THE UNIVERSITY *of* EDINBURGH

Speech processing using digital MEMS microphones

Erich Zwyssig

Doctor of Philosophy

Centre for Speech Technology Research

School of Informatics

University of Edinburgh

2013

Abstract

The last few years have seen the start of a unique change in microphones for consumer devices such as smartphones or tablets. Almost all analogue capacitive microphones are being replaced by digital silicon microphones or MEMS microphones.

MEMS microphones perform differently to conventional analogue microphones. Their greatest disadvantage is significantly increased self-noise or decreased SNR, while their most significant benefits are ease of design and manufacturing and improved sensitivity matching.

This thesis presents research on speech processing, comparing conventional analogue microphones with the newly available digital MEMS microphones. Specifically, voice activity detection, speaker diarisation (who spoke when), speech separation and speech recognition are looked at in detail.

In order to carry out this research different microphone arrays were built using digital MEMS microphones and corpora were recorded to test existing algorithms and devise new ones. Some corpora that were created for the purpose of this research will be released to the public in 2013.

It was found that the most commonly used VAD algorithm in current state-of-the-art diarisation systems is not the best-performing one, i.e. MLP-based voice activity detection consistently outperforms the more frequently used GMM-HMM-based VAD schemes. In addition, an algorithm was derived that can determine the number of active speakers in a meeting recording given audio data from a microphone array of known geometry, leading to improved diarisation results.

Finally, speech separation experiments were carried out using different post-filtering algorithms, matching or exceeding current state-of-the art results.

The performance of the algorithms and methods presented in this thesis was verified by comparing their output using speech recognition tools and simple MLLR adaptation and the results are presented as word error rates, an easily comprehensible scale.

To summarise, using speech recognition and speech separation experiments, this thesis demonstrates that the significantly reduced SNR of the MEMS microphone can be compensated for with well established adaptation techniques such as MLLR. MEMS microphones do not affect voice activity detection and speaker diarisation performance.

Zusammenfassung

Mobile Telefone und Tablett-Rechner durchlaufen in den letzten paar Jahren eine ungemeine Veränderung, indem fast alle ihre analogen Mikrofone durch digitale Silikon-Mikrofone ersetzt werden.

Silikon-, d.h. MEMS-Mikrofone unterscheiden sich leicht in ihren Leistungsmerkmalen verglichen mit konventionellen analogen Kondensator- oder Elektretmikrofonen. Der größte Nachteil von MEMS-Mikrofonen ist das deutlich größere Selbstgeräusch oder verkleinerte Signal-Rausch-Verhältnis (SRV), während ihre größten Vorteile der einfachere Einbau und die leichte Fabrikation sowie vor allem eine stark verbesserte angepasste Empfindlichkeit sind.

Diese Doktorarbeit umfasst eine Forschungsarbeit auf dem Gebiet der Sprachverarbeitung, wobei konventionelle analoge Mikrofone mit den neuen digitalen MEMS-Mikrofonen verglichen werden. Besonders werden Sprachaktivitätserkennung, Sprecher-Fahrplan (wer spricht wann), Sprachtrennung und Spracherkennung genau untersucht.

Verschiedene Mikrofon-Arrays wurden dafür mit den digitalen MEMS-Mikrofonen entwickelt und gebaut, und diese wurden zur Aufnahme verschiedener Korpora genutzt, um damit bestehende und neue Algorithmen und Theorien zu testen. Einige dieser Korpora sollen auch im Frühling 2013 der Öffentlichkeit zugänglich gemacht werden.

Es wird gezeigt, dass der verbreitetste Sprachaktivitäts-Algorithmus, der in den meisten Sprecher-Fahrplan-Systemen eingesetzt wird, einer großen Verbesserung bedarf. D.h. MLP-basierte Sprachaktivitätserkennung funktioniert stets besser als die weitverbreitete und meistbenutzte GMM-HMM-basierte Sprachaktivitätserkennung.

Weiterhin stellt diese Doktorarbeit einen Algorithmus vor, der die Anzahl aktiver Sprecher in einer Sitzung erkennt. Dazu benötigt dieser neue Algorithmus Sprachsignale von einem Mikrofon-Array mit definierter Geometrie. Die Kenntnis der Anzahl aktiver Sprecher in einer Sitzung erlaubt die Erzeugung eines verbesserten Sprecher-Fahrplans.

Zum Schluss werden Sprachtrennungsexperimente beschrieben, in denen verschiedene Nachfilter-Verfahren vorgestellt werden und in deren Analyse gezeigt wird, dass die vorgeschlagenen Verfahren mit dem Stand der Technik gleichziehen oder sogar überlegen sind.

Die Ergebnisse der Untersuchungen der Algorithmen und Methoden dieser Doktorarbeit wurden mit Spracherkennung und einfacher MLLR-Adaption ermittelt. Damit wird garantiert, dass die Resultate als Wortfehler-Rate einfach und sicher zu interpretieren sind.

Zusammenfassend zeigt diese Doktorarbeit, dass das erheblich reduzierte Signal-Rausch-Verhältnis der digitalen MEMS-Mikrofone durch einfache und bekannte Adaptierungs-Algorithmen wie MLLR für Sprechertrennung und Spracherkennung kompensiert werden kann und dass Sprachaktivitätserkennung und Sprecher-Fahrplan-Erzeugung durch MEMS-Mikrofone nicht beeinträtigt werden.

Acknowledgements

Carrying out the research necessary for a PhD and writing a dissertation is a one-man job, or so it may appear, but in reality I am grateful to many people who provided generous help and support.

There are too many people to mention individually, so I apologise for any omissions.

First I would like to thank my wife Lynda for being there “durch dick und dünn” and for reading this thesis not once, not twice, but as many times as I asked for her help with my *Swenglish*.

I would also like to thank my supervisors Professor Steve Renals and Dr. Mike Lincoln for their guidance and support. Their encouragement has given me the opportunity to discover the world of research, something which is very different from the life of an engineer, as has proved a learning experience for all concerned.

My thanks also go to my internal examiners Professor Austin Tate and Professor Simon King for their valuable feedback on the annual reviews and to Maurizio Omologo and Simon King for the very interesting discussion and valuable feedback during the Viva.

During my PhD research I was fortunate enough to meet and work with many interesting people. I would first like to thank Friedrich Faubel from Saarland University for our joint collaboration on speech separation. I am also grateful to Xavier Anguera, Iain McCowan, Marijn Huijbregts, Gerald Friedland, David Imseng, Stéphane Dupont and Mark Sinclair who helped by sharing their work with me.

Being a postgraduate student at the University of Edinburgh allowed me to work with people from the CSTR, ILCC, the complete School of Informatics and to benefit from personal communications, workshops, seminars and courses.

I would also like to thank my family for all their support over the years, in particular my mother Rosa and father Xaver.

Finally, I would like to thank my team leaders at EADS IW, Geoff, Simon and Pablo for their support.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Erich Zwyssig*)

“To live only for some future goal is shallow.
It’s the sides of the mountain that sustain life,
not the top.”

Robert M. Pirsig

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The world we live in offers a wealth of recorded material such as lectures, meetings or TV programmes. The volume of these recordings far exceeds what a human being can possibly search or listen to. What is required are recognition technologies that will produce transcripts which are more readable by human beings and more useful to machines, so that key points or answers to specific questions can be obtained with a relatively low input of time and resources.

