadband signals, which is based on the regular and repeated spatial distance between the microphones. These sidelobes can be significantly reduced by using a irregular array design, where no spatial distance between two microphones is the same. However, finding the right irregular array geometry for maximising the performance for a particular condition is not trivial and the supporting structure of the microphones can be difficult to produce.

In summary, beamforming can produce good speech enhancement results for spatially separated sources. However, the main disadvantage compared to blind source separation or spectral subtraction approaches is that any beamforming algorithm needs information about the relative direction between the array and the source to enhance the signal.
2.6.2 Blind source separation

Blind source separation (BSS) techniques attempt to separate out different source signals from a linear mixture of convolutive audio signals. The term \textit{blind} refers to the fact that these techniques have no prior knowledge of the source signals, the mixing conditions or the sensor alignment (Brandstein and Ward, 2001), and hence can be seen as a black box (see figure 2.6). In principle, the algorithms try to estimate the mixing matrix $A$ of size $K \times N$ that describes the transformation function between the source signal $s(t) = (s_1(t), s_2(t), ..., s_N(t))$ of $N$ sources and the recorded audio $x(t) = (x_1(t), x_2(t), ..., x_K(t))$ of $K$ microphones. The mixing model for instantaneous scenario, a simplified case of the convolutive where all source signals arrive at the microphone at the same time, can be mathematically expressed as:

$$x(t) = As(t) + \eta(t)$$  \hspace{1cm} (2.2)

where $\eta$ is the environmental noise and $t$ is the discrete time index.
Various BSS algorithms are investigated over the years and surveyed in detail in Brandstein and Ward (2001); Hild (2003); Pedersen et al. (2007). Generally, these approaches can be categorised into time and frequency domain techniques that either use information theoretic approaches such as maximum likelihood or non-information theoretic approaches including second order statistics (SOS) or higher order statistics (HOS) (Hild, 2003). Depending on the demixing approach, the source signals need to be non-stationary and not correlated for SOS (Pedersen et al., 2007) or strictly statistically independent for HOS (Hyvärinen and Oja, 2000). Additionally, most techniques require that the number of sources is known and that the number of sources is less than or equal to the number of microphones (Brandstein and Ward, 2001; Pedersen et al., 2007).

While blind source separation has proven its effectiveness when certain assumptions are satisfied, there are a few challenges and issues still to be solved. These are non-stationary acoustic environments, varying number of sources and when the number of sources are higher than the number of microphones (Brandstein and Ward, 2001; Pedersen et al., 2007).

### 2.6.3 Spectral subtraction

A widely applied speech enhancement technique, using only single microphone, is spectral subtraction and was first presented by (Boll, 1979). The general idea of this approach is to estimate and subtract the noise that distorts the recorded speech signal in conditions where the additive noise signal is statistically independent of the speech signal. The noise estimation is usually computed heuristically (Gustafsson et al., 2001; Kamath and Loizou, 2002; Yang and Fu, 2005; Wójcicki et al., 2006; Breithaupt et al., 2007) by averaging out the recorded noise in noise only situations. Ephraim and Malah (1984) investigated an alternative technique by statistically modelling each background coefficient as Gaussian random variable.
As the name indicates, these approaches are performed on the spectral representation of the input signal. Past work has shown (Cohen and Berdugo, 2001; Ephraim and Malah, 1984; Wang and Lim, 1982) that the magnitude components are most distorted by the ambient noise in comparison to the phase. Thus, the subtraction operation is commonly applied only to the magnitude or power spectrum of a short time Fourier transformed signal, referred to as magnitude or power spectral subtraction respectively (Gustafsson et al., 2001). Essentially, a filter or commonly known as gain function $G$ is applied to the input spectrum $X$ of the windowed signal, with time block index $i$ as:

$$
\hat{S}(i, j) = G(i, j)X(i, j)
$$

where $j$ is the frequency index and $\hat{S}$ the estimated speech signal. The gain function is given as:

$$
G(i, j) = \left(1 - \nu_G \frac{P_{\eta}(i, j)}{P_x(i, j)}\right)^{1/a}
$$

where $P_{\eta}$ and $P_x$ is the magnitude spectrum of the noise and input signal respectively. $\nu_G$ is the subtraction factor that controls how strict the filter is and $a$ controls if either magnitude ($a = 1$) or power subtraction ($a = 2$) is used. The general signal flow of the spectral subtraction is shown in figure 2.7.

Figure 2.7: Signal flow for general spectral subtraction algorithm.
CHAPTER 2. BACKGROUND

The main disadvantage of the general spectral subtraction approach is that it assumes static noise conditions (Wójcicki et al., 2006). However, in real environments this is rarely true and even when the noise spectrum is modelled over time, it can never represent the true noise conditions accurately. During rapid noise type changes, the adaptation of the modelled noise requires time, reducing the effectiveness of the approach. During speech sequences, the updating of the noise model is suspended, leading to poor results.

Another common problem is that the remaining noise in the estimated speech signal is very unnatural, known as residual noise or “musical” noise (Cappe, 1994). The rate of the distortion depends on how strict the filter is set. Several approaches have been proposed to compensate for this distortion. The earliest and frequently applied one is the introduction of a spectral floor function proposed by Berouti et al. (1979). Recent approaches have focused on the use of different smoothing techniques (Wójcicki et al., 2006), a low resolution gain function (Yang and Fu, 2005), sub-band filtering and importance weighting (Kamath and Loizou, 2002) or improving the filter coefficients (Breithaupt et al., 2007).

Crucial to any kind of spectral subtraction algorithms is the detection of speech (Gustafsson et al., 2001). In particular, if speech is not detected well the noise estimation degrades, as speech components are falsely used in the update of the modelled noise. Additionally, the parameters of the filter are set depending on the presence of speech. Thus, the perception of the enhanced speech degrades when the parameters are set incorrectly.

2.7 Summary

This chapter presented the advantages and challenges for surveillance systems with particular focus on audio sensors and multi-modal event indexing. A review of relevant literature describing recently proposed surveillance systems and related state of the art tasks for acoustic surveillance was presented. This provides the background and motivation for the investigation into multi-modal event analysis and large scale speech enhancement technique of this thesis.

The chapter began with a review of general surveillance tasks and network architecture, including the strengths and shortcomings of single and multi-modal systems. In particular, the focus was directed on acoustic and visual sensors. Essential techniques for acoustic surveillance tasks followed, starting with details about the importance of selecting suitable acoustic features. Then an overview of commonly used classification strategies was presented. In particular, voice activity detection was discussed as it is a fundamental
part of many surveillance tasks such as speech enhancement. Acoustic source localisation
techniques were reviewed next, highlighting advantages and limitations between the dif-
ferent approaches in terms of computational load, robustness and accuracy. Finally, an
overview of single and multi-channel speech enhancing techniques was presented, because
enhancing a distorted speech signal can contribute greatly in analysing a scene of interest.
Differences between the approaches that use a spatial filtering and demixing algorithms
were described.
3.1 Introduction

In recent years, surveillance and monitoring systems have become increasingly sophisticated and complex with different sensors being deployed to enhance performance. Common choices of sensors include CCTV cameras, infrared sensors, simple switches for doors or windows, pressure sensors or microphones. Cameras are certainly the most used devices to monitor an area of interest because they are easy to install and can cover wide areas. They are predominantly used to identify and track persons of interest, as in (Chen and Rui, 2004; Vermaak et al., 2001; Lo et al., 2004; Beal et al., 2002). However, it is increasingly difficult to observe all the available cameras and therefore this chapter investigates the introduction of an audio array in a video surveillance domain to extract and index “coordinated” audio and video events.

The system presented in this chapter consists of a static CCTV camera and one linear microphone array with uniformly separated microphones that are positioned next to the camera. To make the configuration of the system user friendly and resilient to any drift in the relative alignment of the sensors, the relative orientation of the camera and the microphone array is not known in advance. Therefore, a non-linear calibration function that aligns the audio-video observations is learned. This dynamically estimated calibration function maps the two dimensional video coordinates to the one dimensional angle of arrival of a sound source. For evaluating the system in real environments, it is applied in the surveillance domain in which events are indexed by their video and enhanced audio information. The contribution of this chapter consists of:

- Dynamically learning a non-linear function that maps the one dimensional audio information to the two dimensional video information.
- The coordinated use of video and audio cues to capture and index surveillance events with multi-modal labels.
CHAPTER 3. AUDIO-VIDEO CALIBRATION

One possible scenario where such a system would increase the observed information content is when a gathering of people is observed, for example, when two people are having a conversation in German and at a different location two other people are having a conversation in English. The visual component of the system can then detect the event “gathering” of two or more people and index it with the beam-steered audio for that particular location. Thus, in the described scenario two visual events would be detected, “gathering 1” and “gathering 2” and are indexed with their corresponding beam steered audio. This allows the selection of either event “gathering 1” or “gathering 2” for listening to the enhanced German or English conversation respectively.

The remainder of this chapter is structured as follows: Section 3.2 describes the methodology for calibrating and indexing the audio-video events. The section is organised into sound source localisation, design parameters of the microphone array and sensor calibration. Experiments evaluating the performance of the system are depicted in section 3.3. Section 3.4 concludes the chapter by analysing the presented approach in terms of usability for real surveillance applications.

3.2 Methodology

The proposed surveillance system has two general stages: Learning the sensor alignment and using the knowledge to capture coordinated audio and video to index surveillance events. This section details the task for learning and indexing audio-video events. Firstly, the sound source localisation and the influence of the array construction on the source location estimation is discussed. The calibration of the different types of sensors that are finally used for computing the enhanced audio of a visual surveillance event is then discussed.

3.2.1 Angle of arrival

Sound source localisation relies on estimating the time delay of arrival between two or more spatially separated microphones to compute an angle of arrival (AOA). This AOA in our system is relative to the centre of the linear microphone array as illustrated in figure 3.1 and indicates the source direction. The array configuration is one dimensional with \( K \) uniformly and linearly separated microphones. Therefore, the estimated angle only indicates the source direction of a two dimensional plane where the array is located.
A common technique to estimate the relative time delay between a microphone pair $k = 1: K \in \mathbb{N}$ and $l = 1: K \in \mathbb{N}$ is the generalised cross correlation (GCC) proposed by Knapp and Carter (1976), where the estimated time delay $\hat{\tau}$ is computed by finding the highest peak in the cross correlation function $R_{kl}$. The cross correlation is computed as:

$$R_{kl}(\tau) = \int_{-\infty}^{\infty} x_k(t)x_l(t+\tau)dt$$ \hspace{1cm} (3.1)$$

where $x$ is the input signal, $t$ is the time index and $\tau$ is the inserted time delay. The cross correlation function is related to the cross power density function by the Fourier transformation relationship as:

$$R_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_k(\omega)X_l^*(\omega)e^{j\omega\tau}d\omega$$ \hspace{1cm} (3.2)$$

where $X$ is the input signal that is transformed into the frequency domain via FFT with $\omega$ as frequency component and $*$ is the complex conjugate. $j$ is the square root of -1. The generalised cross correlation is defined by filtering the input signals $X_k$ and $X_l$ to compensate for possible gain variations and weighting the frequency as:

$$R_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega)X_k(\omega)X_l^*(\omega)e^{j\omega\tau}d\omega$$ \hspace{1cm} (3.3)$$

Figure 3.1: Angle of arrival estimation for a linear and uniformly distributed microphone array for a sound source in the far-field region.
where $H$ denotes the general filter function as:

$$H(\omega) = H_k(\omega)H_l^*(\omega)$$ (3.4)

and $H_k$ and $H_l$ denote the filter of microphone $k$ and $l$. A common weighting function is the phase transform (Knapp and Carter, 1976) that weights the importance of the frequency as:

$$H(\omega) = \frac{1}{|X_k(\omega)X_l^*(\omega)|}$$ (3.5)

The estimated TDOA $\hat{\tau}_{kl}$ between microphone pair $k$ and $l$ is the time lag $\tau$ that maximises the generalised cross correlation function as:

$$\hat{\tau}_{kl} = \arg\max_{\tau} \{R_{kl}(\tau)\}$$ (3.6)

where $\tau$ ranges between $\pm \tau_{max}$ and corresponds to a scan-radius of $-\pi/2$ to $\pi/2$, relative to the centre of microphone pair. $\tau_{max}$ is defined by the distance $d$ between the microphones $k$ and $l$ as:

$$\tau_{max} = \frac{d_{kl}}{c}$$ (3.7)

where $c$ is the speed of sound ($\approx 340$ m/s at $15^\circ$C). Reverberation and other sound sources introduce many local maxima in the correlation function $R_{kl}$. Therefore, the weighting function $H$ is important to lower the amount and intensity of such unwanted maxima.

A disadvantage of the GCC method is that it is based on microphone pairs. Therefore, if the microphone array consists of more than two microphones, the AOA must be computed in two-stages. The first stage estimates a pairwise TDOA and the second stage generates hyperbolic curves based on microphone position and the estimated time lag. The sound location is then estimated by the intersection of these curves. Due to measurement precision, reverberations and environmental noise, the intersection must be estimated by an optimisation method such as maximum likelihood.
Similar to the generalised cross correlation approach is the steered response power approach, used for estimating the time delay. SRP is a one-stage approach that uses all \( K \) microphones at once. Therefore, it overcomes the disadvantage of making an early decision based on only one microphone pair. However, this advantage is achieved by an increase in computational complexity. In essence, this algorithm uses beamform techniques to compute a spatial filtered power response to scan an area of interest. The filter and sum beamformer is commonly used for such applied spatial filtering by summing over the filtered and time-alignment input signals. In the time domain, the steered response \( y \) at time instance \( t \) and given time delay \( \tau \) for each microphone \( k = 1 : K \) is computed as:

\[
y(t, \tau_1, \tau_2, ..., \tau_K) = \sum_{k=1}^{K} h_k(t) x_k(t - \tau_k) \tag{3.8}
\]

where \( h \) is the filter function and \( x \) is the input signal. This filter and sum beamformer can be expressed in the frequency domain as:

\[
Y(\omega, \tau_1, \tau_2, ..., \tau_K) = \sum_{k=1}^{K} H_k(\omega) X_k(\omega) e^{-j\omega \tau_k} \tag{3.9}
\]

where \( \omega \) is the frequency component. The steered response power is computed as:

\[
P(\tau_1, \tau_2, ..., \tau_K) = \int_{-\infty}^{\infty} Y(\omega, \tau_1, \tau_2, ..., \tau_K) Y^*(\omega, \tau_1, \tau_2, ..., \tau_K) d\omega \tag{3.10}
\]

By inserting equation 3.9 into equation 3.10 and rearranging it we get:

\[
P(\tau_1, \tau_2, ..., \tau_K) = \sum_{k=1}^{K} \sum_{l=1}^{K} \int_{-\infty}^{\infty} H(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega(\tau_l - \tau_k)} d\omega \tag{3.11}
\]

where \( H \) is expressed as in equation 3.4.

For the proposed system where the microphone alignment is linear and uniformly distributed, the general steered power response can be further simplified under the assumption that the location of the source is in the far-field region. As shown in figure 3.1 the
time delay between the microphone pairs are a multiple of the steering delay $\tau$ to align
the signals to direction $\alpha$ as:

$$P(\tau) = \sum_{k=1}^{K} \sum_{l=1}^{K} \int_{-\infty}^{\infty} H(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega\tau(l-k)} d\omega$$  \hspace{1cm} (3.12)$$

where $\tau$ is the introduced steering delay, $\omega$ is the frequency and $H$ is the frequency weighting function of microphone pair (k,l). $\ast$ denotes the complex conjugate. The weighting function $H$ is important to generate a meaningful steered power function result by normalising the frequency power. We choose the phase transform (PATH) weighting that is defined in equation 3.5. The estimated time delay $\hat{\tau}$ that indicates the source location is obtained as the time lag $\tau$ that maximises the steered response function $P(\tau)$ as:

$$\hat{\tau} = \arg\max_{\tau} P(\tau)$$  \hspace{1cm} (3.13)$$

where $\tau = -\tau_{max}: \tau_{max} \in \mathbb{R}$ corresponds to a scan-radius of $-\pi/2$ and $\pi/2$, relative to the centre of the array. $\tau_{max}$ is defined by the microphone spacing $d$ as:

$$\tau_{max} = \frac{d}{c}$$  \hspace{1cm} (3.14)$$

where $c$ is the speed of sound ($\approx 340$ m/s at 15 $^\circ$C). The estimated time delay $\hat{\tau}$ can then be used to compute the estimated angle of arrival $\hat{\alpha}$ as:

$$\hat{\alpha} = \arcsin \left( \frac{c}{df_s} \hat{\tau} \right)$$  \hspace{1cm} (3.15)$$

where $f_s$ is the sampling frequency. The estimated AOA has quite a large variance due to the reverberations and other ambient noise sources. To reduce the effects of reverberations, that randomly creates local maxima in the steered response distribution $P$, a smoothing function based on the entropy $\Theta$ is applied. The exponential smoothed steered response function $\overline{P}$ of time delay $\tau$ is computed as:
\[ P_i(\tau) = (1 - \frac{\gamma_s(\Theta)}{\nu_s})P_{i-1}(\tau) + \frac{\gamma_s(\Theta)}{\nu_s}P_i(\tau) \]  

(3.16)

where \( i \) is the time block index of computing the steered response distribution. \( \nu_s \) is a normalising factor and \( \gamma_s(\Theta) \) is the update factor that is computed as:

\[ \gamma_s(\Theta) = \frac{1}{1 + e^{(\Theta - \mu_s)/\sigma_s}} \]  

(3.17)

where \( \mu_s \) is the mean and \( \sigma_s \) is the variance. For a better understanding of \( \gamma_s \), two examples are given that describe the two extremes of the steered response distribution \( P \), where without loss of generality the parameter \( \mu_s \) is set to 0.5 and \( \sigma_s \) to 0.5. When \( P \) is uniformly distributed, \( \Theta = 1 \), \( \gamma_s \) is 0. In this scenario the updating is suspended because the function \( P(\tau) \) does not have any local maxima, that would indicate a possible sound source. Conversely, when the distribution \( P \) is not uniform (i.e. very peaked), \( \Theta = 0 \), then \( \gamma_s \) is 1. That scenario describes a perfect localisation that is not influenced by any reverberations or ambient noise. The rate of “fall” in that \( \gamma_s \) changes between 1 and 0 is controlled by \( \sigma_s \) and can be shifted in position by \( \mu_s \), as shown in figure 3.2(a). In fact, this smoothing function is similar to the well known logistic function (Mitchell, 1997).

![Update factor function \( \gamma_s \) for two different sets of parameters.](image1)

![AOA estimation based on \( P_i(\tau) \) and \( P_i(\tau) \).](image2)

Figure 3.2: Smoothing of the steered power response for AOA estimation. Source: Kühnapfel et al. © 2007 IEEE

### 3.2.2 Design parameter of the microphone array

The spatial distance between the microphones is a crucial factor because when the distance increases over a certain point the difference between sampled signals becomes indistinguishable. This is commonly referred as the spatial aliasing effect. A well known temporal
sampling theorem to avoid temporal aliasing effects is the Nyquist (Nyquist-Shannon) theorem (Oppenheim et al., 1999) that describes the minimum sampling frequency $f_s$ in samples per second that is required to sample a continuous signal unambiguously as:

$$f_s \geq 2f_{\text{max}}$$

where $f_{\text{max}}$ is the maximum frequency component of the signal. Related to the temporal sampling is the spatial sampling theorem (Teutsch, 2007) that states the maximal spatial distance $d_{\text{max}}$ between two microphones is given by:

$$d_{\text{max}} \geq \frac{\lambda_{\min}}{2}$$

where $\lambda_{\min}$ is the smallest wavelength of the signal that corresponds to the highest frequency component of the signal. This wavelength is defined by the used sampling frequency $f_s$ as:

$$\lambda_{\min} = \frac{c}{f_s}$$

where $c$ denotes the speed of sound. Figure 3.3 illustrates the importance of the spatial sampling theorem by plotting the beam patterns when the theorem is adhered to (b) and when the theorem is discarded (a). For the case when the spatial sampling theorem is discarded, spatial aliasing occurs that results in two large sidelobes towards $0^\circ$ and $180^\circ$, that significantly decreases the spatial filtering ability.

Not only does the distance between the microphones influence the beam shape but the number of microphones will impact also. In essence, when more microphones are used, the spatial distance between the outer microphones increases. This implies that the source signal is recorded with a higher time delay between the outer microphones, causing a greater shift when aligning the individual signals to a particular direction. Sources from other directions are consequently misaligned causing a reduction that is increased by the number of signals that are summed up. Especially for low frequencies, more microphones significantly increase the spatial filtering performance, as the wavelength $\lambda$ is larger for lower frequencies (equation 3.20). Figure 3.4 illustrates these effects.
CHAPTER 3. AUDIO-VIDEO CALIBRATION

Figure 3.3: Beam patterns for different microphone spacings $d$. Steering angle $\alpha$ is $90^\circ$ and microphone count $K$ is 3.

To summarise, when designing a linear microphone array for creating an audio beam the following points are fundamental for the shape of the directed pattern:

- The number of microphones $K$;
- The inter-microphone spacing $d$;
- The frequency $f$ for which the directed beam pattern is created.

### 3.2.3 Acoustic and vision calibration

Calibrating the CCTV camera with the linear microphone array requires a calibration object that is observed by both sensors. For this purpose we use a speaking person as the calibration object. Therefore, background subtraction (Stauffer et al., 2000) is used to detect the person in the video stream. This approach is suitable for our system because it is able to efficiently segment a person from the observed scene when the background and lighting conditions are constant. As the system is deployed in an indoor environment and the used camera is fixed, we can assume that the background and light conditions are static. Because, dynamic audio-video calibration is the focus of this work and not video segmentation, no further video segmentation algorithms are investigated as the restrictions for using the background subtraction algorithm are adhered to. Generally, foreground
objects are detected when the difference between the current frame $C$ and a learned image of the scene’s static background $B$ is larger than a set threshold $T_B$ as:

$$|C(x,y) - B(x,y)| \geq T_B$$

(3.21)

where $x$ and $y$ are the pixel coordinates of the 2D image plane.

The alignment of the video and audio observation is made by a calibration function that maps the video coordinates to the one dimensional audio information. When estimating such a calibration function, three restrictions apply: 1) During calibration only one cal-
ibration object is detected; 2) the detected object has to emit sound and 3) the sound must be directed towards the microphone array. As in our case where the calibration object is a person, the system must detect speech and that the person is facing the microphone array. Voice activity detection is realised by the well known supported vector machine (Gunn, 1998; Vapnik, 2000), that is trained with features of speech and ambient noise. We adapted the Mel Frequency Cepstral Coefficients (Davis and Mermelstein, 1980; Peeters, 2004; Theimer et al., 2008) of the audio signal as features. For detecting the frontal pose of the person’s face in respect to the camera, a cascade of boosted classifiers developed by Viola and Jones (2001) with improved haar-like features (Lienhart and Maydt, 2002) is used. The function \( \rho(t) \) at time \( t \) computes the likelihood that the person is looking towards the camera and ranges from \( 0.2 \leq \rho \leq 1.0 \) with \( \rho \in \mathbb{R} \). As pointed out in section 3.2.1, this function is similar to the logistic function that is additionally shifted between the defined range. This is essential because the face detection algorithm only detects frontal faces that are not heavily rotated or tilted. Therefore, the function \( \rho \) takes into account that the face is not always detected and that a person cannot significantly change his pose from one frame to another and is expressed as:

\[
\rho(t) = 0.8 \frac{1}{1 + e^{-(z(t) - t - \mu_\rho)/\sigma_\rho}} + 0.2
\]

where \( \mu_\rho \) and \( \sigma_\rho \) control the rate of “fall” and \( z(t) \) represents the last time a frontal face is detected. Therefore, the term \( (z(t) - t) \in \mathbb{N}_0 \) represents the time with no detected faces, where the time is measured in frame numbers. This term is 0 when a frontal face is detected at the current frame and increases by 1 for each following frame that does not have a detected face. Figure 3.5 illustrates such a function of \( \rho \) where the control parameters \( \mu_\rho \) and \( \sigma_\rho \) are set such that the likelihood that the person is facing the camera starts decreasing after about 10 frames from the last time a face is detected. For a frame rate of 15 frames a second, that would mean after about 0.6 seconds of not detecting a face, it is more likely that the person is turning away from the camera.

For estimating the calibration function, the recorded image of the camera is divided into \( U \) vertical (perpendicular to the x-axis of the image) and uniformly distributed bins. This discretisation is used to estimate a mean AOA \( \bar{\alpha} \) for each bin \( u \in 1 : U \) and is computed as:

\[
\bar{\alpha}_u = \frac{\sum_t \alpha(t) \rho(t)}{\sum_t \rho(t)} \quad (3.23)
\]
wherein \( u \) is selected by the location of the centre of the detected person and \( \rho \) is used as a weight that is normalised by the sum over all weights. For speech enhancement the visual based discretisation of the AOA in \( U \) would result in a large step width when steering an audio beam towards a target. Therefore, a polynomial function of degree three is estimated based on the computed mean AOAs to give a smoothed AOA estimation for each vertical image location. The Levenberg-Marquardt algorithm (Marquardt, 1963) is used to fit the calibration function to the estimated AOAs of all bins \( u \).

### 3.2.4 Audio-Video surveillance

The surveillance task of the system is to detect subjects or situations of interest and enhance possible audio resulting from such events. For example, when the distance between two subjects is within a defined spatial threshold, the visual event “gathering” is detected, suggesting a possible conversation. Then the dynamically learned calibration function is used to compute the angle between the microphone array and the visually estimated centre location of the subjects. This is used to subsequently steer a directed beam pattern towards the location for enhancing the audio of that area. The directed beam pattern is created by a filter and sum beamformer explained in equation 3.9. The filter function \( H \) is used to control the beam pattern and applies a scaling that by summing over all microphones the resulting signal adheres to the data range of the audio signal. Because the proposed system estimated the AOA \( \hat{\alpha} \) of the target location, the time lag \( \tau_k \) can be expressed as:
where $d$ is the inter-microphone distance, $c$ the speed of sound and $k = -(K - 1)/2$ : 

$$
\tau_k(\hat{\alpha}) = \frac{d}{c}k \sin(\hat{\alpha})
$$

(3.24)

Substituting $\tau$ in equation 3.9 by equation 3.24 and taking into account the the AOA is estimated based on the centre of the array, the spatially enhanced audio can be computed as:

$$
Y(\omega, \hat{\alpha}) = \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} H_k(\omega)X_k(\omega)e^{-j\omega \frac{d}{c}k \sin(\hat{\alpha})}
$$

(3.25)

### 3.3 Experiments

The proposed surveillance system is investigated in terms of calibration accuracy and speech enhancement. All experiments were conducted in a 5.7 × 6.8 × 2.7 metre (width, depth, height) room with air conditioning and various PC fan noise as ambient noise. The linear microphone array consists of seven omnidirectional electret condenser microphones (SPECO MLM-1) with a uniform spacing of $d = 4$ cm between them. Because speech is the desired sound signal for AOA estimation and enhancement, the audio was sampled at 8000 Hz. That sampling frequency corresponds to telephone quality and adheres to the spatial sampling theorem (equation 3.19). For the AOA estimation the SPR-PHAT algorithm used a frequency range of 500 to 3500 Hz because this range contains the most speech power (Silverman, 1987). The scanning range was ±90° in front of the array, divided into 90 steps. Next to the array was the fixed CCTV camera located so that it faces in approximately the same direction as the microphones to maximise the common observed area. The camera had a wide angle lens and recorded the scene with a resolution of 320 × 240 pixels, captured at 15 frames per second. Sound sources for experiments in section 3.3.1 to 3.3.3 are introduced by a loud speaker and are either white noise, that is further bandlimited for experiment in section 3.3.3, or a speech signal of a female who counts from one to ten.
3.3.1 Angle of arrival evaluation

This experiment evaluates the computed AOA in terms of mean estimation error and standard derivation. The source location is altered in steps of 15° between ±75° and at a distance of 1, 1.5, 2, 2.5 metres from the centre of the array, as shown in figure 3.6. During the recording only one sound source was present that was louder than the common environmental noise. Table 3.1 shows the signal to noise ratio for all sequences. White noise and speech are used as a sound source and are played back through a loud speaker. For each location, the source was stationary for about 20 seconds.

![Figure 3.6: Source location for AOA evaluation.](image)

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Source distance from the microphones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0m</td>
</tr>
<tr>
<td>White Noise 1</td>
<td>13.27</td>
</tr>
<tr>
<td>White Noise 2</td>
<td>7.18</td>
</tr>
<tr>
<td>Speech 1</td>
<td>7.48</td>
</tr>
<tr>
<td>Speech 2</td>
<td>3.50</td>
</tr>
</tbody>
</table>

Table 3.1: SNR (dB) for different signals and distances. The SNR is computed as an average over all directions of a fixed distance.

Figure 3.7 shows the mean AOA estimation for signal *Speech 1*. For simplicity, only *Speech 1* is used to illustrate the general characteristics because the results of the other source signals are similar and speech is the desired source signal. It can be seen that the estimation error, that is defined by the distance between estimation and ground truth is lowest when the source location is towards the centre (0°). Figure 3.8 shows the estimation error...
Figure 3.7: Mean AOA estimation for a single sound source with speech as the signal. The sound source location is changed in 15° steps between ±75°, for distances between 1 and 2.5 metres from the centre of the array. Plotted against the ground truth location, confirming the previous observation. However, this graph also shows that when the distance increases, the error increases as well.

Figure 3.8: Error of the estimated AOA.

The results of the other test signals are shown in table 3.2 as average estimation errors.
CHAPTER 3. AUDIO-VIDEO CALIBRATION

It indicates that the estimation error for white noise is generally lower. However, all test signals show the general characteristic that the error is increasing with larger distance.

<table>
<thead>
<tr>
<th>Source Type</th>
<th>1 - $1_{11}$</th>
<th>1.5 - $1.5_{11}$</th>
<th>2 - $2_{11}$</th>
<th>2.5 - $2.5_{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>White Noise 1</td>
<td>2.86</td>
<td>2.75</td>
<td>5.25</td>
<td>6.69</td>
</tr>
<tr>
<td>White Noise 2</td>
<td>2.93</td>
<td>2.84</td>
<td>2.27</td>
<td>3.61</td>
</tr>
<tr>
<td>Speech 1</td>
<td>2.16</td>
<td>4.05</td>
<td>6.37</td>
<td>7.98</td>
</tr>
<tr>
<td>Speech 2</td>
<td>2.96</td>
<td>2.68</td>
<td>6.77</td>
<td>10.67</td>
</tr>
</tbody>
</table>

Table 3.2: Average estimation error of sound source location.

Another factor that describes the accuracy of the estimated AOA is how much it varies around the mean angle. Figure 3.9 illustrates that the standard deviation of the AOA shows similar characteristics as the estimation error with increasing values for angles further from the centre and generally higher values for speech. The big difference between speech and white noise is expected as speech is a non-stationary signal, changing constantly in amplitude and frequency.

![Figure 3.9: Mean standard deviation of AOA estimation for different sound sources.](image)

3.3.2 Calibration accuracy

This experiment determines the accuracy of the calibration function that maps an image location to a direction of an acoustic source. For learning the calibration function, a target subject moves in front of the array, between 1.5 to 2 metre away. The person holds a loud speaker at shoulder height that is emitting either white noise or a speech signal. Three different walking patterns are used to collect the training data for the calibration function:
• **Walking Pattern 1** (Easy): The person stands stationary in the centre of each video bin facing the sensor setup for about 20 seconds at each location.

• **Walking Pattern 2** (Moderate): The person constantly walks around the field of view of the camera, but faces the camera within small angles.

• **Walking Pattern 3** (Difficult): The person walks freely around the field of view of the camera and turns in any direction. The recording duration for this walking pattern is about two to three times as much as for Pattern 2.

Table 3.3 shows the average SNR value for each walking pattern and for different source signals. As expected the SNR value is highest for Walking Pattern 1 because the sound source is directed towards the microphone. When the source moves, the SNR starts to decrease as the source is not pointing directly at the microphones. This fact is strongest for the Walking Pattern 3 because the sound source is sometimes not directed at the microphones at all, so that no directed speech path is recorded.

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Average SNR value (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Walking Pattern 1</td>
</tr>
<tr>
<td>White Noise 1</td>
<td>11.99</td>
</tr>
<tr>
<td>White Noise 2</td>
<td>7.06</td>
</tr>
<tr>
<td>Speech 1</td>
<td>6.12</td>
</tr>
<tr>
<td>Speech 2</td>
<td>3.83</td>
</tr>
<tr>
<td>Speech 3</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Table 3.3: Average SNR for calibrating the sensor setup.

Figure 3.10 illustrates the arrangement of the image bins that are used to estimate a mean AOA for mapping the calibration function. Such calibration functions for all three walking patterns with source type *Speech 1* are shown in figure 3.11. Because the characteristics are similar over the various source types, only speech as the source signal is shown. To evaluate the calibration functions, the error between the measured ground truth direction of the centre of each video bin and the corresponding value of the calibration function is computed and is shown in table 3.4. Basically, source signals with a higher SNR value are expected to have better accuracy because the signal is dominant over any reverberation effects or other background noises, especially for Walking Pattern 1. However, for this walking pattern *Speech 1* has a slightly higher error than *Speech 2*. The reason is that during data capture the target subject must be located in the centre of each video bin, which covers about 10°, but small deviations from the centre are unavoidable. For Walking Pattern 2 and Walking Pattern 3 the error is mainly the result of the indirect sound path. Therefore, a freely moving sound source in Walking Pattern 3 has the least amount of time where the signal is pointed to the microphones, and hence the highest error rate.
CHAPTER 3. AUDIO-VIDEO CALIBRATION

Figure 3.10: Illustration of the image partitioning into video bins, object detection and face detection. Video bins are separated by yellow vertical lines, foreground objects are marked by a red bounding box and detected faces are highlighted with a white bounding box. The bin that corresponds to the centre of the foreground object is emphasised in yellow.

Figure 3.11: Calibration functions for all walking patterns and Speech 1 as source signal.

The error of the calibration function is directly linked to the accuracy of the estimated AOA of each image bin. An indication of how well the mean AOA is estimated can be made by looking at the standard deviation. Figure 3.12 shows the standard deviation of a speech single used in figure 3.11. Basically, a small standard deviation allows a precise AOA estimation within a few samples. When the standard deviation increases, meaning the estimated AOAs are more spread around the mean angle, more AOA instances are needed to compute a precise mean. It also indicates that this mean may not be a good estimate of the true direction, as experiments have shown that when the standard deviation
CHAPTER 3. AUDIO-VIDEO CALIBRATION

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Walking Pattern 1</th>
<th>Walking Pattern 2</th>
<th>Walking Pattern 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>White noise 1</td>
<td>0.93</td>
<td>1.12</td>
<td>2.62</td>
</tr>
<tr>
<td>White noise 2</td>
<td>0.96</td>
<td>0.79</td>
<td>2.17</td>
</tr>
<tr>
<td>Speech 1</td>
<td>2.11</td>
<td>1.62</td>
<td>3.09</td>
</tr>
<tr>
<td>Speech 2</td>
<td>1.90</td>
<td>3.06</td>
<td>6.92</td>
</tr>
<tr>
<td>Speech 3</td>
<td>5.75</td>
<td>10.02</td>
<td>15.05</td>
</tr>
</tbody>
</table>

Table 3.4: Average error between ground truth AOA of the centre of each video bin and the corresponding calibration value.

As the sound intensity increases, the error increases as well. Table 3.5 shows that for a freely moving calibration target, the standard deviation is generally higher as the sound is not always directed toward the array and therefore results in wrongly estimated source locations.

![Graph](image)

Figure 3.12: Standard deviation of all three walking patterns with *Speech 1* as source signal.

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Average standard deviation (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking Pattern 1</td>
<td>1.80</td>
</tr>
<tr>
<td>Walking Pattern 2</td>
<td>2.47</td>
</tr>
<tr>
<td>Walking Pattern 3</td>
<td>3.36</td>
</tr>
</tbody>
</table>

Table 3.5: Average standard deviation for all walking patterns and source signals.

Overall, this section demonstrates that the system can be calibrated well for *Walking Patterns 1* and 2, with increasing error for *Walking Pattern 3*. The calibration with test signal *Speech 3* demonstrates that the error sharply increases when the signal intensity of the source signal becomes lower. An average error of about $5^\circ$ can be seen as acceptable, considering the purpose of the system and the beam pattern characteristics of lower frequencies (shown in figure 3.16).
3.3.3 Signal enhancement

The signal enhancement capability of the system is evaluated by introducing two band-limited white noise sources into the environment, at a distance of 1.8 metres to the microphones. For source localisation, the image position is used to estimate the direction of the source. During the experiment, different combinations of bandlimited white noise sources are used, as shown in table 3.6, with a bandwidth of 300 Hz. At the beginning, these noise sources are positioned at the far right and left side of the camera view and are moved in about 10° steps in the opposite direction to the other side, as shown in figure 3.13. The signal enhancement is evaluated in terms of removing one of the source signals by maintaining as strong as possible signal content of the other source. Additionally, the source direction estimated, based on the image location, is compared to the angle that generates the best possible signal intensity of the desired source.