One means to improve the information content of conversational data is to create a *rich transcript*. A rich transcript, according to Furui et al. [2012c], is

> "a transcript of a recorded event along with meta-data to enrich the word stream with useful information such as identifying speakers, sentence units, proper nouns, speaker locations, etc. As the volume of online media increases and additional, layered content extraction technologies are built, rich transcription has become a critical foundation for delivering extracted content to down-stream applications such as spoken document retrieval, summarization, semantic navigation, speech data mining, and others."

Rich transcriptions (RT) are therefore rich in the sense that a word transcript alone is not sufficient to convey the information from a conversation, but that additional information is required such as *who was present*, *who spoke when*, *who decided what* or *who is assigned to which task*.

Considerable research effort continues to be invested to devise methods that generate rich transcriptions from lectures, meetings and other forms of discussions and presentations. The principle of rich transcription is shown in Figure 1.1.

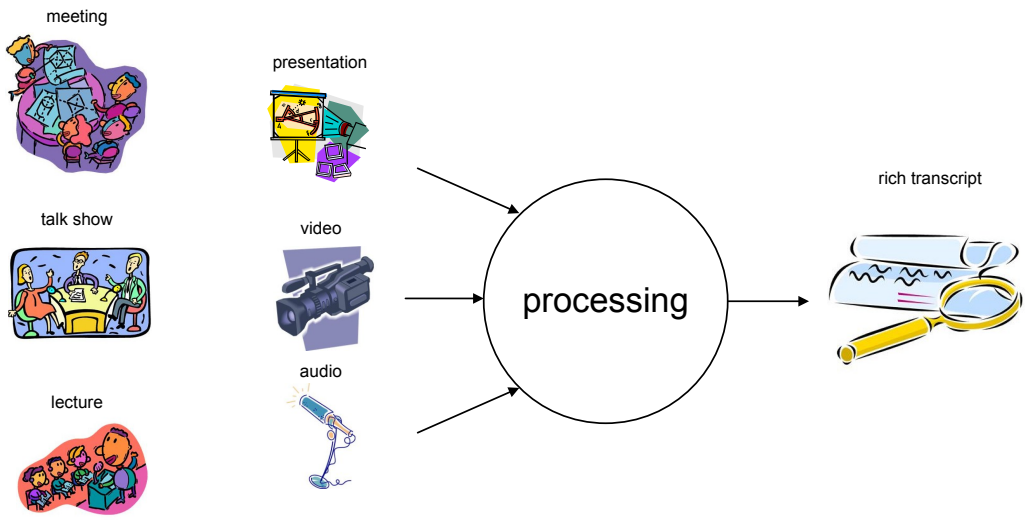


Figure 1.1: Rich transcription (RT)

Ideally the data captured from a meeting, a talk show or a lecture contains the audio and video of the event as well as the information presented, whether through slides or documents either as a file, a series of screenshots or even an additional video.

Traditionally the recordings of these events are performed using sophisticated, purpose-built capture hardware that produces high quality signals. For meeting recording this would contain several close talking and distant microphones, cameras capturing the meeting room from different angles and panoramic cameras on a table in the middle of the participants, and some means to capture whiteboards, blackboards and screens and even note-capturing devices.

The audio, video and presentation information is then processed to generate a rich transcript.

Processing audio signals typically includes components such as audio enhancement, voice activity detection, speaker diarisation, speaker identification, automatic speech recognition and discourse analysis. Audio and speech enhancement is typically done using noise reduction, acoustic beamforming, dereverberation and echo cancelling. The order and combination of these processes is critical and still an open research topic.

Voice (or speech) activity detection and speaker diarisation are also sometimes combined for best performance. Voice activity detection (VAD) aims to find regions of silence, speech and audible non-speech in the conversation. Speaker diarisation is the

task of determining *who spoke when* in a meeting, or meeting recording. Input to the speaker diarisation system is one or more channels of audio. Output is a timetable of *who spoke when* and the number of active speakers present. Generally it is often not known how many people are present and how many people speak or who they are, i.e. there is no primary model for speakers and a speaker diarisation system therefore needs to work in an open-set manner.

Speaker identification tries to determine the ID of the individual speakers, automatic speech recognition generates a text transcript of the recording and discourse analysis aims to post-process this transcript and combine it with all the information present in order to extract a summary, a list of action or a list of decisions from a meeting.

1.1 Contributions

Speech processing using digital MEMS (micro electro-mechanical systems) microphones motivates the research presented in this thesis. Traditionally, speech data used to develop new algorithms and methods and to measure the performance of them is recorded with the best possible audio quality achievable. Robustness of the new algorithms and methods is very often tested by artificially adding noise or reverberation to the (clean) input signal.

The last couple of years have seen a massive shift in the consumer electronics market where almost all (analogue) capacitive microphones are swiftly being replaced with silicon microphones, or MEMS microphones. This, combined with the trend that novel portable devices are getting more and more sophisticated, is leading to the availability of advanced speech processing on portable devices that record sounds (and therefore also speech) using MEMS microphones.

The newly available MEMS microphones have different sound qualities compared to high quality studio recording equipment for example. Most importantly, the SNR (signal-to-noise ratio) performance of the MEMS microphones does not match conventional microphones, leading to audio data with higher noise levels.

The effect of the higher noise level on different speech processing algorithms and in particular the effect of the decreased SNR on speaker diarisation, speech recognition and speech separation is the focus of this research.

The work presented in this thesis has been carried out using digital MEMS microphones. Digital MEMS microphones, compared to analogue ones, contain not only the capacitive sensor and a pre-amplifier, but also an analogue to digital converter (ADC). The performance of the digital and analogue MEMS microphones is identical and any experimental outcome and conclusion presented in this thesis applies for both types of microphones. Details will be presented in Chapter 3.

Starting with a very first prototype with limited functionality, developed as part of my MSc dissertation [Zwyssig, 2009], I have demonstrated the feasibility of a digital MEMS microphone array for speech recognition.

Leading on from this and as part of my PhD, I designed and developed two fully functional circular MEMS microphone arrays, each containing 8 microphones on a diameter of 20 cm and 4 cm. Using these arrays, I carried out several experiments to investigate the effects of the reduced SNR of the digital arrays using noise reduction techniques, beamforming and dereverberation, comparing the digital MEMS microphone arrays (DMMAs) with conventional analogue microphone arrays with identical geometry. The performance of the different arrays was evaluated on the voice activity detection, speaker diarisation, speech separation and recognition tasks.

In the first phase of my research I have analysed different VAD algorithms with regards to their performance on meeting data. I compared well known methods with the most commonly used ones and also with some newly suggested methods and found that neural-network-based (MLP) VAD outperforms the best known GMM-HMM-based method on the NIST RT task.

In the second phase of my research I looked at speaker diarisation. One open research question, which most current state-of-the-art diarisation systems fail to address, is to determine the number of active speakers in a meeting audio stream or recording. Using the TDOA of the individual microphone arrays and the known array geometry I have invented an algorithm which determines the number of active speakers in a recording, aiding the diarisation process significantly. Using the well known NIST RT meeting corpus and diarisation metric I obtained results surpassing state-of-the art diarisation systems, without carrying out segmentation and clustering.

In addition, I recorded a corpus of meetings at the CSTR and carried out experiments to analyse the effect of superdirective beamforming (as against traditional delay-sum beamforming) on the performance of state-of-the-art diarisation systems. The findings

demonstrate that while superdirective beamforming has previously been shown to improve the speech for speech recognition it does not help the diarisation task.