![Figure 3.13: Experimental setup for signal enhancement evaluation. Source locations are denoted as •, microphones by ○ and the camera by □.](image)

Figure 3.13 shows the signal power distribution for two selected test sequences for a scan from −90° to 90°, with the ground truth locations at 46° and −40° for Source 1 and Source 2 respectively. As signal enhancement is the aim, the highest signal power represents the optimal target direction. Table 3.7 shows the error between the estimated source localisation based on the video information and the maximum in the scan of the signal power. For location estimation based on video information, three calibration functions from experiment 3.3.2 of Walking Pattern 1 are used, labelled as Calibration 1 to 3.
CHAPTER 3. AUDIO-VIDEO CALIBRATION

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Source 1</th>
<th>Source 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 1</td>
<td>1050</td>
<td>1550</td>
</tr>
<tr>
<td>Sequence 2</td>
<td>3050</td>
<td>3550</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>2050</td>
<td>2550</td>
</tr>
<tr>
<td>Sequence 4</td>
<td>1550</td>
<td>2550</td>
</tr>
<tr>
<td>Sequence 5</td>
<td>2550</td>
<td>3550</td>
</tr>
<tr>
<td>Sequence 6</td>
<td>550</td>
<td>3550</td>
</tr>
</tbody>
</table>

Table 3.6: Centre frequency of bandlimited white noise sources for all test sequences.

The audio signal for Calibration 1 to 3 is White Noise 1, Speech 1 and Speech 2 respectively. Figure 3.14(b) and results of Sequence 6 of table 3.7 confirm the statement of section 3.2.2 that for low frequencies the signal location cannot be estimated accurately because of the broad characteristics of the beam pattern. Hence, the final error estimation is based on the mean position of all source locations over all test sequences that maximise the signal power for frequencies greater than 900 Hz and is compared to the vision based source location. This error is 1.43°, 1.65° and 2.02° for calibration one, two and three respectively.

Figure 3.14: Normalised signal power of four different white noise sources for a scan between ±90°. Source 1 and Source 2 are located at 46° and −40° respectively. Calibration signal Speech 1 is used for computing the calibration function to estimate the source location, shown as a dashed line.

Table 3.8 shows the difference in signal power between the maximum of the signal power distribution and the steered signal based on the video image. The difference in degrees between the maximum signal power and the steering angle is rather large for low frequencies, however the signal power does not vary as much. The reason is the beam pattern characteristics shown in figure 3.16, illustrates that for low frequencies the pattern is rather broad and therefore the localisation of the source cannot be computed with high accuracy.
Table 3.7: Error between the angle that maximise the signal power and the angle that is computed based on the image location of the source.

<table>
<thead>
<tr>
<th>White Noise</th>
<th>Angle difference (degrees)</th>
<th>Calibration 1</th>
<th>Calibration 2</th>
<th>Calibration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 1</td>
<td>Source 1</td>
<td>5.55</td>
<td>4.02</td>
<td>4.30</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>3.17</td>
<td>2.61</td>
<td>2.70</td>
</tr>
<tr>
<td>Sequence 2</td>
<td>Source 1</td>
<td>2.21</td>
<td>2.70</td>
<td>2.90</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>1.41</td>
<td>1.84</td>
<td>2.20</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>Source 1</td>
<td>1.97</td>
<td>1.43</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>1.87</td>
<td>2.11</td>
<td>2.50</td>
</tr>
<tr>
<td>Sequence 4</td>
<td>Source 1</td>
<td>2.82</td>
<td>1.52</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>1.92</td>
<td>2.11</td>
<td>2.41</td>
</tr>
<tr>
<td>Sequence 5</td>
<td>Source 1</td>
<td>1.74</td>
<td>2.15</td>
<td>2.53</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>1.36</td>
<td>1.93</td>
<td>2.29</td>
</tr>
<tr>
<td>Sequence 6</td>
<td>Source 1</td>
<td>13.17</td>
<td>11.65</td>
<td>11.80</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>1.41</td>
<td>1.84</td>
<td>2.20</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>4.36</td>
<td>3.32</td>
<td>3.50</td>
</tr>
</tbody>
</table>

Table 3.8: Difference in signal power of the enhanced signal when steered based on the image location compared to the direction that maximises the signal.

<table>
<thead>
<tr>
<th>White Noise</th>
<th>Difference in signal power (dB)</th>
<th>Calibration 1</th>
<th>Calibration 2</th>
<th>Calibration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 1</td>
<td>Source 1</td>
<td>0.017</td>
<td>0.008</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>0.034</td>
<td>0.019</td>
<td>0.023</td>
</tr>
<tr>
<td>Sequence 2</td>
<td>Source 1</td>
<td>0.039</td>
<td>0.071</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>0.042</td>
<td>0.065</td>
<td>0.078</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>Source 1</td>
<td>0.023</td>
<td>0.033</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>0.017</td>
<td>0.013</td>
<td>0.019</td>
</tr>
<tr>
<td>Sequence 4</td>
<td>Source 1</td>
<td>0.027</td>
<td>0.037</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>0.017</td>
<td>0.008</td>
<td>0.012</td>
</tr>
<tr>
<td>Sequence 5</td>
<td>Source 1</td>
<td>0.039</td>
<td>0.072</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>0.027</td>
<td>0.036</td>
<td>0.043</td>
</tr>
<tr>
<td>Sequence 6</td>
<td>Source 1</td>
<td>0.042</td>
<td>0.074</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>0.069</td>
<td>0.055</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Although maintaining a strong signal of the target source is important, more significant for the improvement of the signal content, is how well an unwanted signal is removed. Two important factors that control how well an unwanted signal is suppressed are the frequency range and the spatial separation between the target and noise signals. Figure 3.15 shows the signal power of two unwanted sources against the distance between the target and unwanted source. For illustrating the difference between low and high frequency ranges, the centre frequencies of the unwanted sources are 550 Hz and 3550 Hz. In general, a perfectly suppressed source would be indicated by a signal power value of 0. As the
theoretic beam pattern in figure 3.16 shows, higher frequencies have a much narrower mainlobe. Consequently, higher frequencies can be suppressed better with lower spatial distances between a target and the unwanted source. Figure 3.15 shows this characteristic of the ideal beam pattern by a significantly higher reduction of the unwanted source with centre frequency 3550 Hz. It also indicates that the reduction of the signal power between the visual based source direction estimation and the location that maximises the target source is negligible, as the characteristics of the dashed and solid line are nearly identical. Over all six test sequences, and with speech as the calibration signal (Calibration 2), the average difference is 0.0038 dB.

Figure 3.15: Signal power of the unwanted source for source location that is computed with referred calibration function Calibration 2 (solid line) and by the location that maximises the signal power of the target source (dashed line). The signal power is normalised by the signal power of the original signal.

Table 3.9 shows the reduction of the signal power in dB for all test sequences in comparison to the separation of the sources in degrees. Because all three calibration functions used estimate the source direction within a difference of a few degrees, only results that are computed with Calibration 2 are shown, since the difference is negligible. At first, it is interesting to note that most sources are better removed at a spatial distance of 83°, than at the larger distance of 106°. The cause of this can be found again in the characteristic of the beam pattern, shown in figure 3.16. As the pattern illustrates, the signal power does not linearly reduce when the distance from the target direction increases. Basically, it decreases sharply first depending on the frequency and then has many local minima and maxima created by the sidelobes. Hence, the reduction of an unwanted source is dependent on the proximity of the source location to a minima or maxima of the produced beam pattern.
TABLE 3.9: Signal power reduction (dB) of the unwanted source in comparison to the spatial separation in angle between source locations. Source locations are visually estimated using Calibration 2.

![Figure 3.16: Theoretic beam pattern of seven linear aligned microphones with a uniformly distance of 4 cm.](image-url)
3.3.4 Indexing audio-video events

This experiment evaluates the proposed surveillance system for two real world scenarios in terms of speech enhancement of an event of interest. In the first scenario two stationary groups, consisting of two persons each, are detected at different locations. For the second scenario, two persons are engaged in a conversation while moving. The relative distance of the subjects from the microphones ranges between 1.5 metres and 2.5 metres.

Figure 3.17 illustrates the group location for the first scenario. The two persons of the visual detected event Group # 0 have a conversation in English, while Group # 1 converses in German. Two enhanced audio streams were produced that are indexed to the corresponding event by steering an audio beam towards the centre of each event. These enhanced audio streams are then compared to the recording of a single microphone, after all audio streams are normalised to the same signal intensity. For evaluation, the mean opinion score (MOS) is used, rating the speech quality as a number ranging from 1 (Bad) to 5 (Excellent), defined by the ITU-T (1996) standard. This metric is widely used for subjective quality assessment of various forms of voice communication, where it is commonly used to rate the user satisfaction of VoIP communications, conferencing calls or audio codecs. For this experiment seven test subjects were asked to grade the speech quality in terms of understanding by removing the undesired noise and background conversation. Table 3.10 shows the average score of this survey. It clearly demonstrates that the enhanced signals are better perceived by the test subjects than the original audio.

Figure 3.17: Illustration of two visually detected events. Group # 0 is conversing in English and Group # 1 in German. Kühnapfel et al. (2007) © 2007 IEEE

For the second test scenario, a radio is placed in the far right corner of the room to increase the ambient noise. In this experiment the system proves the ability to follow a moving target using video-based tracking by adjusting the audio beam constantly to
CHAPTER 3. AUDIO-VIDEO CALIBRATION

<table>
<thead>
<tr>
<th>Event</th>
<th>MOS Steered</th>
<th>MOS Orig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 0</td>
<td>4</td>
<td>1.7</td>
</tr>
<tr>
<td>Group 1</td>
<td>3.75</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 3.10: Mean opinion score for the first test scenario between the original and the beam steered audio.

the new position of the detected event. Figure 3.18 shows six instances of this scenario where the event Group ≠ 0 is first detected at image (b). From there on the two subjects move in front of the camera from the right to the left side, and back to the middle. The resulting enhanced audio stream was again evaluated by surveying seven subjects using MOS, comparing it to the audio of a single microphone. Again, the MOS for the enhanced audio was higher than for the original, achieving a score of 3.9 compared to 2.

Figure 3.18: Illustration of the second test sequence, where two people are getting together and the visual event Group ≠ 0 is detected. Over time the persons walking around the field of view of the camera.

3.4 Conclusion

The introduction of the microphone array into the video surveillance domain has shown to improve the observed information content of the scene by extracting “coordinated” audio-video events. Such events, like “gathering” of people, are automatically detected
by the CCTV camera and then indexed with the corresponding enhanced audio of that particular direction. This hybrid form of surveillance system is more sophisticated than the usual CCTV systems because it can indicate if the observed area is of interest and thereby giving the operator the ability to assess the situation based on both enhanced audio and video information.

For any multi sensor system it is necessary to know either the relative sensor alignment or be able to learn the calibration function that maps the sensor information to a common space. In this chapter we did not constrain the sensor alignment by pre-defining the relative sensor orientation, to make the system as flexible as possible. Hence, the system needs training data to estimate a calibration function that maps a sound direction to an image location. Therefore, the proposed system uses instances where only one sound producing foreground target is detected. Experiments have shown that the estimated AOA used for computing the calibration function is quite accurate for angles around the centre of the array, with increasing errors for larger angles, and the distance of the source locations to the array. As this system is mainly used for monitoring people, speech is the sound type used for calibration. Experiments show that when using speech as the calibration signal, more data is needed to estimate an accurate source direction because the variance of the estimated AOA is much larger than for stationary, artificial sound types.

This investigation has shown the usefulness of combining video and audio for surveillance tasks. However, experiments have demonstrated that the error of the estimated AOA depends on the distance and angle of the sound source location to the microphone array. Such a restriction is a serious disadvantage when a large area is to be monitored. A solution could be to install several microphone arrays to cover a larger area but that would significantly increase the cost. Additionally, the quality of the enhanced sound signal depends on the noise location. As beamforming is used to enhance the audio, only a spatially separated noise source, be it a conversation in a different language or a radio, can be filtered out.
Chapter 4

Distributed microphone network

4.1 Introduction

Chapter 3 investigated the use of beamforming for enhancing a sound source by separating it out from a mixture of several sound sources. Such algorithms are often used in situations where the source is rather close to the microphone arrangement. The reason for the range restriction is that sound intensity decreases as distance increases. Such restrictions may not be a problem where the target location is confined to a location within a short range to the microphone array. Examples are telephone conferencing or hands-free phone systems in cars. Here, array solutions can enhance a speech signal by separating it from an uncorrelated and spatially separated noise signal. When it comes to outdoor situations, and consequently large areas, such restrictions in range are a major disadvantage. Compensating such a problem can be achieved by installing several microphone arrays, but that becomes very complex and costly making such a solution unrealistic for many acoustic surveillance tasks. Hence, this chapter investigates the use of a distributed microphone network for capturing and enhancing speech in large areas.

An acoustic signal, such as speech, decreases in signal strength when the distance from the source increases. The relationship between the sound intensity (sound intensity per square metre) of an omni-directional sound source and distance from the source in free field environments is described by the inverse square law. Figure 4.1 shows that by doubling the distance, the sound intensity drops to 1/4 because the area at that distance increases four times while the sound intensity is constant. Mathematically, the law is described by the relation of sound intensity $I$ and distance $d$ as:

$$\frac{I_1}{I_2} = \left(\frac{d_2}{d_1}\right)^2$$

(4.1)
CHAPTER 4. DISTRIBUTED MICROPHONE NETWORK

Figure 4.1: Inverse square law $1/d^2$. $S$ is the origin of an omnidirectional source and $d$ is the distance from the source. Red points out a rectangle with the same surface area of the sphere at different distances from the source.

$$I_2 = I_1 \left(\frac{d_1}{d_2}\right)^2$$  

(4.2)

where the subscripts indicate location 1 and 2 of the measurements. Equation 4.2 confirms that the sound intensity decreases to $1/4$ from its original intensity with respect to the area when $d_2$ is twice as large as $d_1$. Equation 4.2 is graphically illustrated in figure 4.2.

The sound intensity level (SIL), also called sound power level, is measured in watt per square metre ($W/m^2$) and generally shown in dB. This measurement is normalised by the lowest sound intensity level $I_0$ (0 dB) that the average human ear can pick up, at a frequency of 1 kHz. The dB level of a sound intensity measurement is computed as:

$$SIL = 10 \log \frac{I}{I_0}$$

(4.3)

with $I_0$ as $10^{-12} W/m^2$. 1 dB is in generally the smallest change in sound power level that the average human can detect. By doubling the distance between two measurements,
the sound power will decrease to a quarter and that is equivalent to a drop of the signal intensity level of about 6 dB (Howard and Angus, 2006).

For real environments, this law is only a generalisation because sound can be reflected and then added to the direct sound path, which increases the sound power. Also, obstruction between the measurement point and the source can lower the estimated sound power. However, this law proves why it is more economical to use distributed microphones for audio surveillance of large areas, rather than several microphone arrays.

With the distributed network of microphones we would like to explore the possibility of continuously capturing the audio signal of a moving subject in the surveillance space, using the concept of a “virtual” microphone. Figure 4.3 illustrates the principle of such a “virtual” microphone. Generally, at any given time a microphone is selectively chosen out of the network to create a new audio stream that is linked to a subject so that the created audio stream consists of the best direct speech path. Hence, the subject can move throughout the area with the resultant audio stream having the best possible signal to noise ratio of the raw and unprocessed audio of the speaking subject.

However, creating an audio stream with reference to a subject does not remove unwanted ambient noise, which will make speech perception difficult. Therefore, spectral subtraction is investigated to reduce ambient noise for enhancing the speech signal. This technique is suitable for our system because it requires only the input of a single microphone. The spectral subtraction approach used is magnitude spectral subtraction.

Spectral subtraction works very well for static noise types (Wójcicki et al., 2006) because
such noise types are constant over time and frequency. Synthetically generated noise is generally stationary: Examples are white noise, pink noise or brown noise. Unfortunately, real environments have rather non-stationary noise characteristics that change over time and frequency. Also noise types change rapidly from one type to another. Examples of such noise types are car noise, construction noise or music. In between non-stationary and stationary noise types is quasi-stationary noise. This noise changes over time and frequency as well but the variance is not as much as for non-stationary noise, for example, during a party or at a café where several people are talking at the same time, creating a mixture of different voices, often called “babble” or “cocktail” noise. The changes for this noise type come from the varying numbers of people talking but the overall noise characteristic stays similar. The reason why the general subtraction approach has limitations when the noise type rapidly changes or the noise is non-stationary, is that it uses only one noise model and updates it over time and frequency during noise only situations. Therefore, the estimation error increases when the noise characteristic constantly changes.
The aim of this chapter is to provide acoustic surveillance in realistic environments for large areas. In such environments the noise type is non-stationary and can change quickly. For example, consider an outdoor café located next to a street. The usual ambient noise is a mixture of several people talking until a car drives by. The car would then become the dominant noise type because car noise is generally more intense. Therefore, the general subtraction approach is modified to compensate for such situations.

The contribution of the work presented in this chapter is as follows:

- The generation of an audio stream that is connected to a subject via a “virtual” microphone. This allows the person to move freely around the network, maintaining the best possible audio capture at any time instance.

- Improving the spectral subtraction approach by modelling a separate noise estimation for each known ambient noise source to quickly adapt to rapidly changing noise types. Such a behaviour implies that the background noise has to be classified first.

- Improved voice activity detection that not only detects speech sequences but identifies within the sequence elements of high and low speech power, because speech consists of gaps where humans breathe or take short breaks to separate words. Thus, low speech energy elements can be used to update the noise model.

The layout of the chapter is as follows: Section 4.2 describes the methodology for creating the “virtual” microphone and enhancing the speech. The speech enhancing part is separated in noise classification, voice activity detection and the spectral subtraction approach. Experiments in section 4.3 demonstrate the effectiveness of the methodology. Section 4.4 concludes the chapter by reflecting on the results of the proposed system.

4.2 Methodology

This section explains the methodology of using the distributed network for acoustic surveillance, creating a “virtual” microphone and enhancing speech by removing the ambient noise.
CHAPTER 4. DISTRIBUTED MICROPHONE NETWORK

### 4.2.1 Generating a “virtual” microphone

Creating the proposed “virtual” microphone, the system must be able to detect microphone \( k' \) with the best direct speech path to the target. This is achieved by computing a weighting factor between all \( K \) microphones similar to Grbić and Nordholm (2002). This weighting factor is based on the power spectral density because the microphone with the best direct path has a higher signal to noise ratio compared to other microphones. The ambient noise may not have the same intensity over all microphones and thus the input signal of each microphone is normalised to level out the recorded ambient noise over all microphones. Hence, all microphones have similar importance when only background noise is present.

For computing the weighting factor, the input signal \( x_k(t) \) at microphone \( k \) is normalised and windowed with a Hamming window of 512 samples. The windowed signal \( x_k'(i) \) at time block index \( i \) is then transformed into the frequency domain \( X_k(i,j) : j = 1 : J \) via FFT, with \( j \) as frequency index. The cross power spectrum \( \Gamma \in R^{K \times K} \) at time block index \( i \), and over a frequency range is computed as:

\[
\Gamma(i) = \sum_{j=j_{\text{min}}}^{j_{\text{max}}} \tilde{X}^T(i,j)\tilde{X}(i,j) \tag{4.4}
\]

where \( T \) is the transpose and \( \tilde{X}(i,j) = [X_k(i,j) : k = 1 : K] \). Further, the cross power spectrum is only computed over the certain frequency range defined by \( j_{\text{min}} \) and \( j_{\text{max}} \). This is essential, because we are interested in detecting the best direct speech path and speech has a general frequency range of 300 to 3500 Hz. \( \Gamma \) is then used to compute the eigenvectors \( V \in R^{K \times K} \) via PCA. The eigenvector \( V' \) represents the eigenvector with the highest corresponding eigenvalue and is used as a \( 1 \times K \) dimensional weighting vector \( w \) as:

\[
w(i) = \nu_w \| V'(i) \| \tag{4.5}
\]

where \( \nu_w \) is a normalising factor. This weighting vector can be used to rate the importance of each microphone in terms of the best direct speech path. For creating the best possible audio stream of a single freely moving target, the recorded audio information of the highest ranked microphone \( k' \) at time block index \( i \) can be dynamically selected as:
\[ k'(i) = \arg\max_k \{ w(i, k) \} \] (4.6)

Such an approach works well for one target but has the disadvantage that a particular
target is not linked to a spatial location where the microphone is positioned. Therefore,
when multiple targets are present at different locations, multiple microphones will have
a higher weighting factor, and it is not possible to associate one particular target to the
corresponding microphone, especially if one or more targets are moving. Extending such
a system with a CCTV camera system could solve this problem by using visual tracking
techniques to identify the target location (Lim et al., 2007) and map it to the best possible
microphone. Such a mapping function can be learned by using sequences where only
one person is present and mapping its location with the highest ranked microphone \( k' \).
Because audio surveillance of larger areas of interest is the scope of this work, we will no
longer investigate any visual tracking approaches but will focus our attention on speech
enhancement.

4.2.2 Speech enhancement

The entire speech enhancement approach consists of three distinct parts: Noise classifi-
cation, voice activity detection and speech enhancement by spectral subtraction. As the
proposed magnitude spectral subtraction method uses \( N \) noise models, noise classification
is essential. Therefore, the entire network is used to estimate the general noise source in
order to overcome a possible false classification of a single microphone. The next stage is
voice activity detection, which is a prerequisite for speech enhancement. The VAD result
is computed by the signal power estimation in combination with the noise classification
result of the microphone of interest. Finally the spectral subtraction approach utilises the
noise classification result to select the corresponding noise model and the VAD result to
set the parameters for updating the noise model and enhancement of the speech signal.

4.2.2.1 Noise classification

For noise classification, the audio signal \( x_k \) at microphone \( k \) is split into the windowed
signal \( x'_k(i) \) with a window size of 512 samples (\( i = \) time block index). FFT is then applied
to \( x'_k(i) \) to transform the signal into the frequency domain \( X_k(i, j) : j = 1 : J \), with \( J \) as
frequency index count. The final feature vector \( D_k(i) \in \mathbb{R}^{1 \times M} \) is computed by correlating
CHAPTER 4. DISTRIBUTED MICROPHONE NETWORK

a triangular Mel scale filter bank with $X_k$, where $M$ denotes the number of filters and $M << J$. This filter bank computes a non-linear frequency representation for frequency components which was proposed by Steven et al. (1937).

For learning the characteristics of the known ambient noise sources, training data $D^n_k$ for each noise type $n$ is used to compute the mean feature vector $\overline{D}_k^n$ for microphone $k$ as:

$$\overline{D}_k^n = \frac{1}{\delta + 1} \sum_{i}^{i+\delta} D^n_k(i)$$  (4.7)

where $\delta$ is the number of time instances used. The first step for classifying the global background noise type is to classify the noise type for each microphone separately. Therefore, the normalised cross correlation coefficient $R_{kn}$ between the current features $D_k(i)$ and each noise model $\overline{D}_k^n$ is computed. The classification decision $n'_k(i)$ for the estimated background noise type at microphone $k$ and time block $i$ is computed as:

$$n'_k(i) = \arg\max_n \{ R_{kn}(i) \}$$  (4.8)

The correlation value $R_{kn}(i)$ of the classified noise type $n'_k(i)$ can also be used to detect if only background noise is present for time block $i$. For example, if the observation $D_k$ consists only of background noise, the correlation value with respect to noise model $\overline{D}_k^n$ will be close to 1. When speech and noise is mixed, the correlation values drops significantly for any noise model, depending on the strength of the speech component. If the background noise varies slightly over time, these characteristics are still valid. However, to compensate for changing ambient noise, the noise model $\overline{D}_k^n$ of the classified noise type $n'_k$ will be constantly updated. Such updating is performed when the correlation value is strong (e.g. above 0.95). The update is computed as:

$$\overline{D}_k^n(i) = (1 - \gamma_D) \overline{D}_k^n(i-1) + \gamma_D D_k(i)$$  (4.9)

where $\gamma_D$ is an exponential updating factor.

The final noise classification decision $n^*$ is based on the entire microphone network. Therefore, all microphones that do not contain any speech instances are used to compute a count
4.2.2.2 Voice activity detection

Voice activity detection utilises two measurements, the signal to noise ratio and the result of the correlation $R_{kn}$ between the noise model $n$ and the input signal at microphone $k$. Figure 4.4 shows the system flow of the proposed voice activity detection.

![System flow of the voice activity detection.](image)

We use the signal to noise ratio to estimate speech sequences. The general idea is that the signal power $P_k$ at microphone $k$ is higher when speech is present. Therefore, the average noise power $\overline{P}_k$ is estimated over time and all instances that are higher than a multiple of the estimated noise power are labelled as speech. $C^P$ is the result of the signal power based speech detection result at microphone $k$ and time block $i$ as:

$$C^P_k(i) = \begin{cases} 1, & \text{if } P_k(i) \geq T^P_k(i) \\ 0, & \text{if } P_k(i) < T^P_k(i) \end{cases}$$  \hspace{1cm} (4.11)

with $T^P$ as noise level threshold at microphone $k$ and speech is labelled 1. The noise level threshold must be dynamically set because the background noise level can change over time. Therefore, $T^P_k$ is computed as:
where \( \nu_p \) is a scaling factor to compensate for smaller variances in \( P_k \). Unfortunately, VAD based on signal power is quite sensitive to changes in background noise (Wu and Wang, 2006), because changes in signal power of non-stationary ambient noise result in falsely classified speech. Therefore, the correlation value \( R_{kn'} \) between the classified noise model of noise type \( n' \) and the extracted features at microphone \( k \) is utilised to estimate speech. \( C^R \) indicates this at microphone \( k \) as:

\[
C^R_k(i) = \begin{cases} 
1, & \text{if } R_{kn'}(i) < T_C \\
0, & \text{if } R_{kn'}(i) \geq T_C 
\end{cases}
\]  

\[ (4.13) \]

where \( T_C \) is a threshold where speech is labelled as 1. The threshold value for \( T_C \) is constant over all microphones and time because the minimum accuracy of the classification value is always the same. The combination \( o_k \) of both the previously described VAD measurements at microphone \( k \) and time block index \( i \) is computed as:

\[
o_k(i) = C^P_k(i) + C^R_k(i)
\]

\[ (4.14) \]

For improving the speech classification result and to make it more resistant to false classification, \( o_k(i) \) is smoothed out as:

\[
\bar{o}_k(i) = \frac{1}{\delta + 1} \sum_{\epsilon=i-\frac{\delta}{2}}^{i+\frac{\delta}{2}} o_k(\epsilon)
\]

\[ (4.15) \]

where \( \delta \) is the smoothing window.

The voice activity detection process not only labels the input data as speech or non-speech, but also identifies within a speech sequence occurrences with low (\( \psi_{\text{low}} \)) and high (\( \psi_{\text{high}} \)) speech power. Such a differentiation between high and low speech power is useful because one can capitalise on the fact that updating of the noise model of the speech enhancement approach can take place between suitable speech breaks. Therefore, the speech label \( v_k \)
at microphone $k$ is defined as:

$$
\psi_k(i) = \begin{cases} 
\psi_{no}, & \text{if } \sigma_k(i) < T_v \\
\psi_{low}, & \text{if } \sigma_k(i) \geq T_v \& C^P_k(i) = 0 \\
\psi_{high}, & \text{if } \sigma_k(i) \geq T_v \& C^P_k(i) = 1
\end{cases}
$$

(4.16)

where $T_v$ is a threshold.

### 4.2.2.3 Spectral subtraction

For enhancing the speech signal, magnitude spectral subtraction is applied to remove the background noise. Past work has shown (Cohen and Berdugo, 2001; Ephraim and Malah, 1984; Wang and Lim, 1982) that the magnitude components are mostly distorted by the ambient noise in comparison to the phase. Generally, if a speech signal $s$ is corrupted by an uncorrelated noise signal $\eta$, the recorded signal $x$ at microphone $k$ is an additive mixture of both signals as:

$$
x_k(t) = s(t) + \eta(t)
$$

(4.17)

where $t$ is the time index. To remove the noise from the mixed signal a time varying filter $g$ is applied on the recorded signal. For the time domain this would be realised by a convolution operation as:

$$
\hat{s}(t) = g_k(t) \ast x_k(t)
$$

(4.18)

where $\ast$ is the convolution sum and $\hat{s}$ is the estimated speech signal. Applying this formula in the frequency domain, the convolution operation is replaced by a multiplication operation. Hence, the windowed input signal $x'_k(i)$ at time block $i$ and window size of 512 samples is transformed into the frequency domain $X_k(i,j) : j = 1 : J$ via FFT, with $j$ as frequency index. The estimated speech signal $\hat{S}$ is computed by applying the gain function $G$ to the short-term frequency spectrum of $X_k$ as:
\( \hat{S}(i, j) = G_k(i, j)X_k(i, j) \) \hspace{1cm} (4.19)

where \( i \) is the time block index. The gain function is defined by the magnitude spectrum \( \mathcal{P} \) of the noise magnitude spectrum \( \mathcal{P}_{\eta_k} \) and the input signal magnitude spectrum \( \mathcal{P}_k \) at microphone \( k \) as in Gustafsson et al. (2001):

\[ G_k(i, j) = 1 - \nu_G \frac{\mathcal{P}_{\eta_k}(i, j)}{\mathcal{P}_k(i, j)} \] \hspace{1cm} (4.20)

with \( \nu_G \) as subtraction factor. Depending on the chosen parameter \( \nu_G \), such a filter introduces a distortion of the speech signal, known as residual noise or “musical” noise. This occurs when \( \nu_G \) is larger than 1 and hence some speech frequencies are removed by increasing the estimated noise. To reduce this speech distortion, a spectral floor function \( \beta \) is introduced by Berouti et al. (1979) as:

\[ G_k(i, j) = \max \left\{ 1 - \nu_G \frac{\mathcal{P}_{\eta_k}(i, j)}{\mathcal{P}_k(i, j)}, \beta \right\} \] \hspace{1cm} (4.21)

The parameters \( \nu_G \) and \( \beta \) of the gain function are set depending on the voice activity detection result \( v_k \) (equation 4.16). For noise only sequences, \( v_k = \psi_{no} \), the parameter can be set rather strictly to remove noise components by about 10 to 20 dB. During detected speech sequences, \( v_k = \psi_{high} \) and \( v_k = \psi_{low} \), the gain function is set less strictly to minimise the distortion of the speech signal. By adaptively updating the gain function over time, the quality of the estimated speech signal is further improved, because the variance of the estimated noise spectrum to the real noise spectrum is reduced as:

\[ \mathcal{P}_{\eta_k}(i, j) = (1 - \gamma_{\eta}) \mathcal{P}_{\eta_k}(i - 1, j) + \gamma_{\eta} \mathcal{P}_k(i, j) \] \hspace{1cm} (4.22)

where \( \gamma_{\eta} \) is the exponential updating factor. The value of the updating factor depends on the voice activity detection \( v_k \). For noise only sequences, \( v_k = \psi_{no} \), \( \gamma_{\eta} \) is set to a shorter averaging time because a clean noise signal is recorded. During speech sequences, the updating is suspended for \( v_k = \psi_{high} \) or set to a rather long updating time for \( v_k = \psi_{low} \).

A disadvantage of this speech enhancement approach is that during speech sequences the
gain function is not updated, \( v_k = \psi_{\text{high}} \), or only marginally for \( v_k = \psi_{\text{low}} \). The reason is that the noise signal is mixed with speech and updating the modelled noise would introduce errors. Especially when the noise type changes during speech, this causes a big variance between the estimated spectrum of the noise \( P_{n_k} \) and the real noise. Hence, the proposed approach is modified by modelling all known noise sources. Thus, if the noise source changes rapidly, the noise model is instantly switched instead of updating it slowly until it matches the current ambient noise characteristics. This technique reduces the variance of the modelled noise spectrum significantly. Therefore, the new gain function is computed as:

\[
G_k(i,j) = \max \left\{ 1 - \nu \frac{P_{n_k}^{n^*}(i,j)}{P_k(i,j)}, \beta \right\}
\]

(4.23)

where \( n^* \) is the classified noise source as in equation 4.10.

### 4.3 Experiments

We evaluate the two proposed techniques of creating a “virtual” microphone and enhancing a speech signal in two separate sets of experiments. Section 4.3.1 explores the ranking approach for the “virtual” microphone in terms of ground truth location and numbers of active targets. The evaluation is based on the acoustic and spatial importance of a microphone. In section 4.3.2 the speech enhancement approach is assessed by a range of experiments. As the proposed speech enhancement approach relies on noise classification and speech activity detection, these components are evaluated first. Experiments in section 4.3.2.3 numerically examine the advantage of utilising multiple noise models when the ambient noise type is changing during a speech sequence. These results are confirmed by a qualitative evaluation shown in experiment 4.3.2.4.

#### 4.3.1 “Virtual” microphone - Experimental setup

The first set of experiments investigates the proposed generation of a “virtual” microphone based on a distributed microphone network. Data for these experiments is collected in a 5.7 × 6.8 × 2.7 metre (width, depth, height) reverberant room. The usual background noise in this room comes from the air conditioning ducts as well as ventilation noises from
several PCs in the room. To simulate a more realistic environment and to increase the ambient noise level, café noise is played back into the room by two loudspeakers. This noise was previously recorded at a busy café during lunch time and consists of people having conversations and occasional cutlery noise. Audio was captured by five arbitrarily distributed omnidirectional condenser microphones (Behringer ECM 8000) with a sampling rate of 8 kHz and a resolution of 16 bits per sample for audio. Such a rather low sampling rate is chosen because the target signal is speech and the energy of speech is mainly in the lower frequency ranges (Silverman, 1987; Baken and Daniloff, 1991). Video that is used for ground truth location estimation is recorded with a resolution of 320x240 pixel and 15 frames per second by a CCTV camera located at the ceiling of the bottom left corner, as shown in figure 4.5.

4.3.1.1 One moving target

These experiments demonstrate the ability to detect the closest microphone to a sound source. The target source in this case is an adult male speaking continuously and moving around the microphone configuration. Figure 4.5 outlines the layout of the microphone distribution, the location of the noise sources and the sequences of paths the person takes for two different path patterns. The ground truth of the speaker location with respect to the microphones is estimated by a calibrated camera system, as shown in figure 4.6 for microphone positions and the target path.

![Path pattern 1](image1.png)

![Path pattern 2](image2.png)

Figure 4.5: Outline of the microphone and noise source locations for the first and second walking path pattern. The target motion is delineated in light blue, with the direction starting from one. Microphone locations are denoted by $\mathcal{O}$, noise source locations by $\mathbb{B}$ and the camera location by $\mathbb{C}$.

A distance function captured from the camera measurements is used to evaluate the acoustic importance weighting $w$. The camera measurements are projected back into real world coordinates and a two dimensional Gaussian distribution is placed at each microphone location with a standard deviation of 1.1 metre in x and y directions (illustrated in figure 75).
4.6). A Gaussian function is used as a distance measure over a Euclidean function because of the inverse square law.

(a) Path pattern 1. The area with insufficient microphone coverage, that also is shown in figure 4.7, is highlighted by a read ellipse.

(b) Path pattern 2

Figure 4.6: Gaussian distance function and ground truth (shown as black line) of the target path. Position estimations are based on camera measurements that are projected back into real world coordinates. The target path is illustrated in black and the Gaussian distribution has a standard deviation on 1.1 metres in x and y directions.

The numerical evaluation of the first (figure 4.7) and second (figure 4.8) path pattern
is made by comparing the acoustic weighting $w$ (a) to the Gaussian distance weighting (b). In both test scenarios only ambient noise was present at the beginning, hence all microphones have similar acoustic importance and no importance is shown for the distance measurement. When the test subject starts speaking, the closest microphone becomes dominant and consequently the weighting increases for both measurements.

Figures 4.7 and 4.8 show that for both scenarios the weighting functions have similar characteristics. For a numerical error comparison, the acoustic and distance weightings
are compared by evaluating the highest ranked microphone. An error is defined when the highest ranged microphone based on the distance metric does not match the highest ranged microphone based on the acoustic weighting. The overall error is 8.1% and 9.2% for the first and second path patterns, respectively. It must be pointed out that the comparison is made between distance and acoustic measurement, where the accuracy of acoustic measurements depends on the room reverberation and the direction of the emitted speech signal. The distance measurement only depends on the proximity between the subject and the microphone. It shows that the majority of errors occur during the change from one microphone to another, as well as when the spatial microphone coverage is insufficient. This is demonstrated during the time interval of 95 to 120 seconds in figure 4.7. It occurs when the test subject walks slowly from microphone 2 to 4, passing microphone 3 within a short distance but does not talk directly towards the microphone. The distance function indicates the distant proximity but the acoustic function does not indicate this as clearly. In this situation speech is directed to microphone 4 and therefore gets picked up earlier than the distance measurement would indicate. Theoretically, if this period is taken out of the error estimation, based on insufficient microphone coverage, the error rate would drop to 4.9%.