In the third and final phase of my PhD research, I collaborated with Friedrich Faubel from Spoken Language Systems, Saarland University, in which we carried out speech separation experiments. For this I first recorded a novel corpus of single and overlapping speech, the 2012_MMA corpus, motivated by the work presented from the MC-WSJ-AV corpus. Using state-of-the-art speech separation, acoustic beamforming techniques, post-filtering and simple constrained MLLR adaptation, Friedrich Faubel and I have obtained baseline WERs matching the current best performing systems on the distant single speaker task, and demonstrated improved recognition accuracy on the overlapping speech separation and recognition task.

We have also demonstrated that the 2012_MMA corpus is a valuable extension to the existing MC-WSJ-AV corpus, allowing research in speech separation on natural speech using recordings from four different microphone arrays, including digital MEMS microphones. We are currently working with the Linguistic Data Consortium (LDC) to publish the 2012_MMA corpus in 2013.

The following list summarises the contribution of this thesis:

- analysis of different VAD algorithms for meeting diarisation
- design of DMMA.2 and DMMA.3
- analysis of effect of SNR and acoustic beamforming on meeting diarisation
- invention of a novel algorithm to determine the number of active speakers in a meeting recording
- collection of 2012_MMA corpus of single and overlapping speech which is to be released for research
- analysis of speaker localisation and speech separation algorithms on single and overlapping speech

1.2 Publications

The following publication were derived from the work and results presented in this thesis:

- Determining the number of speakers in a meeting using microphone arrays [Zwyssig et al., 2012a]
- On the effect of SNR and superdirective beamforming in speaker diarisation in meetings [Zwyssig et al., 2012b]
- Signal processing method and apparatus [Zwyssig, 2012]
- Recognition of overlapping speech using digital MEMS microphone arrays [Zwyssig et al., 2013]
- The Sheffield Wargames Corpus [Fox et al., 2013]

1.3 Outline

The outline of this thesis is as follows: first, prior work and some background on methods and algorithms used for the research presented in this thesis are reviewed. This includes microphone arrays and acoustic beamforming, automatic speech recognition, distant (and robust) speech recognition, voice activity detection and speaker diarisation, and concludes with speech corpora and open source software available.

This is followed by an introduction to MEMS microphones and a description of the microphone arrays that were developed for the work presented in this thesis after which the 2012_MMA (multi-microphone array) corpus is introduced.

Multiple voice activity detection algorithms and their suitability for meeting diarisation are presented and analysed next, which is followed by the description and analysis of a novel algorithm that is capable of determining the number of active speakers in a meeting recording.

Finally, results from speech separation experiments using single and overlapping speech from digital MEMS and analogue microphone arrays will be presented.

This thesis closes with a chapter summarising and concluding the work presented and proposes possible future work.

Chapter 2

Speech processing for meeting recordings

This chapter presents a review of prior work and gives background information on methods and algorithms used for speech processing of meeting recordings. The aim is to provide a foundation for understanding the methods and algorithms investigated in the following chapters and to give a comprehensive review of speech processing as used for the rich transcription of meetings.

First, microphone arrays and acoustic beamforming are reviewed. This is followed by a discussion of automatic speech recognition and distant speech recognition, including their robustness to acoustically adverse environments. Next, state-of-the-art voice activity detection and speaker diarisation algorithms and systems are reviewed. Last, speech corpora and open source software used for the research in this thesis are presented.

2.1 Microphone arrays and beamforming

Multiple microphones are often used to perform distant recordings of conversations. The signals from many microphones can be combined in an array to perform acoustic beamforming, a versatile method for spatial filtering. The advantage of a microphone array over a single close-talking microphone is the hands-free signal acquisition and the benefit of acoustic beamforming is improved noise robustness. The most basic form of beamforming is delay-sum beamforming and the most basic delay-sum beam-

former is a stereo microphone. Signals from the front (or back) are amplified while signals from the side cancel out. Applying a signal delay to one channel enables the look-direction of the array to be steered. This principle can be applied to any number of microphones, allowing fine-steering of the beam. Beamforming is subject to wave theory and spatial as well as frequency aliasing [Ward et al., 2001]. A typical beamforming directivity pattern is shown in Figure 2.1.

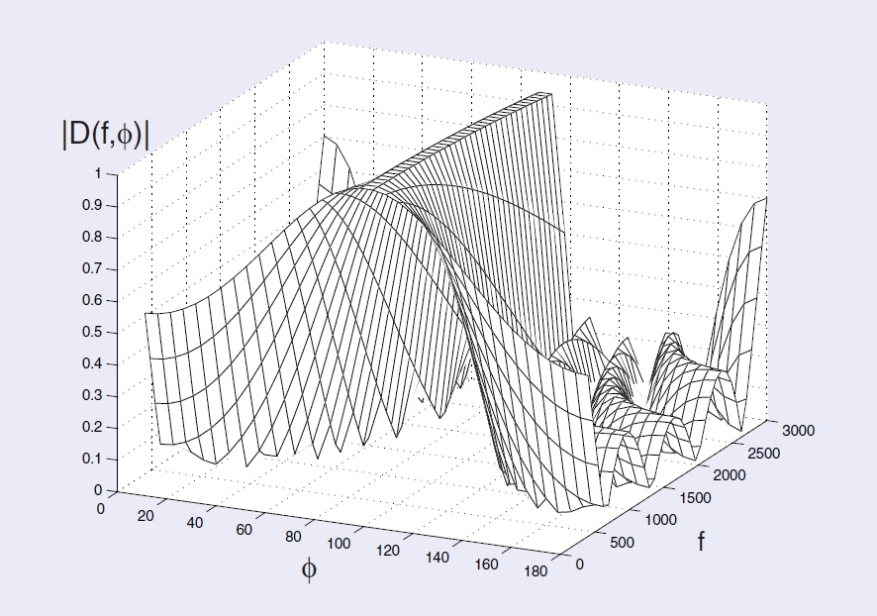


Figure 2.1: Beamforming directivity pattern for 400 Hz $<$ f $<$ 3 kHz (with kind permission of McCowan [2001])

Figure 2.1 shows the absolute value of the directivity $|D(f,\phi)|$ as a function of the input signal angle of arrival ϕ and input frequency f. The beam is not of constant width over the audio frequency range which can cause spatial aliasing problems. It is recommended to use as high a sample rate as possible for the audio signal acquisition in order to get the best-performing beamformer, i.e. 48 kHz sampling rate is preferred over 16 kHz (which is typically used for speech processing).

Note that the layout (e.g. linear, circular) and the distance between adjacent microphones of a microphone array represent one of the most important parameters that characterise its behaviour, i.e. its spatial and frequency response, as shown in Figure 2.1.

Improvements to delay-sum beamforming are adaptive or steered beamforming methods such as filter-sum beamforming, constant-beamwidth beamforming, generalised

sidelobe cancelling (GSC) beamforming and minimum variance distortionless response (MVDR) beamforming [Ward et al., 2001, Bitzer and Simmer, 2001, Elko and Meyer, 2008].

Adaptive beamforming allows the beam to be steered and adapted to different acoustic scenarios. One adaptive beamforming option is to use filter-sum beamforming where the sampled signal delay chain is used as an FIR (finite impulse response) filter before it is summed. For filter-sum-beamforming the weight vectors of the signal delay chain (i.e. the coefficients of the FIR filters) are optimised for a higher directivity and therefore better noise suppression than ordinary delay-sum beamformers. Minimum variance distortionless response (MVDR) beamforming, a special case of filter-sum beamforming achieves super-directivity (or super-gain) by minimising the output power of the array and therefore suppressing the white noise gain [Cox et al., 1987, Bitzer and Simmer, 2001].