4.3.1.2 Multiple stationary targets

To evaluate the system for multiple targets the previous experiment is repeated with three stationary persons. The microphone layout is identical to the one used in the previous section. To test the ability of detecting multiple speakers, two scenarios with different speech patterns were tested:

- **Sequence 1**: Speaker 1, 2 and 3 are located near microphone 1, 4 and 5 respectively. Following are the sequences of the talking patterns from the speakers: First, there was silence followed by Speaker 1, 2 and 3 talking in consecutive order, each for a duration of about 6 seconds. Next a group of two speakers (e.g. Speaker 1 and 2) were recorded at the same time for about 4 seconds. This is followed by another group of two speakers (e.g. Speaker 2 and 3) talking. Prior to the ending was a further 5 seconds of silence.

- **Sequence 2**: Speaker 1, 2 and 3 are located near microphone 2, 4 and 5 respectively. Following are the sequences of talking patterns from the speakers: First, there was silence followed by only Speaker 1, 2 and 3 talking in consecutive order, each for a duration of about 10 seconds. Next a group of two speakers (e.g. Speaker 2 and 3)
were recorded talking at the same time for 10 seconds. After that another group of two speakers (e.g. Speaker 1 and 3) began to talk, followed by another group of two speakers (e.g. Speaker 1 and 2) talking. At the end all three speakers were talking simultaneously for a duration of 12 seconds.

(a) Sequence 1: Speakers are located at microphone 1 (blue), 4 (black) and 5 (yellow). Speech patterns are: Single speech from 9 to 26 seconds and two people speaking at the same time from 26 to 37 seconds.

(b) Sequence 2: Speakers are located at microphone 2 (red), 4 (black) and 5 (yellow). Speech patterns are: Single speech from 14 to 50 seconds, two people speaking from 50 to 86 seconds and three people speaking from 86 second to the end.

Figure 4.9: Importance weighting $w$ of each microphone for multiple active target.

Figure 4.9 shows that a single active target is indicated as a strong peak at the respective microphone. When two people talk simultaneously, the peaks at both respective microphones are clearly visible but are not as strong as for a single speech sequence. This is expected because both microphones have similar importance and the sum over all weights is normalised to 1. However, this does not mean that both microphones have the exact same weighting because the weighting factor depends on the directionality and loudness of the speech signal. The reduction in the general weight continues when three active targets are present, as is shown in figure 4.9(b) between 86 and 100 seconds. However, all three microphones have a higher weighting than microphones with only ambient noise.
4.3.2 Speech enhancement - Experimental setup

This section evaluates the performance of the proposed speech enhancement approach for different scenarios. All experiments used audio data that is sampled at 16 kHz with a resolution of 16 bits per sample. For evaluating the approach in real ambient noise, audio was captured at a crowded café and from a running scooter. These audio streams are illustrated in figure 4.11.

Experiments in section 4.3.2.1, 4.3.2.2 and 4.3.2.4 are conducted in a 5.7 × 6.8 × 2.7 metre (width, depth, height) reverberant room. It is the same room as in section 4.3.1 with the same environmental noise sources. In addition, either café or scooter noise was played back into the room to increase the background noise level and to simulate a more realistic environment. In experiment 4.3.2.2, two omnidirectional condenser microphones (Behringer ECM 8000) are distributed in the room, where one microphone dominantly recorded the ambient noise and the other microphone the ambient noise and a speech signal that was played back by a loudspeaker located next to the microphone. Experiments in section 4.3.2.1, 4.3.2.2 and 4.3.2.4 use five of the same types of microphones that are distributed over the entire room as described in figure 4.10. Two people, one male (Speaker 1) and one female (Speaker 2), were asked to stand next to two different microphones and talk using the speech pattern shown in figure 4.14. The spatial separation between the microphones next to Speaker 1 and 2 was about 2.7 metres.

![Figure 4.10: Experiment setup for evaluation of the proposed system under realistic conditions. Microphone locations are denoted by ⊙ and noise source locations by □.](image)

Experiments in sections 4.3.2.2 and 4.3.2.3 are conducted using playbacks of a recorded speech signal to provide the ability to control the signal power of the speech during data capturing. This makes it possible for experiments in section 4.3.2.2 to use the same speech sequence with the same ambient noise sequence repeatedly but with varying SNR values. For comparing the results of experiments in section 4.3.2.3 the speech signals are normalised to the same signal power. In total, speech of two female and three male adult
persons was captured.

Test data for experiments in section 4.3.2.3 are created by mixing a noise signal and the speech signal using Audacity (2008). The masking noise signal is either a synthetically generated bandlimited white noise, or previously recorded real noise from a busy café during lunch time or a running scooter.

![Café noise](image1)

![Scooter noise](image2)

Figure 4.11: Captured audio sequences of a café during lunch time and of a scooter.

4.3.2.1 Noise classification

The noise classification approach is evaluated on two data sets as follows:

- **Unmasked data**: Several test sequences with a length of 10 to 12 seconds each are randomly selected from the pre-recorded café and scooter noise. These noise sequences are used to create six test sequences with two noise changes, containing scooter and café noise in arbitrary order.

- **Recorded data in realistic environment**: To verify the results from the first data set for a reverberant environment with constant ambient noise, the previously used noise types are played back into the room that is described in section 4.3.2. Because two person were talking during the data capturing next to two microphones, the classification is based on the remaining microphones that do not contain any speech. Such an input signal of one microphone is shown in figure 4.12.
Figure 4.12: Introduced noise pattern with noise labels \( N1 \) for café and \( N2 \) for scooter noise. Instances where the noise level changes are labelled from \( C1 \) to \( C3 \). \( C1 \) and \( C2 \) are rapid noise level changes. \( C3 \) labels a time frame where the noise level slowly increases and the decreases.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Prediction</th>
<th>Café (%)</th>
<th>Scooter (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq 1</td>
<td>Café</td>
<td>98.66</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>Scooter</td>
<td>0.86</td>
<td>99.14</td>
</tr>
<tr>
<td>Seq 2</td>
<td>Café</td>
<td>99.53</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Scooter</td>
<td>11.32</td>
<td>88.68</td>
</tr>
<tr>
<td>Seq 3</td>
<td>Café</td>
<td>98.60</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>Scooter</td>
<td>0.18</td>
<td>99.82</td>
</tr>
<tr>
<td>Seq 4</td>
<td>Café</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Scooter</td>
<td>6.82</td>
<td>93.18</td>
</tr>
<tr>
<td>Seq 5</td>
<td>Café</td>
<td>94.71</td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td>Scooter</td>
<td>0.55</td>
<td>99.45</td>
</tr>
<tr>
<td>Seq 6</td>
<td>Café</td>
<td>98.28</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>Scooter</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Average</td>
<td>Café</td>
<td>98.2</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>Scooter</td>
<td>1.46</td>
<td>98.54</td>
</tr>
</tbody>
</table>

Table 4.1: Noise classification results of the six test-sequences from the first data set consisting of arbitrarily selected sub-sequences of café and scooter noise.
Table 4.1 shows the classification result in terms of a confusion matrix, of the first data set. For this two-class problem, the average classification rate for correct and incorrect classified samples is about 98.37% and 1.54% respectively.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Prediction</th>
<th>Café (%)</th>
<th>Scooter (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Café</td>
<td>99.33</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Scooter</td>
<td>2.76</td>
<td>97.24</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Noise classification result of the second data set shown in %.

The confusion matrix in table 4.2 shows the noise classification result of the second data set. It shows that the correct and incorrect classification rate is comparable to the first data set, even though measurement noise, reverberations and additional environment noise is added during the recording. Results of both data sets show that intensity changes of the test signal does not influence the classification result.

### 4.3.2.2 Voice activity detection

The voice activity detection approach is assessed on two data sets that are recorded in the described test environment in section 4.3.2. The data sets are recorded as follows:

- **Speech recorded under varying SNR**: This data set evaluates the speech classification in terms of the SNR for the base methods and the proposed combined approach. Therefore, a speech signal is played back and masked sequentially by café and scooter noise. The noise intensity stayed constant during the recording but the speech signal is lowered in signal power each time.

- **Speech recorded from two person in a real environment**: The second data set is a recording between a conversation of two people in a noisy environment. This data is used to evaluate the proposed VAD approach for a realistic situation when both persons adjust dynamically their speech level during the conversation.

Results of the VAD for different SNR situations are shown for the proposed approach, compared to the methods based on signal power and noise classification only. The evaluation is made on true positive (TP) and false positives (FP) rates, shown in figure 4.13 and table 4.3 respectively. The true positive rate for our classification problem is the percentage of correctly identified speech samples that is defined as:
Figure 4.13: Voice activity detection results in the presence of scooter (a) and café (b) noise in terms of true positive rates for various SNR levels.

\[ TP = \frac{\text{correctly classified speech samples}}{\text{total number of speech samples}} \]  

(4.24)

The false positive rate indicates how many samples are misclassified as:

\[ FP = \frac{\text{incorrectly classified speech samples}}{\text{total number of non-speech samples}} \]  

(4.25)
CHAPTER 4. DISTRIBUTED MICROPHONE NETWORK

In essence it is desirable to achieve a TP rate of 1, where all speech samples are correctly classified as speech, and a FP rate of 0, where no noise samples are misclassified as speech.

Figure 4.13 illustrates that the proposed approach outperforms both based methods as long as at least 30% of speech instances are detected based on the noise classification result. For scooter noise, figure 4.13(a) shows that after a SNR value of -6 dB the VAD based on noise classification drops under 30%, which causes the proposed approach to fall below the performance of the signal power based approach. The reason being that the proposed approach utilised the poor VAD result of the noise classification and therefore discarded classified speech samples based on the signal power. This poor classification result is caused by an increase of the correlation value of the noise classification due to a decrease of the speech content in the mixture of noise and speech. Because scooter noise has a broader spectrum than café noise, this effect is more distinct.

As the noise level is unchanged during this experiment, the misclassification rate (FP) varies only insignificantly and is therefore shown as an average value in table 4.3. It shows that café noise is more problematic than scooter noise, manifesting in a higher FP rate. Although, VAD based on the noise classification has a low misclassification rate, it is still not usable on its own because of the low TP rate, as it can be seen in figure 4.13. Therefore, the proposed approach combines the advantages of both methods and performs well for SNR situations of -6 dB and better.

<table>
<thead>
<tr>
<th>VAD Approach</th>
<th>Speech masked by scooter noise (%)</th>
<th>Speech masked by café noise (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal Power</td>
<td>7.99</td>
<td>16.14</td>
</tr>
<tr>
<td>Noise Classification</td>
<td>0.38</td>
<td>3.68</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td>1.88</td>
<td>10.41</td>
</tr>
</tbody>
</table>

Table 4.3: Average false positive rate for voice activity detection with café and scooter noise as ambient noise.

Figure 4.14 illustrates the VAD results for the second data set. Only microphones that detected any speech sequences are illustrated. Consequently the microphones located closest to Speaker 1 and Speaker 2 are shown. For evaluation only the proposed approach is tested because the previous experiment has already shown that it is superior to the signal power or noise classification based method. It can be seen that a rapid increase in signal intensity does not result in detected speech. Also, the café noise type that consists of a mixture of multiple persons talking is not detected as speech. The reason is that this noise type is known and therefore classified as a background noise. However, as soon as Speaker 1 or 2 starts talking, speech is detected because it does not fit the noise model and it also increases the usual signal power of that noise type. Table 4.4 shows the
classification results compared to the ground truth. In general, the speech detection has 
a offset of about 110 milliseconds because of the applied smoothing, and therefore speech 
is not detected at the beginning of a speech sequence, lowering the true positive value.

<table>
<thead>
<tr>
<th>Test Sequence</th>
<th>TP (%)</th>
<th>FP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>90.88</td>
<td>0.88</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>95.41</td>
<td>1.73</td>
</tr>
</tbody>
</table>

Table 4.4: Voice activity detection result for Speaker 1 and 2 in terms of true positive 
and false positive classification rate.

Figure 4.14: Result of the voice activity detection for microphones that are closest to the 
speaker. Speaker 1 is male and 2 is female. Dark red (1) marks speech sequences, active 
speech labelled as 0.1. Light blue (2) indicates speech with high energy. Dark blue (3) is 
the recorded signal and s indicates the ground truth of speech. Source: Kühnapfel et al. 
© 2008 IEEE

4.3.2.3 Noise changes during speech

The effects of using multiple noise models for enhancing a speech signal that is masked 
by different noise types is evaluated in this section. At any time only one noise type is 
present and the noise change occurs during the speech sequence. This fact maximises the
error of the enhanced speech signal when the modelled noise does not match the current noise type. The evaluation is conducted on two separate datasets as follows:

- **Synthetic noise**: A speech signal is sequentially masked by bandlimited white noise with centre frequencies of 800 Hz (White Noise 1), 2700 Hz (White Noise 2) and 5200 Hz (White Noise 3), all with a bandwidth of 1200 Hz.

- **Real noise**: Five speech sequences of about 5 seconds from two female and three male speakers are masked by either café or scooter noise.

For the first experiment, synthetically generated white noise is chosen because it is stationary and the bandlimits can be set. In this dataset the noise sources do not have any overlapping frequency ranges, and therefore, any difference in the modelled noise type and the current noise type of the test signal is shown instantly. Figure 4.15 shows the SNR value of the original speech sequence and the speech sequence masked by the white noise. When the speech sequence is mixed with the noise signal, the relative signal to noise ratio drops from about 25 dB to 3 dB.

![Figure 4.15: Speech masked by synthetic noise.](image)

The results of the speech enhancement of the speech signal masked by white noise are shown in figure 4.16 with increasing numbers of noise models from (a) to (c). For the noise sequence White Noise 1, all resulting audio streams show the same SNR, as all approaches have modelled the current noise type correctly. Differences between the enhanced audio
streams first occur during the speech sequence when the noise type changes to White Noise 2 at about 1.6 seconds. In figure 4.16(a) only one model noise source is used that does not match the new noise type. Hence, the speech signal can not be estimated accurately as the model noise cannot be updated during speech sequences. After the speech sequence, the model is updated and converges to the same result as for the approach using multiple noise models. In this case the noise model needs about 0.45 seconds to adapt to the new noise type. The same result can be seen at about 3 seconds where the noise changes again to White Noise 3. Now, only the approach that models all three noise sources is able to recover the speech signal, as seen in figure 4.16(c). In essence, the estimated speech signal is best when the noise model count equals the number of noise sources. Additionally, any of these enhanced signals has a maximum noise reduction of about 17 dB that is controlled by the floor function of the filter.

Figure 4.16: Comparison of speech enhancement results for increasing numbers of modelled noise sources.

A numerical evaluation of the previous observations from figure 4.16 is shown in table 4.5. The error in SNR between the original speech signal and the estimated speech signal is computed during speech sequences. It confirms that the error is lowest if the number of modelled noise sources equals the number of noise types.


table 4.5: Average error between the estimated speech signal to the original speech signal for synthetic noise with varying numbers of modelled noise sources.

<table>
<thead>
<tr>
<th>Number of modelled noise sources</th>
<th>Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.41</td>
</tr>
<tr>
<td>2</td>
<td>4.82</td>
</tr>
<tr>
<td>3</td>
<td>2.05</td>
</tr>
</tbody>
</table>

The second data set evaluates the speech enhancement ability for real noise types to confirm the findings of the previous experiment for real world situations. Therefore, five speech sequences are used that are masked by mainly scooter noise. At the beginning, the ambient noise type is café noise and changes rapidly to scooter noise during the beginning of the speech sequence. The original speech input and noisy speech signal is shown in figures 4.17 and 4.18 for a male and female voice respectively.

![Original speech signal](image1)

![Noisy speech signal](image2)

Figure 4.17: Input signal of the speech enhancement approach for real world noise situations with a rapid noise type change during speech for a male speaker.

For evaluating the enhanced speech signal, the noise reduction is estimated for modelling one or both noise types. As an example, figures 4.19 and 4.20 show the SNR of the enhanced speech results for a male and female voice. These figures illustrate that when both noise types are modelled (sub-figure (b)), the estimated speech signal is much closer to the original signal in comparison to the enhanced signal with only one modelled noise type (sub-figure (a)). Sub-figure (a) also shows that at about 5.7 seconds the modelled noise needs about 0.45 seconds to adapt to the current noise source after the speech sequence.
Figure 4.18: Input signal of the speech enhancement approach for real world noise situations with a rapid noise type change during speech for a female speaker.

Figure 4.19: Estimated speech signal for real world noise situations with a rapid noise type change during speech for a male speaker.
Table 4.6 confirms the visual observation numerically by computing noise reductions for one and two modelled noise sources. The noise reduction is estimated by averaging out the local minima during the speech sequence. These minima correspond to the small speech breaks with mainly noise content. When matching the noise source number with the modelled noise type number the reduction of the noise signal is greatest. On average, the noise is reduced by about 4.24 dB more when modelling all noise types.

<table>
<thead>
<tr>
<th>Test sequence</th>
<th>Count of modelled noise sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-7.20</td>
</tr>
<tr>
<td>2</td>
<td>-7.25</td>
</tr>
<tr>
<td>3</td>
<td>-7.11</td>
</tr>
<tr>
<td>4</td>
<td>-7.43</td>
</tr>
<tr>
<td>5</td>
<td>-8.24</td>
</tr>
</tbody>
</table>

**Table 4.6:** Noise levels (dB) during speech sequence for varying numbers of modelled noise sources.

### 4.3.2.4 Qualitative evaluation of the enhanced speech

The final evaluation of the speech enhancement approach is conducted by a qualitative evaluation between the original recording and the enhanced signal. Therefore, two speech
sequences of a male (Speaker 1) and female (Speaker 2) are recorded during rapid changing noise with noise pattern shown in figure 4.12.

![Figure 4.21: Audio streams for Speaker 1 (male) and Speaker 2 (female). The original audio signal is on top and the enhanced speech signal is directly beneath.](image)

Figure 4.21 shows the two relevant signals for the recorded input signal and the result of the speech enhancement. It can be seen that the noise is significantly reduced. To further evaluate the enhancement result, eleven test subjects were asked to rank the quality of the original and enhanced audio stream. As in experiment 3.3.4 the MOS scale is used for numerically grading the speech quality, ranking from 1 to 5, where 1 is the lowest possible quality (bad) and 5 is the highest perceived audio quality (best). Table 4.7 shows the results of this survey. It indicates that by removing the ambient noise, the speech quality improves from an average of 2.2 for the original audio stream to 4.3 for the enhanced audio stream.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Original</th>
<th>Enhanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>2.3</td>
<td>4.3</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>2.1</td>
<td>4.3</td>
</tr>
</tbody>
</table>

Table 4.7: Mean opinion score for the original and speech enhanced signal.

### 4.4 Conclusion

In large area acoustic surveillance, dealing with a distributed microphone network as compared to an array based deployment offers significant performance, cost and design
We first investigate the creation of a “virtual” microphone that links a freely moving subject with its audio signal. This is realised by ranking all microphones in terms of the best direct speech path. The virtual audio stream is then constructed based on this ranking by using the highest ranked microphone. The experiments show that the ranking approach not only detects one direct speech path but also those of multiple targets. However, for the generation of the virtual audio stream, only one active target is supported because only the best ranked microphone is used to create the virtual audio stream. For multiple detected targets, no association between the microphone ranking and the targets is made. A solution for this problem is to incorporate a CCTV camera for target localisation based on visual techniques. The proposed ranking technique can then be used for calibrating such a system in situations where only one target is present.

The introduction of the “virtual” microphone has solved the problem of recording an acoustic signal of a moving subject in a cost effective way. However, it does not solve the problem of ambient noise that makes the speech perception difficult. Therefore, a modified spectral subtraction approach was investigated that is able to enhance a speech signal in environments with rapid changing ambient noise. This is achieved by modelling all known noise types and switching quickly to the detected noise type. This approach outperforms the general subtraction method, especially when the noise change occurs during speech.

The performance of the speech enhancement result is closely related to the noise classification and voice activity detection result. Therefore, a noise classification technique was investigated that utilised the entire network and also gives an indication if speech is present at a particular microphone. This enables the VAD approach to identify situations where the signal intensity is changed based on varying noise situations. This reduces the false detection rate and makes the system more reliable in real world situations.

The entire system is mainly based on single microphone techniques and therefore is economically efficient to apply to any audio surveillance of larger areas. In addition, it makes such a system easily scalable. Only the global noise classification requires more than one microphone to reliably detect the ambient noise type, therefore using only microphones with no speech detected.

This initial investigation has proven the effectiveness of a distributed single microphone network for audio surveillance. Nevertheless, the proposed system still has shortcomings in real world situations. The reasons are:
• The noise classification is based on a rather small feature set of a Mel scaled power density spectrum. This means, if the number of known noise types increases, the classification accuracy could decrease. The reason for this is the features may not be enough to separate the characteristics of the different noise types.

• In real environments, the noise sources are rather non-stationary which heavily influences the results of the speech enhancement. An example is street noise where cars, trucks, motor bikes or buses emit different ambient noise, as well as the constantly changing noise level. Such a noise scenario would degrade the speech quality crucially because the noise model is not substantially updated during speech sequences.
Chapter 5

Improved speech enhancement

5.1 Introduction

In the previous chapter we showed how a network of distributed microphones is used to monitor a large area of interest. Such a configuration covers a larger spatial area than a single microphone array. However, the limitation of this approach is that the performance of classification and speech enhancement degrades in the presence of non-stationary ambient noise. Unfortunately, in real world environments, the characteristics of ambient noise is usually non-stationary with rapidly changing noise characteristics. For example, consider a café located next to a street. The ambient noise consists of conversation sounds mixed with sounds made by people using cutlery. This is compounded when a car drives by. In this situation the ambient noise would change from a mixture of voice and cutlery noise to car noise, as generally the car noise is louder. These noise types are non-stationary with higher variance associated with car noise. Car noise not only changes significantly in signal power but also in frequency due to the Doppler effect.

For enhancing a desired speech signal or a conversation, several solutions have been proposed based on a single microphone (Gustafsson et al., 2001; Benesty et al., 2005; Wójcicki et al., 2006; Breithaupt et al., 2007) or multiple microphones aligned in an array (Nordholm et al., 2004; Zotkin and Duraiswami, 2004; Dmochowski et al., 2007a; Cha et al., 2008). The array solution has its advantages when it comes to speech separation, but it also has drawbacks in having limited range and high setup costs when covering larger areas. In this chapter we will address these shortcomings by modifying the speech enhancement to compensate for non-stationary noise sources. Again, the system consists of a distributed, single microphone configuration because sound intensity of a acoustic source, such as a talking person, decreases in signal intensity when the distance between the subject and the microphone increases. The contributions in this chapter are:

- The reliable classification of multiple, non-stationary sound sources to improve the noise model selection for spectral subtraction. This classification is achieved by
CHAPTER 5. IMPROVED SPEECH ENHANCEMENT

projecting the audio features into a sub-space for each predefined noise type via PCA. The classification decision is based on the Mahalanobis distance to clusters of known noise types.

• Improved voice activity detection by utilising the distance metric of the noise classification approach in combination with the signal power. This combination ensures reliable speech detection with low false detection rates in non-stationary noise backgrounds.

• Improved dynamic updating of ambient noise models by utilising the input of microphones where no speech is detected selectively for update.

• Improved speech enhancement realised by removing the ambient noise via a modified spectral subtraction approach.

The layout of the remainder of this chapter is as follows: Section 5.2 describes the methodology of the approach, organised in noise classification, voice activity detection and speech enhancement. Experiments follow in section 5.3 confirming the usability of the proposed methodology in real world situations. Section 5.4 reflects on the system and the results of the experiments.

5.2 Methodology

A network of $K$ microphones are used to enhance the speech at any given microphone. The generic signal flow of microphone 1 for speech enhancement is shown in 5.1. Noise classification and power estimation is performed first and is then utilised for voice activity detection. For enhancing the speech at microphone 1, the spectral subtraction algorithm selectively uses the audio information of microphones where no speech is detected. If a noise only sequence is detected, the spectral subtraction algorithm uses the audio signal of microphone 1 to update the detected noise model.

For noise classification, the audio features are projected into a sub-space via PCA to reduce dimensionality. The Mahalanobis distance of the reduced feature set to the clusters of known noise sources is computed as a distance metric for classification. The voice activity detection also uses this distance measure in combination with the variance of the current signal power to the estimated noise power. The VAD output is fed back to the noise power estimation process to allow the suspension of the estimation during speech. We discuss each model in greater detail.
5.2.1 Noise classification

In Chapter 4.2.2.1, the raw input data is fed to a Mel scale filter bank. Such a feature set describes the spectral characteristics of the signal within a logarithmic scaled frequency range. Hence, the classification accuracy can drop when the number of ambient noise sources increases, mainly caused by the limited discrimination of the chosen features. Due to the complexity of non-stationary noise sources, a more complex feature set is required. Therefore, this chapter uses a larger variety of features from the time and frequency domain for classifying the ambient noise. The following features are used:

- Zero crossing rate (ZCR): This time domain feature counts the number of zero crossings of the amplitude of the raw audio signal. In general, speech has a higher ZCR than ambient noise source. This feature has been used in many voice activity detection algorithms (Saunders, 1996).

- Spectral sub-bands: The separation of the frequency spectrum into linear, logarithmic scales, such as octaves, or combinations, such as Mel-scale are useful when sound sources have different frequency components (Lu, 2001).

- Spectral centroid: This feature is the weighted mean of the spectrum, with the magnitude value as weight (Peeters, 2004). It can be described as the “balancing point” or “brightness” of the frequency spectrum and indicates whether the sound signal contains more higher or lower frequency components. Generally, music has more higher frequencies than speech, wherein the fundamental frequencies are fairly
• Spectral spread: The spectral spread describes the shape of the power spectrum by computing the variance around the spectral centroid (Peeters, 2004). This variance is generally higher for speech than for noise such as engine noise.

• Spectral skewness: The skewness measures how symmetric or asymmetric the frequency distribution is around the spectral centroid (Peeters, 2004). A value of 0 indicates a symmetric distribution and a value larger or smaller than 0 indicates an asymmetric distribution to the left or right respectively.

• Mel Frequency Cepstral Coefficient: The Cepstral coefficients are spectral shaped features and describe the spectrum of the signal within a few coefficients. For computing these coefficients a Mel scale filter bank is applied to the short-term spectrum. This scale is linear for frequencies below 1000 Hz and logarithmic for frequencies above 1000 Hz. Then a cosine transform is applied to the logarithmically scaled real values of the result of the Mel scaled filter bank spectrum. Research has shown that the MFCC features are very useful for speech detection and recognition (Vergin et al., 1999).

The attribute range of the feature types varies significantly. Therefore, all features must be normalised (Han and Kamber, 2006). This ensures that features with a larger attribute range do not “outweigh” features with smaller attribute ranges. We normalise all features into a range between 0 and 1. Even after normalisation, using more features does not automatically guarantee better classification. Therefore, for each sound source a feature set must be found that describes the characteristics of this particular source with relevant features. This can be achieved by reducing the dimensionality of the normalised feature set via PCA. PCA computes a new sub-space that describes the original feature set within a lower dimensional space, such that the data retains maximum discrimination.

Given an audio signal $x$, a short-term FFT is applied to the windowed signal with a window size of 512 samples. From each window, a vector of features, containing 15 energy sub-bands and 15 MFCCs, are extracted and referred to as $A(i) \in R^{a \times 1} \ (a = 34)$, where $i$ as time block index. During training, given $I$ feature vectors $A^n$ of noise source $n$, PCA is performed to derive the eigenvectors $V^n \in R^{a \times a}$. For reducing the dimensionality, a sub-set of the eigenvectors is selected based on their corresponding eigenvalues, described as $V'^n \in R^{a \times a'}$ with $a' < a$. Projecting $A^n(i)$ into the sub-space described by $V'^n$ reduces the dimensionality of the feature vector to $a'$ and is computed as:
$A^n(i) = (V^n)^T A^n(i)$

where $T$ denotes the transpose of matrix $V^n$ and $A^n$ is the reduced feature vector. For the distance metric it is crucial that for each noise class the dimensionality of $V^n$ is the same, meaning $a'$ is constant. For each noise source the covariance $\Sigma^n$ and mean $\mu^n$ is computed in the reduced sub-space $[A^n(i) : i = 1, 2, 3, ..., I]$. These parameters are later used for classification for computing the distance metric between a test sample and the cluster centre of each noise source.

For classification, the dimensionality of the extracted feature vector $A(i)$ of the test signal at time block $i$ is reduced to $A^n(i)$ by projecting into the sub-space of each noise source. The Mahalanobis distance is computed and compared to the learned cluster centre of each noise source respectively. The Mahalanobis distance $r$ as proposed by Maesschalck et al. (2000) is chosen over the standard euclidean distance because it incorporates the variance of the training features and is computed as:

$$r^n(i) = \sqrt{ (A^n(i) - \mu^n)^T \Sigma^n^{-1} (A^n(i) - \mu^n) }$$

where $\Sigma^n$ and $\mu^n$ are the covariance and mean of the training samples of noise source $n$. The classification result $n'$ for a single microphone is based on the shortest distance measure $r$ over all noise sources as:

$$n'(i) = \arg\min_n r^n(i)$$

Unfortunately, this classification result can be influenced if the signal contains both noise and speech. In cases where the signal to noise ratio is low, the classification result can still be correct because the noise source has a higher proportion of the recorded input signal. However, the classification degrades significantly if the power of the speech signal increases. Therefore, the final noise classification of the area of interest is based on a majority voting process over the entire microphone network, defined as $e^n$. This method takes into account only the classification results of microphones where no speech is detected. Hence, it increases the positive classification rate by rejecting samples with a mixture of speech and noise. Therefore, the final noise classification $n^*$ is computed as the highest count $e^n$ of noise type $n$ as:
CHAPTER 5. IMPROVED SPEECH ENHANCEMENT

\[ n^*(i) = \arg\max_n e^n(i) \] (5.4)

5.2.2 Voice activity detection

Voice activity can be detected in many different ways, such as by modelling speech characteristics statistically, learning the feature distribution with classifiers or analysing the signal to noise ratio. Each of these methods have their advantages and disadvantages. This section proposes a VAD based on the combination of a) the signal to noise ratio and b) the distance measurement \( r^n \) of the noise classification. These methods are detailed below.

5.2.2.1 VAD based on signal to noise ratio estimation

A similar approach as in section 4.2.2.2 is used to estimate speech sequences based on the signal to noise ratio. This general approach is improved by looking at certain frequency ranges wherein speech has dominant characteristics. These characteristics, however, can vary depending on language, sex and age. A study by Baken and Daniloff (1991) shows that the fundamental frequency of the human voice for adult males and females ranges between 60 and 280 Hz.

The voice activity detection scheme adapted here uses the frequency range from 50 Hz to 525 Hz separated into 6 sub-bands: 50-100 Hz; 100-175 Hz; 175-250 Hz; 250-325 Hz; 325-425 Hz and 425-525 Hz. Sub-bands with lower frequencies have a smaller frequency range because of the fundamental frequency of speech. The average noise power \( P \) at microphone \( k \) in sub-band \( b \) is estimated as:

\[ P_b^k(i) = (1 - \gamma_p) P_b^k(i - 1) + \gamma_p P_b^k(i) \] (5.5)

where \( \gamma_p \) is the smoothing factor. This smoothing factor is set to zero where speech sequences are detected to ensure that the noise power is only updated when speech is absent. Based on the ratio of the noise power to the current signal power, voice activity \( T \) is detected for microphone \( k \) and sub-band \( b \) as:
where $\nu_p$ is an empirical threshold which is set to maintain a desired ratio between noise and speech power. It regulates the sensitivity of the power based voice activity detection, meaning the lowest SNR where speech can be detected. The final VAD result $C^p$ at microphone $k$ is a combination of all sub-bands and is computed as:

$$C^p_k(i) = \begin{cases} 
1, & \text{if } P^p_k(i) \geq \nu_p \cdot P^p_k(i-1) \\
0, & \text{if } P^p_k(i) < \nu_p \cdot P^p_k(i-1) 
\end{cases}$$

(5.6)

$\nu_p$ is an empirical threshold which is set to maintain a desired ratio between noise and speech power. It regulates the sensitivity of the power based voice activity detection, meaning the lowest SNR where speech can be detected. The final VAD result $C^p$ at microphone $k$ is a combination of all sub-bands and is computed as:

$$C^p_k(i) = \begin{cases} 
1, & \text{if } \sum_b \nu_b T^b_k(i) \geq 0.5 \\
0, & \text{if } \sum_b \nu_b T^b_k(i) < 0.5 
\end{cases}$$

(5.7)

where $\nu_b$ is a weighting factor, with $\sum_b \nu_b = 1$. This factor can be used to increase the weight of certain frequency ranges. For example a higher weight between 60 Hz and 280 Hz or language specific adjustments can be defined. A sample is classified as speech when $C^p_k(i) = 1$.

### 5.2.2.2 VAD based on noise distance measure

Section 5.2.1 identified the global noise classification as $n^*$. Then $r^{n^*}(i)$ denotes the distance measurement of the test signal at time block $i$ to the cluster centre of noise source $n^*$. The voice activity detection result $C^N$ is based on this classification measurement at microphone $k$ and is computed as:

$$C^N_k(i) = \begin{cases} 
1, & \text{if } r^{n^*}(i) \geq \nu_r \cdot \overline{r}^{n^*} \\
0, & \text{if } r^{n^*}(i) < \nu_r \cdot \overline{r}^{n^*} 
\end{cases}$$

(5.8)

where $\overline{r}^{n^*}$ is the average inter-class distance of noise class $n^*$ and $\nu_r$ is an empirical threshold. Speech is detected if $C^N_k(i) = 1$. $\nu_r$ is used to specify the sensitivity of the voice activity detection and therefore the minimal variance to $\overline{r}^{n^*}$ that is needed to label a sample as speech (see later in 5.3.2). For classifying non-stationary noise sources the variance is naturally higher than for quasi-stationary noise. A new noise type is indicated by a large variance compared to all known noise sources.
5.2.2.3 Combined VAD

The final voice activity detection scheme is based on the two previous approaches to detect speech (section 5.2.2.1 and 5.2.2.2). Algorithm 1 describes the approach to combine both measurements for reducing the false detection rate caused by non-stationary noise. The VAD result of the signal power $C^P_k$ is used to detect the beginning and the end of any speech sequence. The reason for this is that $C^P_k$ indicates any increase in signal power which is caused by the additive speech signal. A major disadvantage however is that for non-stationary noise, misclassifications readily occur. Therefore, step 2 of the algorithm verifies that a certain amount of the entire speech sequence is also detected by using the distance measurement of the noise classification, $C^N_k$. The last assumption is that any given speech sequence must have a certain length, to remove misclassifications that occur for short time periods. For experiments in section 5.3 a value of at least 40% detected speech elements by $C^N_k$ and a minimum speech time of 1 second is chosen.

Algorithm 1: VAD combination for microphone $k$

1) Find start ($i_s$) and end ($i_e$) of speech sequence based on $C^P_k$

   $i_s = \text{first time block } i \text{ where } C^P_k(i) = 1$
   $i_e = \text{last time block } i \text{ where } C^P_k(i) = 1 \text{ with } i > i_s$

2) Verify detected speech sequence

   a) Has speech sequence at least 40% of speech elements based on $C^N_k$

      \[
      \text{count} = \sum_{i=i_s}^{i_e} C^N_k(i) = 1
      \]

      \[
      \frac{\text{count}}{i_e - i_s} \geq 0.40
      \]

   b) And is speech sequence at least 1 second long

      \[
      i_e - i_s \geq 1 \text{ second}
      \]

In comparison to the voice activity labels of section 4.2.2.2, this voice activity detection algorithm uses only two labels, speech and non-speech. The reason for this is that the modified speech enhancement approach uses a different microphone input for updating the noise model during detected speech sequences.
5.2.3 Speech enhancement

Speech enhancement is achieved using a similar approach described in Chapter 4. The approach described earlier had the disadvantage that the noise cannot be modelled during speech because speech distorts the noise signal. The error in the modelled noise then degrades the quality of the estimated speech signal, especially for non-stationary ambient noise situations. In this chapter we utilise all available microphones to compensate for this problem.