An MVDR beamformer therefore optimises its steering vector, beamforming frequency spectrum and spatial characteristic for any input signal that is not considered uniformly distributed white noise. In the case of one active speaker this will lead to a much improved audio signal from the one source.

For speech processing of meetings we cannot assume uniformly distributed noise or fixed speaker position over the entire duration of the meeting. Speaker localisation is therefore a necessary component of meeting analysis.

The primary goal of a speaker (or source) localisation system is accuracy. The speaker position estimates must be reasonably correct and updated frequently for best acoustic output over the entire speech frequency band as the sensitivity of the beam width is much greater for higher frequencies, as illustrated in Figure 2.1 (see DiBiase et al. [2001] for details).

Three general source localisation methods have been extensively investigated based on

- extracting time-difference of arrival (TDOA) information,
- maximising the steered response power (SRP) and
- estimating a high-resolution spectrum (e.g. eigenanalysis used for the MUSIC [Schmidt, 1986] algorithm)

The first two are looked at in greater detail in what follows.

2.1.1 Delay-sum beamforming (DSB)

The way in which acoustic beamforming is performed depends on whether the direction of the sound source and therefore the direction of the beam is specified at the start or whether it needs to be determined from the incoming sound. In the first case the acoustic beam is steered into the desired direction, in the second methods such as GCC, GCC-PHAT, SRP and SRP-PHAT can be used to determine the most likely direction of arrival of the sound(s) after which beam-steering is carried out.

This section first introduces delay-sum beamforming (DSB) and superdirective beamforming (SDB) and then defines GCC (generalised cross correlation), GCC-PHAT (generalised cross correlation with phase transform), SRP (steered response power) and SRP-PHAT (steered response power with phase transform).

Before defining beamforming and SRP, four assumptions in array processing used in this thesis are presented. First, we assume that there is uniform propagation of all sound in all directions in the air. Second, we assume that our algorithms are carried out in a far field environment, i.e. the distance from the source of the sound (and any reflection) to any microphone is much greater than the size of the array and that there is plane wave propagation. The third assumption is that the noise and signal recorded with our arrays have zero mean, that the noise is white and that the signal and noise are uncorrelated. Finally, we assume that the microphones have no coupling, that the sensitivity, self-noise and amplification are matched and that the relative microphone positions are known precisely.

Beamforming provides an elegant way to extract the signal from a desired source through spatial filtering. Under the far-field assumption, for a microphone array with N microphones, a delayed and noise corrupted source signal $x(t)$ is present at each individual microphone. The task of the acoustic beamformer now is to spatially filter a speaker located at direction

$$\mathbf{a} = [\cos\theta\cos\phi \;\; \sin\theta\cos\phi \;\; \sin\phi]^T, \tag{2.1}$$

with θ and ϕ denoting the azimuth and elevation in relation to the array.

The direction of $\mathbf{a}$ translates to time delays

$$\tau_n = \frac{-\mathbf{a}^T \mathbf{m}_n}{c} \tag{2.2}$$

at the microphone positions $\mathbf{m}_n$, $n \in \{1,\ldots,N\}$, where N denotes the number of microphones in the array and c the speed of sound. The delay-sum beamformer then aligns the individual signals in time and sums them.

Delay-sum beamforming is therefore defined as:

$$y(t,\mathbf{a}) = \sum_{n=1}^{N} x_n(t+\tau_n), \tag{2.3}$$

where τ_n are the steering delays for a defined spacial location $\mathbf{a}$.

For fixed beam beamforming the steering delays τ_n are calculated from the desired angle of arrival of the sound source and the microphone positions.

The output of an N-channel delay-sum beamformer can be defined in the frequency domain as

$$Y(\omega,\mathbf{a}) = \sum_{n=1}^{N} X_n(\omega) e^{j\omega\tau_n}, \tag{2.4}$$

where X_n is the Fourier transform of the n-th microphone signal x_n.

Often DSB is described as a multiplication of the input $\mathbf{X}$ by a weight vector $\mathbf{w}$ as

$$Y(\omega,t) = \mathbf{w}^H(\omega) \cdot \mathbf{X}(\omega,t) \tag{2.5}$$

with H denoting the Hermitian conjugate. For the delay-sum beamformer, $\mathbf{w}(\omega)$ is defined as

$$\mathbf{w}(\omega) = \frac{1}{N}\mathbf{v}(\omega) \tag{2.6}$$

where $\mathbf{v}$ denotes the array manifold vector

$$\mathbf{v}_n(\omega) = [e^{-j\omega\tau_{n,1}} \quad \cdots \quad e^{-j\omega\tau_{n,N}}]. \tag{2.7}$$

$\mathbf{v}(\omega)$ translates the time delays $\omega\tau_i$ to phase shifts v_i.

In the presence of white, uncorrelated noise, delay-sum beamforming (DSB) is optimal and gives signal-to-noise ratio (SNR) improvements of 3 dB for every doubling of the number of microphones in the array. Enhancement is achieved by constructively adding the signals from the look direction and suppressing interference from other sources.

2.1.2 Superdirective beamforming (SDB)

One commonly used measure of the performance of beamforming techniques is the array gain G which shows the improvement of the SNR of the array compared to an individual sensor:

$$G = \frac{SNR_{array}}{SNR_{sensor}} \tag{2.8}$$

By optimising the array gain, more sophisticated methods, known as superdirective beamformers, can be used to improve the beamformer's directional selectivity, further cancelling undesired sources.

A number of superdirective beamformers found their application in speech processing [Bitzer and Simmer, 2001, Elko and Meyer, 2008]. Examples include filter-sum, constant-beamwidth, generalised sidelobe cancelling (GSC) and minimum variance distortionless response (MVDR) beamformers, each being differentiated by the method employed to optimise G.

As shown in Figure 2.1 above, acoustic beamforming is subject to wave theory and spatial as well as frequency aliasing. For constant-beamwidth beamforming the physical placement of the microphones has to be such that spatial aliasing can be overcome. For a fixed beamwidth from 500 Hz to 8 kHz this requires an endfire[1] microphone array with a distance of 2–16 cm in between the microphones and a total width of over 1.5 m (see Elko and Meyer [2008] for details) which is infeasible for mobile meeting recording.

[1] An endfire array is an array of N (preferably equidistant) microphones placed in line with the direction of the arriving sound, compared to a broadside array, where the (equidistant) microphones are placed perpendicular to the direction of the sound.

Generalised sidelobe cancelling beamformers aim to improve their output by steering a null into the direction of an interfering source. This requires knowledge of the location of the undesired sound source, something which cannot be assumed for this research (see Elko and Meyer [2008] for details). In addition, working with speech signals, the statistics of the desired source are not precisely known or are highly non-stationary.

MVDR beamforming, also know as Capon's beamforming, is a well known and extensively used beamforming technique that offers a good spectral characteristic of the output and is therefore well suited to acoustic beamforming and speech enhancement [Bitzer and Simmer, 2001, Elko and Meyer, 2008]. MVDR beamforming is looked at in detail in what follows.

The aim of MVDR beamforming is to minimise the power of the output signal of the array while maintaining unity gain in the look direction and also maximising the white noise gain. MVDR beamforming is based on filter-delay-sum beamforming and its frequency domain output signal Y is defined as:

$$Y(e^{j\Omega}) = \sum_{m=0}^{M-1} {w_m}^*(e^{j\Omega}) X_m(e^{j\Omega}) = \mathbf{w}^H \mathbf{X}, \tag{2.9}$$

where $w_m(e^{j\Omega})$ denotes the filter coefficients of the beamformer for sensor m at frequency Ω, $\mathbf{X}_m(e^{j\Omega})$ are the microphone input signals and $[\cdot]^H$ denotes the matrix transpose conjugate.

Cook et al. [1955] proposed to minimise the total output power $Y(e^{j\Omega})$ under the assumption of a diffuse noise field in order to optimise spatial filtering with respect to reverberant environments. This leads to the superdirective beamformer whose weight vector is:

$$\mathbf{w}(\omega) = \frac{T^{-1}(\omega)\mathbf{v}_\omega)}{\mathbf{v}^H(\omega)T^{-1}(\omega)\mathbf{v}(\omega)} \ . \tag{2.10}$$

$T_{i,j}(\omega)$ denotes the coherence of a spherically isotropic noise field with

$$T_{i,j}(\omega) = \mathrm{sinc}(\frac{\omega}{c}\|\mathbf{m}_i - \mathbf{m}_j\|), \quad i,j \in \{1,\ldots,N\}. \tag{2.11}$$

$\|\mathbf{m}_i - \mathbf{m}_j\|$ is the absolute magnitude matrix of the microphone inter-distances.

The output of the acoustic beamformer, i.e. the spatially filtered speech signal $y(t)$, is recovered from $Y(e^{j\Omega})$ using inverse Fourier transformation followed by overlap-and-add (see Bitzer and Simmer [2001] for details).

2.1.3 TDOA and GCC-PHAT

Beamforming as described so far requires knowledge of the direction of the sound. For adaptive beamforming the steering delays can be calculated using GCC, GCC-PHAT, SRP or SRP-PHAT, which are defined below.

The most important parameter for delay-sum beamforming is the direction of arrival (DOA) of the speech signal. Knowing the DOA, the delay-sum beamformer coefficients can be calculated and the beam steered in the direction of the speaker. An established method to determine the DOA is to use generalised cross correlation with phase transform (GCC-PHAT) to obtain the time difference of arrival (TDOA) of each microphone pair of the microphone array.

Note that while GCC can be used for DOA estimation this is rarely done in practise due to poor performance of GCC alone compared to GCC-PHAT [Omologo and Svaizer, 1994].

For DOA estimation the array microphone with the highest energy level is generally used as the reference. If a signal is detected on that microphone then, using GCC-PHAT, the TDOAs can be calculated and the beam steered (see Knapp and Carter [1976] and Brandstein and Silverman [1997] for details).

The generalised cross correlation with phase transform of two signals is defined as

$$\hat{G}_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|}, \tag{2.12}$$

where $x_i(t)$ and $x_j(t)$ are two discrete signals in the time domain and $X_i(f)$ and $X_j(f)$ their DFT (discrete Fourier transform); $X_i(f)X_j(f)$ is the element-by-element product of $X_i(f)$ and $X_j(f)$; $[X(f)]^*$ is the conjugate complex of $X(f)$, and $|f(x)|$ is the amplitude of the complex number.

The TDOA $\hat{d}_{PHAT}(i,j)$ of the two signals $x_i(t)$ and $x_j(t)$ is estimated as the maximum

value of the inverse Fourier transform of $\hat{G}_{PHAT}$, i.e.

$$\hat{d}_{PHAT}(i,j) = \arg\max_{d}(\hat{R}_{PHAT}(d)), \tag{2.13}$$

where $\hat{R}_{PHAT}$ is the inverse Fourier transform of $\hat{G}_{PHAT}$. Note that a valid $\hat{d}$ ranges from the minimum to the maximum possible delay determined by the distance between the microphone pair. In practise though, due to noise, $\arg\max(\hat{R}_{PHAT})$ might well be outside the valid range of $\hat{d}$. This can be detected and corrected if the distance of the two microphones is known.

GCC-PHAT as a source localisation algorithm performs time-delay estimation (TDE) in pairs of microphones at very little computational cost. Unfortunately, pairwise TDE techniques suffer considerably in reverberant environments due to the increase of peaks in the cross correlation function from the individual echoes, leading to a lower direct-to-reverberant ratio (DDR), i.e. the ratio between the energy of the direct sound and reverberation. This problem can partially be overcome by increasing the amount of data, i.e. the input audio segment length. This again has practical limitations, as for best beamforming and therefore acoustic output, the look direction of the acoustic beamformer is best tracked (or updated) ten times a second or more (for details see DiBiase [2000], Section 1.1 and reference therein). Using an error-prone pair-wise TDE method is therefore not ideal for sound source localisation.

2.1.4 SRP and SRP-PHAT

Steered response power (SRP) and steered response power with the phase transform (SRP-PHAT) are much better suited to sound source localisation than TDE-based methods such as GCC and GCC-PHAT.

For SRP, a beamformer is used to search over a predefined spatial region looking for a peak (or peaks) of higher signal power. This is computationally much more expensive than pairwise methods but it combines the signal from all inputs and not just two at a time. Using SRP-based sound source localisation it is possible to locate one or more speakers simultaneously on much shorter speech segments compared to TDE-based techniques. Multiple active speakers appear as multiple peaks in the SRP map [Do and Silverman, 2008].

For source localisation using the SRP, a simple delay-sum beamformer searches a predefined spatial region looking for peaks in the power spectrum. The SRP can therefore be defined as:

$$P(\mathbf{q}) = \int_{-\infty}^{+\infty} |Y(\omega)|^2 \mathrm{d}\omega \tag{2.14}$$

The location estimate is then found from

$$\hat{\mathbf{q}}_s = \arg\max_{\mathbf{q}} P(\mathbf{q}) \tag{2.15}$$

In reality the integral of Equation 2.14 is not calculated as such, but a power map is generated over the space required. For meeting recordings using a circular array, an SRP map would be calculated on the circular plane with a reasonably good azimuth resolution of e.g. 2° and on the elevation from 0° to 30° with a resolution of e.g. 5°. An SRP map with these requirements involves creating 1267 beamformers, applying them to the acoustic input, calculating the SRP of each of them and finding the maximum SRP(s) which corresponds to the speaker localisation(s). To complete the process, SDB is applied in the desired direction(s) after which some means of postfiltering is performed.

The SRP as defined in Equation 2.14 is the real value of the power of a 3-D spatial vector obtained by steering a delay-sum beamformer, and the location estimates $\hat{\mathbf{q}}_s$ should ideally be the point sources of our active speakers even under very noisy and highly reverberant conditions.

However, under adverse conditions, DiBiase showed that applying the weighting function from the GCC-PHAT (see Equation 2.12) to the SRP sharpens the peaks in the phase transform and therefore the steered response power, making SRP-PHAT a superior localisation method compared to SRP, especially under higher noise conditions (see DiBiase [2000] for details).

DiBiase suggested steering a DSB into each possible direction $\mathbf{a}_1(\phi,\theta)$ and then calculating the total power at the beamformer output as

$$P\{\boldsymbol{a}(\phi,\theta),t\} = \int_{-\infty}^{\infty} \|\boldsymbol{w}^H(\omega)\cdot\tilde{\boldsymbol{X}}(\omega,t)\|^2 d\omega \tag{2.16}$$

where

$$\tilde{X}(\omega,t) = \frac{X(\omega,t)}{|X(\omega,t)|}. \tag{2.17}$$

$\tilde{X}_i(\omega,t)$ is a whitened version of $X_i(\omega,t)$ and $\boldsymbol{w}^H(\omega)$ for the DSB can be calculated as per Equations 2.6 and 2.7.