In general, we assume that the speech signal has been degraded by statistically independent additive noise (Cohen and Berdugo, 2001), as defined in equation 4.17. Therefore, as shown in equation 4.19 of Chapter 4, a time varying filter with gain function $G$ can be applied to estimate the short-term frequency spectrum of the speech signal $\hat{S}$. $G_k$ at microphone $k$ is defined as:

$$G_k(i, j) = \max \left\{ 1 - \nu_c \frac{\mathcal{P}_n(i, j)}{\mathcal{P}_k(i, j)}, \beta \right\}$$  \hspace{1cm} (5.9)

where $\mathcal{P}_k(i, j)$ is the magnitude spectrum of the input signal at microphone $k$ at time block index $i$ and frequency component $j$. $\beta$ is the floor function and $\nu_c$ is the subtraction factor. The quality of the estimated speech signal heavily relies on the accuracy of the modelled noise spectrum $\mathcal{P}_n(i, j)$ for classified noise source $n^*$ at microphone $k$. As previously mentioned, this is the main problem for the general spectral subtraction approach when it comes to non-stationary ambient noise situations. For this reason the gain function is modified to use the entire network to update the noise model during speech sequences, using only microphones with no detected speech. Such a system is able to handle significant changes in amplitude and frequency spectrum of the noise source. The updating is given as:

$$\mathcal{P}_{n^*k}(i, j) = \begin{cases} (1 - \gamma_n)\mathcal{P}_{n^*k}(i - 1, j) + \gamma_n \mathcal{P}_q(i, j), & \text{if speech} \\ (1 - \gamma_n)\mathcal{P}_{n^*k}(i - 1, j) + \gamma_n \mathcal{P}_k(i, j), & \text{otherwise} \end{cases}$$  \hspace{1cm} (5.10)

where $q$ is the microphone closest to $k$ with no detected speech and $\gamma_n$ is the exponential smoothing factor. To determine the closest microphone $q$, the microphone with similar noise characteristics, a correlation matrix is computed for noise only sequences. This
matrix is based on the correlation coefficients of the feature sets $A^{n,*}$ over all microphones of a time span over 1 second. A longer time span in comparison to single feature set is needed to compute meaningful correlation coefficients and to account for the propagation time of sound.

5.3 Experiments

In this section we investigate the proposed system in terms of noise classification, voice activity detection and speech enhancement. Therefore, four real world ambient noise sources (scooter, café, street and beach) were recorded and used as known noise types. These noise types are then either synthetically added to a speech signal using Audacity (2008) or played back as ambient noise during sound recording in the anechoic chamber at the West Australian Telecommunications Research Institute (WATRI). All experiments used audio files that are sampled at 16 kHz with a resolution of 16 bits per sample.

5.3.1 Noise classification

The noise classification is evaluated on five test signals. Within these test signals, no speech is present and the noise pattern as well as the length of the noise sequence is arbitrarily chosen as shown in figure 5.2. Test signal (a) was recorded in the anechoic chamber at WATRI and test signals (b) to (d) are created by adding the noise sequences together using Audacity (2008). Four common noise types (scooter, café, street and beach noise) are present in any test signal. The signal plots in these graphs illustrate that street and beach noise are far more non-stationary than scooter or café noise.

Before any test signal can be classified into one of the predefined noise types, the extracted audio features $A$ are projected into the sub-space $V^{m}$ of each noise source. We use the first 12 eigenvectors for each sub-space, based on empirical experiments that minimised the overall noise classification error and maximised the class separation for training sequences of each noise type. Furthermore, these 12 eigenvectors represent about 97% of the information content of the noise sources. Figure 5.3 shows a graph of the eigenvalues corresponding to the eigenvectors and confirms that the information content after the 12th eigenvector is not significant.
Figure 5.2: Test signals that are used for noise classification.
Figure 5.3: Eigenvalue of each eigenvector for scooter noise \( n1 \), café noise \( n2 \), street noise \( n3 \) and beach noise \( n4 \). Kühnapfel et al. © 2009 IEEE

Figure 5.4 shows the result of the classification if all four introduced noise types are known. As defined in equation 5.3, the lowest Mahalanobis distance measure \( r'^n \) indicates the classified noise type \( n' \) for the corresponding test sequence.

The large distance variance associated with street and beach noise reflects the observation made earlier about the non-stationary characteristics of these noise types. Street noise has the largest variance, not only because the amplitude changes the most, as visualised in figure 5.2, but also because the frequency spectrum between 0 to 3000 Hz has a large variance, as shown in figure 5.5. A quantitative evaluation of the classification results based on the ground truth of the test signals is presented in table 5.1. An overall correct classification result (TP) of 97.89% is achieved with a misclassification rate (FP) of 2.12%. The classification errors generally occur during class changes due to smoothing effects of \( r^n \). Hence, the first noise type of the test sequence is correctly classified over the entire length (TP = 100%).

<table>
<thead>
<tr>
<th>Test Sequence</th>
<th>Score</th>
<th>Café</th>
<th>Street</th>
<th>Beach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP(%)</td>
<td>FP(%)</td>
<td>TP(%)</td>
<td>FP(%)</td>
</tr>
<tr>
<td>a</td>
<td>100.0</td>
<td>0.0</td>
<td>98.0</td>
<td>2.0</td>
</tr>
<tr>
<td>b</td>
<td>100.0</td>
<td>0.0</td>
<td>98.2</td>
<td>1.8</td>
</tr>
<tr>
<td>c</td>
<td>97.1</td>
<td>2.9</td>
<td>97.5</td>
<td>2.5</td>
</tr>
<tr>
<td>d</td>
<td>99.1</td>
<td>0.9</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>e</td>
<td>96.0</td>
<td>4.0</td>
<td>97.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Average</td>
<td>98.44</td>
<td>1.56</td>
<td>98.14</td>
<td>1.86</td>
</tr>
</tbody>
</table>

Table 5.1: Noise classification result for test sequences (a) to (e).
Figure 5.4: Mahalanobis distance $r^n$ of the noise classification for test sequences (a) to (e).
CHAPTER 5. IMPROVED SPEECH ENHANCEMENT

Figure 5.5: Power spectral density of noise type scooter (a), café (b), street (c) and beach (d).

In situations where not all ambient noise sources are known, the distance metric also indicates that a new noise source may be present. To demonstrate this effect, the street noise is removed from the known noise types during training for test sequence (a). Then the classification task is repeated for test sequence (a) as shown in figure 5.6.

Figure 5.6: Mahalanobis distance $r^n$ of the noise classification with one unknown source.

Figure 5.6 shows the result of this classification. It clearly illustrates that in the time frame from 63 to 97 seconds, no noise source is detected that has a considerably lower distance.
measure $r^n$ for one particular noise type. It is true that a strong speech signal mixed with a known noise type would have a similar effect. However, it would be quite unlikely that all $K$ microphones of the network would pick up a speech sequence at the same time and for the same time span. Additionally, if this time span is rather long (as in this example about 35 seconds) it is more likely that the cause of this poor noise classification results is an unknown noise source.

\subsection{5.3.2 Speech parameters estimation}

In this section, we will investigate the effect of threshold parameters $\nu_p$ and $\nu_r$ on speech classification. An audio signal of about 25 seconds is selected from each noise type and synthetically added with a 4 second speech signal. The resulting signals have a signal to noise ratio of -4.65dB, -4.26dB, 0.9dB and -2.02dB for scooter, café, street and beach noise respectively. Due to the characteristics of street and beach noise, the SNR values are higher because the signal power changes rapidly for short periods of time, and further, the speech signal is generally longer than a car passing by or an incoming wave. At the peak of the signal power changes for street and beach noise the SNR values are -5.4dB and -5.84dB respectively. The SNR value set for this experiment is important to ensure that further speech sequences with the same or a greater SNR value are detected.

Figure 5.7 shows two ROC (Receiver Operating Characteristic) graphs for the parameter range of $\nu_p$ and $\nu_r$ that describes the trade-off between the true positive and false positive rate. In general, an ROC graph describe the performance of a classifier for a two-class problem (speech/noise). Such a graph has three distinct points, the lower right corner $(0,0)$, the upper left corner $(0,1)$ and the upper right corner $(1,1)$. Classification results near $(0,0)$ and $(1,1)$ indicate a poor classifier. Point $(0,0)$ describes the situation where a positive sample (speech) is never classified correctly and at point $(1,1)$ any sample is classified as a positive sample. A perfect classifier is one when the classification rate is located at point $(0,1)$ where all positive samples are correctly classified and no negative sample (noise) is falsely classified.

Figure 5.7 illustrates that voice activity detection based on only one measurement, either noise classification or signal power, has weaknesses, especially when it comes to beach or street noise. For VAD based on noise classification, graph (a), beach noise performs poorly because the spectrum of this noise type is masking out the speech signal. This is indicated by the low true positive classification result for nearly the entire value range of $\nu_r$. Only when the false positive rate reaches 25%, the true positive rate is in an acceptable range of
over 80%. Street noise on the other hand is problematic when speech is estimated based on signal power, graph (b). The reason is quite obvious, because if a car is passing by, the signal power quickly increases, which is classified as speech. This is again indicated by a low true positive rate for a false positive range of below 25%. Even so, each of these approaches has its downside but the combination of both methods provides a far more reliable and accurate classifier. Table 5.2 presents the final combined voice activity detection result. For combining both approaches, the parameters $\nu_r$ and $\nu_p$ of the classifier are set to 1.2 and 1.5 respectively based on the results of figure 5.7.

<table>
<thead>
<tr>
<th>Noise</th>
<th>TP (%)</th>
<th>FP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scooter</td>
<td>97.24</td>
<td>3.16</td>
</tr>
<tr>
<td>Café</td>
<td>97.23</td>
<td>3.76</td>
</tr>
<tr>
<td>Street</td>
<td>98.03</td>
<td>4.56</td>
</tr>
<tr>
<td>Beach</td>
<td>82.65</td>
<td>10.53</td>
</tr>
</tbody>
</table>

Table 5.2: Speech detection result for the proposed VAD approach measured in true and false positive detection rate. Kühnapfel et al. © 2009 IEEE
CHAPTER 5. IMPROVED SPEECH ENHANCEMENT

5.3.3 Voice activity detection

The proposed voice activity detection approach is evaluated by comparing it against the VAD result of the advanced front-end feature extraction algorithm (ES 202 050) by ETSI (2007) and the support vector machine (Gunn, 1998; Vapnik, 2000). The ES 202 050 standard is developed for mobile communication and one of its aims is to reduce the data transfer rate by detection speech sequences. The SVM is a well known classifier and uses 20 MFCC as features with a radial basis function as kernel function. Six sequences are used to evaluate all three approaches, composed of one synthetic and five real noise sequences. The synthetic test sequence consists of white noise with one speech sequence of -6 dB. The five test signals for real noise have the same noise pattern as in experiment 5.3.1, containing scooter, café, street or beach noise with one speech sequence during each noise type as shown in figure 5.8.

Table 5.3 presents the result of the comparison that is measured by true positive and false positive rates. For white noise situations, all methods achieve a positive hit rate of over 95% whereas the proposed approach has the lowest false positive rate. When it comes to real noise situations, our VAD method reaches about the same average true positive rate of 97.45% with only a slight increase in the average false positives of 3.84% when compared with synthetic noise. On the other hand, the results of ES 202 050 are clearly unacceptable, classifying on average 93.29% of noise samples as speech. SVM has in comparison to ES 202 050 a far lower false positive rate but only detected on average 83.40% of the speech sequences.

<table>
<thead>
<tr>
<th>Test Sequence</th>
<th>ES 202 050</th>
<th>SVM</th>
<th>Proposed VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP (%)</td>
<td>FP (%)</td>
<td>TP (%)</td>
</tr>
<tr>
<td>White noise</td>
<td>99.03</td>
<td>6.41</td>
<td>97.94</td>
</tr>
<tr>
<td>Sequence (a)</td>
<td>100.00</td>
<td>91.45</td>
<td>81.82</td>
</tr>
<tr>
<td>Sequence (b)</td>
<td>100.00</td>
<td>95.26</td>
<td>84.50</td>
</tr>
<tr>
<td>Sequence (c)</td>
<td>100.00</td>
<td>98.42</td>
<td>80.70</td>
</tr>
<tr>
<td>Sequence (d)</td>
<td>90.39</td>
<td>88.32</td>
<td>87.65</td>
</tr>
<tr>
<td>Sequence (e)</td>
<td>99.84</td>
<td>93.01</td>
<td>82.32</td>
</tr>
</tbody>
</table>

Table 5.3: Comparison of VAD methods measured in true positive and false positive classification results.

5.3.4 Speech enhancement

In this section we investigate the result of the speech enhancement. Therefore, the test sequence (a) from the previous experiment (section 5.3.3) with real noise types is chosen
(a) Noise pattern as shown is figure 5.2(a). The audio signal is recorded in the anechoic chamber at WATRI.

(b) Noise pattern as shown is figure 5.2(b). Noise and speech signal are mixed with Audacity.

(c) Noise pattern as shown is figure 5.2(c). Noise and speech signal are mixed with Audacity.

(d) Noise pattern as shown is figure 5.2(d). Noise and speech signal are mixed with Audacity.

(e) Noise pattern as shown is figure 5.2(e). Noise and speech signal are mixed with Audacity.

Figure 5.8: Test sequences for voice activity detection. Speech sequences are labelled with s.
as test input. The evaluation is done between the enhanced signals based on the approach of Chapter 4 and the proposed approach of this chapter. For this test scenario the approach of Chapter 4 is very similar to the general magnitude spectral subtraction approach because the noise model is not updated during speech, and therefore referred to as “general approach”. Figure 5.9 illustrates the original and both enhanced audio signals. Only a subsection is shown for both enhanced signals indicated by a red rectangle in the original signal (a). This subsection visualises the difference in noise reduction with highly non-stationary ambient noise types for the approach of Chapter 4 (b) and the modified approach (c). It can be seen that in (b) more ambient noise remains than in (c).

Figure 5.9: (a) shows the original audio signal and the subsection, marked by a rectangle, of the enhanced signal by the general spectral subtraction approach (b) and the proposed method (c). Source: Kühnapfel et al. © 2009 IEEE

The visual observation is further verified by an auditory survey where eleven subjects were asked to grade numerically the quality of the speech and the level of reduction of ambient noise. As in experiment 3.3.4 the MOS scale is used, ranking from 1 to 5. Where 1 is the lowest possible quality (bad) and 5 is the highest perceived audio quality (best).

As shown in table 5.4 according to the test subjects, the proposed approach is able to remove more of the background noise than the general approach. The average score for the proposed method is 4.24 compared to 3.59 for the general method. During highly non-stationary noise types, such as street and beach noise, the difference is greatest. This fact is expected because only the proposed approach is able to update the noise mode
5.4 Conclusion

Chapter 4 has demonstrated the effectiveness of distributing single microphones for acoustic surveillance of large areas of interest. However, the problem lies with enhancing speech when non-stationary ambient noise types are present. Therefore, we have shown in this chapter an approach for classifying non-stationary noise types, detecting voice activity and enhancing speech by using the entire network.

The first task was to reliably detect the background noise type and identify any unknown noise types. This was achieved by extracting a set of audio features from the raw input signal and learning a reduced sub-space for each noise type to reduce the dimensionality of the features. Therefore, each sub-space represents the characteristics of the individual noise type and a test signal that is projected into such a sub-space will only be close to the class centre if the test signal is close to that noise type. As a distance metric, the Mahalanobis distance measure was chosen. Experiments for a four class problem have shown a mean TP classification accuracy of 97.89% for noise only sequences. In addition,
the classification result is used to indicate an unknown noise source.

Before speech could be enhanced, sequences containing speech elements must be found. This was realised by a combination of power based VAD and the distance measure of the noise classification result. Even so, power based voice activity detection has its problems with non-stationary noise situations, the combination with the noise classification result is shown to significantly reduce this problem. In comparison with ES 202 050 standard and SVM, the proposed approach was able to achieve a mean TP classification rate of 97.45% for speech with a low average false detection rate of 3.84% for real noise situations.

The final goal of the presented system is to enhance speech in various ambient noise situations. Therefore, the spectral subtraction approach of Chapter 4 has been modified to handle non-stationary noise types. The crucial adjustment is that the algorithm utilised the entire network for updating the noise model during speech sequences. Due to the modifications the algorithm was able to remove more ambient noise with the consequence of better speech enhancement when highly non-stationary noise is present. For removing noise during street noise, our approach achieves a higher score on the MOS scale when compared to the general approach.

The contribution of this work is the presentation of a system that is able to enhance speech in challenging noise situations. It also solves the problem in a cost effective way by using a network of single microphones to minimise the setup cost for large areas of interest.
Chapter 6

Conclusion

This thesis investigated the use of acoustic information for surveillance purposes. In particular, the problem of integrating acoustic with CCTV information and speech enhancement of a moving target in noisy environments was explored. The first aim of this thesis was to develop a technique for the detection and analysis of multi-modal surveillance events. Thus, an investigation into a generic framework for combining audio information of a linear microphone array and a single CCTV camera was conducted. Experimental results expose the limitation of the linear microphone array geometry in terms of speech enhancement for large areas of interest. Thus, the second aim of this thesis was the investigation of an economical speech enhancement algorithm for large areas and environments with non-stationary noise types. Hence, a distributed microphone network architecture was investigated that can be scaled for large scale surveillance tasks. A technique for creating a "virtual" microphone to capture the audio of a moving target was presented and an improved speech enhancement approach was proposed based on a single-channel speech enhancement technique.

The introduction of a microphone array into the video surveillance domain was described in Chapter 3. A generic framework was presented that captures, synchronises and extracts high level information from different modalities. The extracted high level information, such as sound source direction and object location in the video stream were used to dynamically learn a mapping function that maps the one dimensional acoustic source direction to the two dimensional target location in the video image. Thus, no knowledge of the relative sensor alignment was needed in advance and the approach dynamically adjusts to progressive drifts in the sensor suite. Additionally, a novel smoothing technique was described that minimises the error for the acoustic source localisation that are caused by reverberation effects and other ambient noise sources. Experiments demonstrated that this framework can enhance a speech signal of a desired target by localising the target position in the video stream and generate a multi-modal label for the event.

Chapter 4 highlighted the limitations of using array based speech enhancement techniques in terms of the signal power reduction when the distance between sensor to source in-
CHAPTER 6. CONCLUSION

creases. Therefore, this chapter proposed a technique for recording the acoustic signal of a moving target and enhancing speech sequences, utilising a distributed microphone network. Such a network is a cost effective solution for observing large areas in comparison to array based solutions. First, a technique was described to create a “virtual” microphone for producing the best possible sound recording of a moving target by dynamically selecting the microphone that records the sound with the best signal to noise ratio. Next, an improved spectral subtraction approach was presented to enhance the speech signal that is masked by noise. The modification of the spectral subtraction algorithm consists of modelling multiple noise conditions and utilising the entire network to classify the current noise conditions, using only microphones with no detected speech. Experiments proved that the proposed algorithm is able to minimise errors during rapid noise changes by selecting the right noise model instead of slowly updating it over time.