2.1.5 Postfiltering

MVDR beamforming theoretically provides the optimum solution for a given sound field, but only from a narrowband point of view. Speech, however, is a broadband signal and postfiltering of the beamformed output has been found to bring significant improvements [Simmer et al., 2001]. Figure 2.2 shows the typical filtering processes applied to speech signals in the pre-processing stage, i.e. before voice activity detection, diarisation and speech recognition.

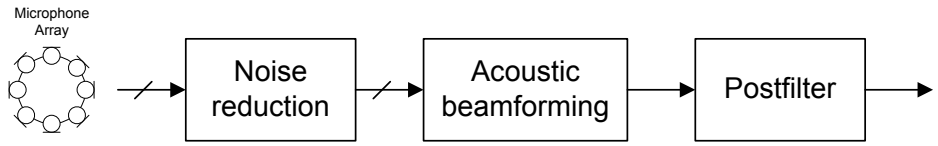


Figure 2.2: Flow diagram for typical speech filtering

It is common to apply noise reduction to the single microphone channels after which acoustic beamforming is performed in which the multiple channels are reduced to one channel only (see e.g. Friedland et al. [2012], Huijbregts and van Leeuwen [2012], Lincoln et al. [2005], Zwyssig et al. [2010], etc.). This single audio channel is then further enhanced by postfiltering.

The best possible linear filter for speech enhancement is the Wiener filter which is also very effective at reducing the input signal noise before sound source localisation and beamforming.

Assume we receive a distorted input signal $y(n) = x(n) + v(n)$ where $x(n)$ is a zero-mean clean input signal and $v(n)$ is a zero-mean noise signal (white or coloured but uncorrelated to $x(n)$). Using Wiener filtering we try to estimate $x(n)$ from $y(n)$. The output of the Wiener filter is therefore defined as $\hat{x}(n)$ and the error signal $e_x(n)$ is

given by

$$e_x(n) = x(n) - \hat{x}(n) = x(n) - \mathbf{h}^T y(n), \tag{2.18}$$

where $(.)^T$ denotes the transpose of a vector or matrix and $\mathbf{h}$ is the Wiener filter which is optimised to minimise e_x. Reducing the error signal $e_x(n)$ is a difficult problem as neither the clean input $x(n)$ signal nor the noise $v(n)$ are known.

Possible solutions are:

- the noise statistics are known a priori
- using a simple speech activity detection algorithm, regions of pure noise can be identified
- knowing the statistics of speech, pure noise regions can be identified

Implementation of noise reduction is often done using sub-band processing and noise-only regions are detected using inactive bands. $\mathbf{h}$ is then determined using the famous Wiener-Hopf equation. For details please refer to Simmer et al. [2001], Chen et al. [2006], Zelinski [1988] and the references therein.

Postfiltering can be carried out using different filter types, some of which modify the phase of the input signal. Good TDOA estimation using GCC-PHAT and SRP-PHAT relies on accurate phase information (cf. Section 2.1.3 and 2.1.4) . The research presented in this thesis follows the signal flow as used by Lincoln et al. [2005] and Zwyssig et al. [2010] in order to change as few parameters as possible and to compare the results.

It is a common choice in the speaker diarisation community to apply noise reduction before performing acoustic beamforming, particularly as the recordings in the NIST RT evaluations were carried out with microphones located arbitrarily, generally far from each other in a noisy meeting room environment (see e.g. Friedland et al. [2012] and Huijbregts and van Leeuwen [2012]).

Note also that in the deliverable D4.6 of the AMI project [Renals, 2010] it is stated that “Wiener filtering is applied to each channel to remove the stationary background noise”, i.e. post-filtering is applied before acoustic beamforming.

The experiments presented in this thesis follow the process used for the NIST RT evaluations and AMI/DA recordings in order to be compatible with these. Unfortunately

this might be suboptimal for good TDOA estimation, as Wiener filter are non-linear phase filters which modify the phase of the signal.

Note also that sample rate reduction can be applied before or after noise reduction. If sample rate conversion was required for the research presented in this thesis then either 'sox'[2] or the Matlab™ 'resample' programme were used. In both cases the settings were chosen so that only linear filtering was applied to the audio signal, therefore guaranteeing that the phase of the audio signals remained unchanged.

2.2 Automatic speech recognition (ASR)

The task of a speech recogniser is to find the most probable sequence of words $\hat{W}$ given a series of acoustic observations O. This is a statistical classification problem where a decoder tries to find the most likely sequence of words $\hat{W}$ given an observations O, as defined in Equation 2.19. However, Equation 2.19 is difficult to model but, using Bayes' theorem, it can be reformulated as Equation 2.20. When searching for $\arg\max(.)$, $P(O)$ can be omitted because it is constant for a given series of O, resulting in Equation 2.21.

$$\hat{W} = \arg\max_{W} P(W|O) \tag{2.19}$$

$$= \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} \tag{2.20}$$

$$= \arg\max_{W} P(O|W)P(W) \tag{2.21}$$

The most probable sequence of words $\hat{W}$ can therefore be calculated from $P(O|W)$, the probability of a feature observation given a word sequence W and $P(W)$, the probability of a word (within the context of the sequence of words). Therefore, the most probable sequence of words $\hat{W}$ is determined from the acoustic model $P(O|W)$ and the language model $P(W)$.

The success of state-of-the-art automatic speech recognition (ASR) was achieved through the application of hidden Markov models (HMMs) and Gaussian mixture models (GMMs) to finding $P(O|W)$, i.e. to the acoustic modelling task.

[2] http://sox.sourceforge.net

A review of GMMs, HMMs, how to decode and train them as well as the performance of state-of-the-art GMM-HMM-based large-vocabulary ASR systems is given next. See also Rabiner and Juang [1986] for a detailed introduction to HMMs and Young [2008] for a more up-to-date review of HMMs and current speech recognition technologies.

A schematic overview of ASR is shown in Figure 2.3.