In Chapter 5, the proposed speech enhancement approach for a distributed microphone network was further improved to account for real world noise situations. The focus is directed at improving the noise classification and the enhancement ability for multiple and non-stationary noise situations. The classification technique is improved by incorporating a variety of acoustic features and using PCA to reduce the feature dimensionality. The enhancement algorithm uses the entire network to update the noise model, significantly improving the speech enhancement for non-stationary noise environments. Experiments on real noise recordings, such as street or beach noise confirmed the effectiveness of the presented technique.

6.1 Future work

This thesis presented an initial investigation into creating a “virtual” microphone for recording an optimal sound signal of a freely moving target based on a distributed microphone network. Given the limitation of multiple moving sources, the proposed approach would benefit from incorporating CCTV data for target association, and thus create multiple “virtual” microphones. The ranking approach for estimating the microphone importance can be used to calibrate such a system dynamically by using situations where only a single target is detected by both media. Experiments on multiple stationary targets have shown that our algorithm is able to detect multiple microphones that are important. However, the association between a specific target person and a microphone cannot be made. Thus, utilising the location information of detected target in the video image enables the system to compute the likelihood of associating a microphone based on the localisation estimation and the importance ranking of the microphones.
Chapter 3 investigated the use of multi-modal labels for surveillance events for a small sensor setup. The results suggest that the introduction of a CCTV camera into the distributed microphone network will improve the automatic scene analysis beyond sound classification and speech enhancement. The main advantage is that the speaker or the sound event can be visually identified and maintained by tracking. However, accurate target localisation mostly suffers from occlusion, thus a multi-camera approach would be most useful.

The association of the enhanced speech signal with the visual observation of the target improves the identification of a target and the situation awareness. In addition, the enhanced speech signal can be used in combination with an automatic speech recognition algorithm to determine the language or creating a text record. The main reason to increase the amount of high level information is that a data mining algorithm can be used more efficiently to analyse the events and highlight situations of interest.
Bibliography


ETSI (2007). ES 202 050 V1.1.5: Speech processing, transmission and quality aspects (STQ): Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms.


BIBLIOGRAPHY


BIBLIOGRAPHY


*Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.*
Appendix A

Audio Features

Raw sound signals in time or frequency domains are unusable for classification and detection of desired acoustic events. There are two main reasons: 1) the number of samples or frequencies that are required to characterise even a short audio signal is quite large and 2) the amplitude or a single frequency value does not discriminate sound events well. Therefore, the signal of the time domain or frequency domain is generally characterised by temporal or spectral features respectively.

A.1 Pre-processing

Principally, any system that uses an audio signal to learn or detect certain acoustic characteristics performs a sequential analysis of the data. Therefore, the audio stream of the time domain $x(t)$ with $t$ as discrete time index or sample, is windowed in blocks or frames of samples $x'(i)$ that often overlap by 30% - 50% of the window size $T$, where $i$ is the time block index. The window size for segmenting the audio stream depends on the acoustic characteristics that need to be detected. For example, sound events of a short duration, such as shooting or breaking glass, generally have shorter window sizes compared to events with a longer duration, such as an alarm signal or other constant background noise. Depending on the features that are extracted from this windowed data, different windowing functions are applied. A brief overview of the most commonly used window functions $W$ (Mertins, 1999) follows:

Rectangular window:

$$W(t) = \begin{cases} 1, & t = 0, 1, \ldots, T - 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{(A.1)}$$

Hanning window:

$$W(t) = \begin{cases} 0.5 - 0.5 \cos \left( \frac{2 \pi t}{T-1} \right), & t = 0, 1, \ldots, T - 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{(A.2)}$$
APPENDIX A. AUDIO FEATURES

Hamming window:

\[ W(t) = \begin{cases} 
0.54 - 0.46 \cos \left( \frac{2\pi t}{T-1} \right), & t = 0, 1, \ldots, T - 1 \\ 
0, & \text{otherwise.} 
\end{cases} \] (A.3)

Blackmann window:

\[ W(t) = \begin{cases} 
0.42 - 0.5 \cos \left( \frac{2\pi t}{T-1} \right) + 0.008 \cos \left( \frac{4\pi t}{7-1} \right), & t = 0, 1, \ldots, T - 1 \\ 
0, & \text{otherwise.} 
\end{cases} \] (A.4)

Bartlett:

\[ W(t) = \begin{cases} 
\frac{2}{T-1} \left( \frac{T-1}{2} - |t - \frac{T-1}{2}| \right), & t = 0, 1, \ldots, T - 1 \\ 
0, & \text{otherwise.} 
\end{cases} \] (A.5)

A.2 Domain transformations

In this section we give a brief overview of the theory of transforming a signal between the time and frequency domain. A detailed explanation of this material is given by Walker (1996); Oppenheim et al. (1999). The well known Fourier transform describes the transformation of a continuous signal \( x \) at time \( t \) into the frequency domain \( X \) with \( f \) as frequency as:

\[ X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \] (A.6)

where \( j \) is the square root of -1. The inverse transformation is defined as:

\[ x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df \] (A.7)

However, a recorded audio signal consists of discrete sampled observations, and hence the theory of the Fourier transformation cannot be directly applied as it only applies to a continuous case. Therefore, a special kind of the Fourier analysis, the discrete Fourier transformation (DFT), is introduced. Consider a discrete time series \( x(t) = x_0, x_1, x_2, \ldots, x_{T-1} \) of \( T \) samples, that is outside the range of \( 0 \geq t \geq T - 1 \) extended \( T \)-periodic, the transformation to the frequency domain \( X(j) : j = 0 : T - 1 \), with \( j \) as frequency index , is given as:

\[ X(j) = \sum_{t=0}^{T-1} x(t) e^{-j2\pi jt/T}, \quad \text{for } j = 0, \ldots, T - 1 \] (A.8)

The inverse transformation is defined as:
\[ x(t) = \frac{1}{T} \sum_{j=0}^{T-1} X(j)e^{-j2\pi jt/T}, \quad \text{for } t = 0, \ldots, T - 1 \quad (A.9) \]

The drawback of the DFT is its computational complexity, because the entire sequences need to be traversed for computing one frequency value or sample. Thus, Cooley and Tukey (1965) presented a fast Fourier transformation algorithm that factorises the DFT approach by using smaller sequences, increasing the performance significantly. This approach is defined where the sequence length (FFT size) \( T \) is a power of two. However, there are also algorithms that can handle sequences that are not of powers of two. There is one more commonly used transformation technique in the area of digital signal processing, called short time Fourier transformation (STFT). In principle, this algorithm uses the same theory as above but also applies a window function to the sequence such that only a small period of time is analysed.

### A.3 Temporal features

All temporal features do not need much pre-processing as they are based on the discrete amplitude values of the recorded audio signal. Usually, only windowing of the audio signal is performed beforehand. This section describes the most commonly applied audio features; more temporal features are introduced in work by Lu (2001), Peeters (2004) and Theimer et al. (2008).

#### A.3.1 Zero crossing rate

The zero crossing rate measures the frequency of how often successive samples of a discrete signal have different signs, and therefore cross the zero value of the amplitude axis (Zhang and Kuo, 2001; Tzanetakis and Cook, 2002). Generally, harmonic signals have less zero crossings than a noise signal. The zero crossing rate \( Z \) of a audio signal with the total number of samples \( T \) is defined as:

\[
Z = \frac{1}{2} \sum_{t=0}^{T-1} |\text{sgn}[x(t)] - \text{sgn}[x(t-1)]| \quad (A.10)
\]

where \( x(t) \) is the amplitude of sample \( t \) and \( \text{sgn}[x(t)] \) will be 1 when \( x(t) \) is positive and -1 for negative values of \( x(t) \).

#### A.3.2 Short time energy

The short time energy (STE) is used to measure the signal power over an audio sequence of a given, static time frame (Zhang and Kuo, 2001). The chosen time frame is generally
very short in comparison to the entire audio signal. The reason is that these measurements are used to compute the signal power envelop of the audio signal. 

The energy $E$ of an audio sequence $x$ is directly determined by summing over the squared amplitude values as:

$$E_{ste} = \frac{1}{T} \sum_{t=0}^{T-1} x(t)^2$$  \hspace{1cm} (A.11)$$

where $t$ is the sample and $T$ the total number of samples of the audio sequence.

An alternative technique to measure the signal energy is the root mean square (RMS) method. Here the result of the squared and summed up amplitude values are normalised by the numbers of samples $T$ and the root is taken as:

$$E_{rms} = \sqrt{\frac{1}{T} \sum_{t=0}^{T-1} x(t)^2}$$  \hspace{1cm} (A.12)$$

The physical energy measurement is analogous to the psychoacoustical loudness property, relating to the human perception of the amplitude level of an audio signal.

A.3.3 Silence ratio

The silence ratio (SR) defines the amount of silence of an audio sequence in comparison to the length (Chen et al., 2006). Determining the amount of silence is achieved by splitting the audio sequence in several smaller frames and computing the STE. All frames that have a smaller STE value below a set threshold are considered as silence. This, however, is the downside of SR because this threshold is a critical factor of this feature. The threshold can be set to a fixed level heuristically or by selecting a reference silence sequence. Then, this feature cannot adjust to noise changes. Thus, the third option is to dynamically estimate an average noise energy level over time to set the threshold. Similar to ZCR, this feature can be used to distinguish between noisy and harmonic signals. The reason is that harmonic signals usually have a higher SR value than noisy signals.

A.3.4 Auto-correlation

The auto-correlation (Brown, 1998) is a special kind of the cross-correlation function where the signal is correlated with itself. It can be used to find recurring patterns of the signal, and hence how harmonic or noisy a signal is. The correlation $R$ of an audio signal $x$ of $T$ samples is computed as:

$$R(k) = \sum_{t=0}^{T-k-1} x(t)x(t + k)$$  \hspace{1cm} (A.13)$$
where \( t \) is the sample and \( k \) is the signal offset, with usually \( k \ll \mathcal{T} \).

**A.4 Spectral features**

Spectral features are based on the frequency components of the audio signal. Thus, the audio signal is first transformed into the frequency domain via FFT, where the FFT size determines the discrete resolution of the frequency representation. However, a high resolution increases the computational demand when computing the final spectral features. This section gives an overview on most frequently applied spectral features. Further features are described in Lu (2001), Peeters (2004) and Theimer et al. (2008).

**A.4.1 Energy sub-bands**

The energy content of defined sub-bands is a useful information for distinguishing sounds, if the sounds have different frequency characteristics (Lu, 2001). For example, speech has its most significant frequency component in lower frequency ranges, depending on sex and age of the person (Baken and Daniloff, 1991). The frequency distribution \( X \) of an audio sequence can directly be used to compute the energy \( E \) of sub-band \( k \) as:

\[
E_k = \frac{1}{J_{\text{max}} - J_{\text{min}} + 1} \sum_{j=J_{\text{min}}}^{J_{\text{max}}} |X(j)|^2
\]  

(A.14)

where \( j \) is the frequency index and \( J_{\text{min}} \) and \( J_{\text{max}} \) are the lower and upper frequency index of the sub-band respectively. The sub-band limits are dependent on the sound characteristics, and are hence often chosen heuristically. However, if imported sound characteristics are in lower frequency ranges, the octave bands could provide a general and logarithmic sub-band allocation (Tzanetakis and Cook, 2002). The resolution, and thus the bandwidth, is defined as:

\[
\frac{f_h}{f_l} = 2^r
\]  

(A.15)

where \( f_l \) and \( f_h \) are the lower and upper frequency boundary and \( r \) is the octave ratio with \( r = 1 \) as full octave and \( r = 1/3 \) as one-third octave. Table A.1 shows the centre frequency and bandwidth of the full and one-third octave. Where the centre frequency \( f_c \) and bandwidth \( b \) is given as:

\[
f_c = f_l \sqrt{2}, \quad \text{respectively} \quad f_c = \frac{f_h}{\sqrt{2}}
\]  

(A.16)

\[
b = f_h - f_l
\]  

(A.17)
## APPENDIX A. AUDIO FEATURES

### A.4.2 Mel filter bank

Similar to energy sub-band features are the features of the Mel filter bank, modelling the human acoustic perception by a set of overlapping triangular band filters. The filters are based on the frequency representation of the Mel scale. This scale is a logarithmic wrapping $\mathcal{M}$ of the linear frequency space, giving lower frequencies more importance and is described as:

<table>
<thead>
<tr>
<th>Centre frequency (Hz)</th>
<th>Bandwidth (Hz)</th>
<th>Centre frequency (Hz)</th>
<th>Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>11</td>
<td>16</td>
<td>3.6</td>
</tr>
<tr>
<td>20</td>
<td>4.6</td>
<td>25</td>
<td>5.8</td>
</tr>
<tr>
<td>31.5</td>
<td>22</td>
<td>31.5</td>
<td>7.3</td>
</tr>
<tr>
<td>40</td>
<td>9.2</td>
<td>50</td>
<td>11.5</td>
</tr>
<tr>
<td>63</td>
<td>44</td>
<td>63</td>
<td>14.6</td>
</tr>
<tr>
<td>80</td>
<td>18.3</td>
<td>100</td>
<td>22.9</td>
</tr>
<tr>
<td>125</td>
<td>89</td>
<td>125</td>
<td>29</td>
</tr>
<tr>
<td>160</td>
<td>37</td>
<td>200</td>
<td>46</td>
</tr>
<tr>
<td>250</td>
<td>178</td>
<td>250</td>
<td>58</td>
</tr>
<tr>
<td>315</td>
<td>73</td>
<td>400</td>
<td>92</td>
</tr>
<tr>
<td>500</td>
<td>455</td>
<td>500</td>
<td>105</td>
</tr>
<tr>
<td>630</td>
<td>145</td>
<td>800</td>
<td>183</td>
</tr>
<tr>
<td>1000</td>
<td>710</td>
<td>1000</td>
<td>231</td>
</tr>
<tr>
<td>1250</td>
<td>291</td>
<td>1600</td>
<td>365</td>
</tr>
<tr>
<td>2000</td>
<td>1420</td>
<td>2000</td>
<td>461</td>
</tr>
<tr>
<td>2500</td>
<td>579</td>
<td>3150</td>
<td>730</td>
</tr>
<tr>
<td>4000</td>
<td>2840</td>
<td>4000</td>
<td>919</td>
</tr>
<tr>
<td>5000</td>
<td>1156</td>
<td>6300</td>
<td>1456</td>
</tr>
<tr>
<td>8000</td>
<td>5680</td>
<td>8000</td>
<td>1834</td>
</tr>
<tr>
<td>10000</td>
<td>2307</td>
<td>12220</td>
<td>2910</td>
</tr>
<tr>
<td>16000</td>
<td>11360</td>
<td>16000</td>
<td>3650</td>
</tr>
</tbody>
</table>

Table A.1: Centre frequency and bandwidth of 1 and 1/3 octave spectrum.

### A.4.2 Mel filter bank

Similar to energy sub-band features are the features of the Mel filter bank, modelling the human acoustic perception by a set of overlapping triangular band filters. The filters are based on the frequency representation of the Mel scale. This scale is a logarithmic wrapping $\mathcal{M}$ of the linear frequency space, giving lower frequencies more importance and is described as:

$$\mathcal{M}(f) = \frac{2595 \log_{10}(f)}{f_0}$$

where $f_0$ is the reference frequency (usually set to 300 Hz).
APPENDIX A. AUDIO FEATURES

\[ \mathcal{M}(f) = 2595 \log(1 + \frac{f}{700}) \]  
(A.18)

where \( f \) is the linear frequency value. The centre frequencies \( f_c \) of the triangular filter are given as (ETSI, 2007):

\[ f_c(m) = \mathcal{M}^{-1} \left( \mathcal{M}(f_l) + m \frac{\mathcal{M}(f_h) - \mathcal{M}(f_l)}{M + 1} \right) \]  
(A.19)

where \( m = 1 : M \in \mathbb{N} \) is the filter index of \( M \) triangular filter. The frequency range of the filter is defined by the lowest used frequency \( f_l \) and highest frequency \( f_h \). An example of a Mel filter bank for \( M = 20 \) triangular Mel filters is shown in figure A.1.

![Figure A.1: Mel scale filter bank for \( M = 20 \) filter channels.](image)

A.4.3 Spatial centroid

The spectral centroid represents a weighted mean of the spectrum, with the magnitude value as weights (Peeters, 2004). This feature is also known as “balancing” point, “brightness” or “gravity” of the frequency distribution and indicates whether the sound sequence contains more higher or lower frequency components. Generally, the centroid of street noises is lower than of speech, which itself is lower than that of music. Mathematically, the spectral centroid \( C \) of the discrete frequency distribution \( X \) is defined as:

\[ C = \frac{1}{\sum_{j=1}^{J} |X(j)|} \sum_{j=1}^{J} |X(j)| \mathcal{G}(j) \]  
(A.20)

where \( j = 1 : J - 1 \in \mathbb{N} \) is the frequency index of \( J \) frequency components and the function \( \mathcal{G} \) gets the centre frequency of frequency index \( j \). The first value of the frequency distribution \( X(0) \) denotes the average of the audio sequence, the DC value, and is discarded when computing the centroid.
A.4.4 Spectral spread

The spectral spread describes the shape of the power spectrum by computing the variance around the spectral centroid (Peeters, 2004). Thus, it is an indication of the bandwidth that determines if the main frequency components are close to the spectral centroid or not. The spread $S$ is a discrete frequency distribution $X$ is given as:

$$S = \frac{1}{\sum_{j=1}^{J} |X(j)|} \sum_{j=1}^{J} |X(j)| (G(j) - C)^2$$ (A.21)

where $j = 1 : J - 1 \in \mathbb{N}$ is the frequency index of $J$ frequency components. $C$ is the spectral centroid and the function $G$ gets the centre frequency of frequency index $j$. The first value of the frequency distribution $X(0)$ denotes the average of the audio sequence, the DC value, and is discarded when computing the spread.

A.4.5 Mel Frequency Cepstral Coefficient

The Mel Frequency Cepstral Coefficient are spectral shaped features and describe the amplitude spectrum of the signal within a few coefficients (Davis and Mermelstein, 1980; Peeters, 2004; Theimer et al., 2008). For computing these coefficients a Mel scale filter bank is applied to the frequency spectrum. Then, a cosine transform is applied to the logarithmically scaled real values of the result of the Mel scaled filter bank spectrum. Figure A.2 illustrates the signal flow to extract the coefficients.

$$\frac{1}{\sum_{j=1}^{J} |X(j)|} \sum_{j=1}^{J} |X(j)| (G(j) - C)^2$$

Figure A.2: Signal flow of MFCC extraction.

Because the Mel frequency Cepstral Coefficients are based on the Mel scale, which models the human perception of speech, these features are commonly used for speech detection (Foote, 1997), speaker recognition (Murty and Yegnanarayana, 2006) and word recognition (Pearce and Hirsch, 2000).