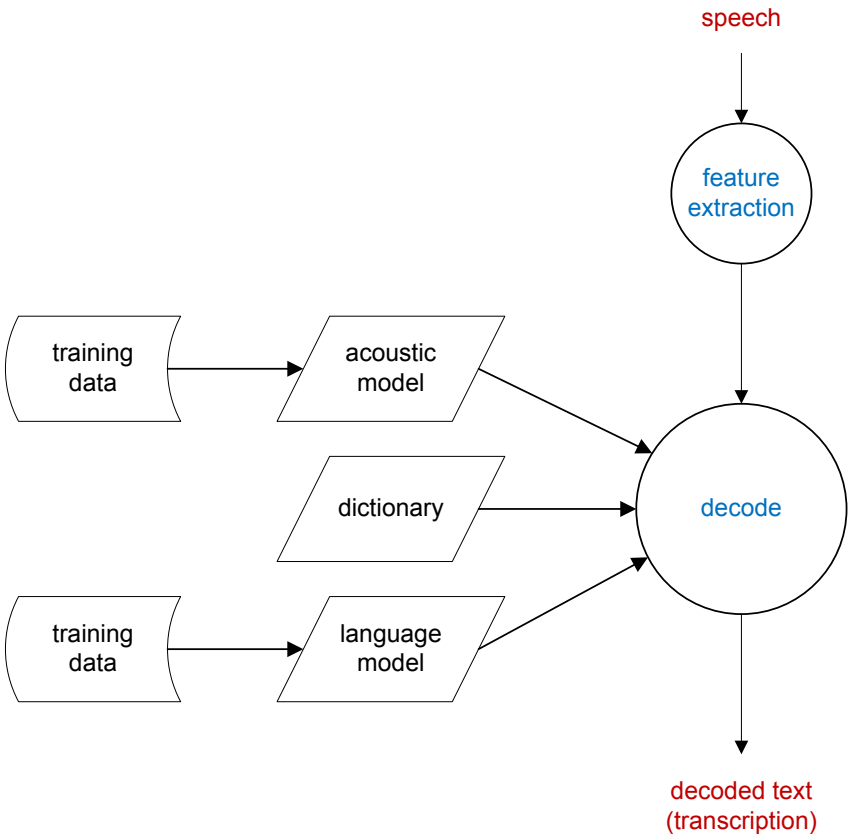


Figure 2.3: Schematic overview of automatic speech recognition

Input to the ASR system is speech, output is the decoded text, or transcription. The incoming speech is analysed and enhanced by means of speech feature extraction, or generation. In the first step, speech features are extracted from the incoming audio signal with two main aims: (1) to reduce the redundancy of the incoming signal to extract the essential information and (2), to adapt the incoming features to best match the training data for best recognition performance.

Inputs to the decoding task are not only the enhanced speech features but also an acoustic model, a dictionary and a language model. Two principal methods to carry out the

acoustic modelling task are to create the GMM-HMM model network dynamically (from the acoustic and language model) or to statically compile the network beforehand using the finite state transducer (FST) approach (see e.g. Mohri et al. [2002] for details).

The sections below describe feature generation, the hidden Markov model and decoding and training hidden Markov models followed by language models. Afterwards the detailed implementation of phone models and how they are adapted for working under adverse conditions is presented. The error metric to measure the performance of an ASR system is then defined followed by a review of state-of-the-art ASR.

2.2.1 Speech feature generation

The speech for speech recognition systems is usually sampled at 16 kHz in 16-bit values using the linear pulse-code modulation (LPCM) format. The highly redundant and correlated nature of waveform data means that it is not well suited for practical speech processing. Speech therefore needs to be compacted and encoded, i.e. its features extracted in order to obtain the best discriminative features while retaining only useful information.

Two schemes dominate current speech systems: these use MFCCs (Mel frequency cepstral coefficients) and PLP (perceptual linear prediction) coefficients. The principle for generating MFCC and PLP features is shown in Figure 2.4.

Figure 2.4 actually shows three different feature extraction schemes, MFCC, PLP and MF-PLP feature generation.

For MFCC generation, a time domain sampled signal $o[n]$ is transformed into the frequency domain using discrete Fourier transformation (DFT) giving complex values $O[k]$. The absolute powered values of $O[k]$ are binned in Mel-frequency bands (typically 12 + energy) and the logarithmic values are decorrelated using the inverse DFT, IDFT. Speech processing (for both MFCCs and PLPs) typically uses the Cepstral coefficients $c[n]$ and their first and second order derivatives Δ and $\Delta\Delta$. A detailed description can be found in Jurafsky and Martin [2009a].

PLP coefficient generation differs in principle from MFCC in that Hermansky [1990] derives

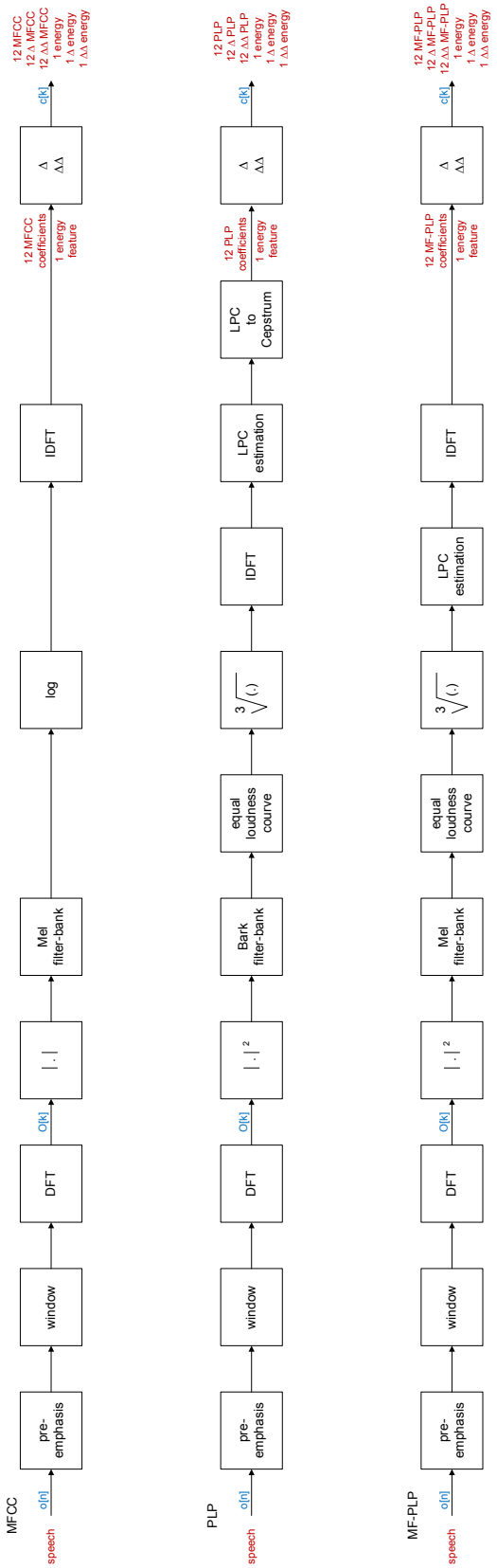


Figure 2.4: Speech feature generation

> "... three concepts from the psychophysics of [human] hearing [to get] an estimate of the auditory spectrum: (1) the critical-band spectral resolution, (2) the equal loudness curve and (3) the intensity-loudness power law. The auditory spectrum is then approximated by an autoregressive all-pole model."

For PLP coefficient generation, the Bark scale is used instead of the Mel scale (satisfying 1 above) after which the coefficients are weighted by an equal-loudness curve (2) and then compressed by taking the cubic root (3). Following this, the inverse Fourier transform is calculated and linear prediction coefficients (LPC) are estimated from the resulting auditory spectrum, i.e. all-pole filter coefficients are approximated from the spectrum. The LPC coefficients are then converted to cepstral coefficients. A detailed description can be found in Hermansky [1990].

The HTK toolkit implements PLP coefficient generation slightly differently, shown as MF-PLP in Figure 2.4. For MF-PLP coefficient generation, the Mel frequency bands are used and an equal-loudness curve and cubic root compression applied. LPC estimation is then carried out in the frequency domain, and cepstral coefficients are generated from the LPCs using the DCT. Please refer to the HTKBook [Cambridge University Engineering Department (CUED), 2012] for details.

Feature generation is still an active area of research and multiple comparative studies have been carried out to determine which features perform best on what task (see e.g. Davis and Mermelstein [1980], Woodland et al. [1997], Beyerlein et al. [2002] and Zolnay et al. [2005]. Young [2008] summarises these as:

> " in practice, PLP can give small improvements over MFCCs, especially in noisy environments and hence it is the preferred encoding [method] for many systems."

I carried out experiments using MFCC and MF-PLP coefficients and was unable to measure a significant difference. Experimental results presented in this thesis use MFCCs if not stated otherwise, as these are the most commonly used speech features for diarisation (see e.g. Friedland et al. [2012], Huijbregts and van Leeuwen [2012], etc.) and in order to compare all my results with those presented previously (see e.g. Lincoln et al. [2005], Zwyssig et al. [2010], etc.).

Speech features can be greatly enhanced by applying normalisation, i.e. by adapting them to the features used for training the GMM-HMMs. Two well-known and established feature enhancing methods are cepstral mean and variance normalisation, and

vocal-tract-length normalisation (VTLN).

For mean normalisation, the average feature value is removed, leading to the test data being more similar to the training data and therefore to improved speech recognition accuracy. For cepstral variance normalisation, each feature coefficient is set to have unit variance, leading to reduced sensitivity to additive noise [Young, 2008]. In a real application the means and variances are normalised over the longest possible speech segment for which the speaker and environment conditions are constant.

Vocal-tract-length normalisation (VTLN) aims to compensate for pitch (F0) and formant frequency shifts observable in between e.g. male and female speech by compressing or expanding the frequency scale. Assuming the HMMs of a speech recogniser are trained with male-only speech, recognising female speech can be significantly improved by applying VTLN [Lee and Rose, 1996].

Please see Young [2008] and references therein for a detailed description of speech feature adaptation.

2.2.2 Hidden Markov model

As mentioned above, the success of state-of-the-art ASR was achieved with the application of hidden Markov models (HMMs) and Gaussian mixture models (GMMs) to decode the text from speech. A hidden Markov model is a special case of a finite state machine (FSM, or FSA - finite state automaton) containing a set of states Q, their transition probabilities a_{ij} (representing the probability of moving from state i to state j) and a sequence of observation likelihoods $b_j(o_t)$, expressing the probability of the observation $O = o_1, o_2, ..., o_T$ being generated from the state q_i, as shown in Figure 2.5.

Please note that for the HTK tool the entry and exit states of an HMM for speech recognition are non-emitting. This allows easy concatenation of phone models to make word models.

The output observation distributions $b_j(o_t)$ are usually mixtures of Gaussians (i.e. Gaussian mixture models – GMMs). The probability of an observation o, given a GMM λ is therefore defined as

$$p(\boldsymbol{o}|\lambda) = \sum_{m=1}^{M} w_m \, g(\boldsymbol{o}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m), \tag{2.22}$$

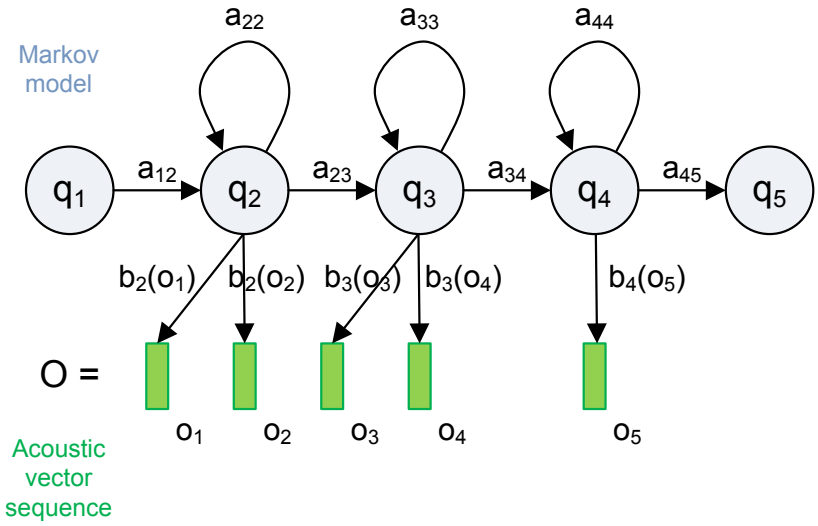


Figure 2.5: HMM-based phone model $\boldsymbol{Q}$ and acoustic observation vector $\boldsymbol{O}$

where $\boldsymbol{o}$ is a D-dimensional data vector (e.g. 13 MFCCs plus first and second order derivatives Δ and $\Delta\Delta$) with M Gaussian densities; $\boldsymbol{w}$ are mixture weights with $\sum_{m=1}^{M} w_m = 1$ and $g(\boldsymbol{o}|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$ are Gaussian component densities.

HMMs make two assumptions. First, that the probability of a state depends only on the previous state, i.e.

$$P(q_i|q_1 \ldots q_{i-1}) = P(q_i|q_{i-1}) \tag{2.23}$$

and second, that the probability b_i of an observation o_t depends only on the state q_i that produced the observation.

The three principal problems that now need to be addressed are:

- likelihood: given an HMM $\Lambda = (A,B)$ and an observation sequence O, find the likelihood $P(O|\Lambda)$, [3]
- acoustic modelling: given an observation sequence O and an HMM $\Lambda = (A,B)$, find the best hidden state sequence Q,
- learning: given an observation sequence O and the set of states in the HMM, learn the HMM parameters w, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.

[3] A is the matrix of all a_{ij} coefficients of the HMM and B is the matrix of all b_j coefficients.

An efficient way of calculating the likelihood of an HMM is by using the forward algorithm. Decoding is carried out using the Viterbi algorithm and training is best done using a special case of the expectation-maximisation (EM) algorithm, the forward-backward algorithm (or Baum-Welch algorithm). These algorithms are explained in detail in what follows (see also Jurafsky and Martin [2009b] for examples and further references).

Please note that in practise the covariances $\boldsymbol{\sigma}_i$ are usually constrained to be diagonal in order to reduce the computational effort necessary to perform speech recognition, since the dimensionality of the acoustic input vector $\boldsymbol{o}$ can be relatively high.

2.2.3 Viterbi decoding

The Viterbi algorithm is an efficient decoding algorithm applied to HMMs which finds the optimal sequence of hidden states. Given an observation sequence O, it returns the state path through the HMM with the maximum likelihood

$$v_t(j) = \max_{i-1}^{N} v_{t-1}(i) a_{ij} b_j(o_t), \tag{2.24}$$

where $v_{t-1}(i)$ is the previous Viterbi probability, a_{ij} is the transition probability and $b_j(o_t)$ is the state-observation likelihood.

The applied use of the Viterbi algorithm is explained in the context of HMMs and ASR in Section 2.2.5 below. Note that token-passing is used for Viterbi decoding in order to return the state path through the HMMs. In practise the N-best paths – and not just the best possible output – are returned in the form of a word lattice.

2.2.4 EM algorithm

The expectation-maximisation (EM) algorithm is a forward-backward algorithm to train HMMs with the aim of learning the parameters of an HMM given the observation sequence.

As the name implies, optimisation of the HMM parameters is carried out in two steps, the expectation step and the maximisation step. First the HMM parameters a_{ij} and b_j are initialised (either with an arbitrary or a random number). Then, in the expectation step, the expected values of the state occupancies are calculated using the forward (Equation 2.25) and backward algorithms (Equation 2.26).

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t) \tag{2.25}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \tag{2.26}$$

Using the above likelihoods $\alpha_t(j)$ and $\beta_t(i)$, the state occupancy counts $\gamma_t(j)$ and the expected state transition counts $\xi(i,j)$ are calculated as follows:

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{P(O|\lambda)} \qquad \forall \quad t \text{ and } j; \tag{2.27}$$

$$\xi(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\alpha_T(N)} \qquad \forall \quad t, i, \text{ and } j; \tag{2.28}$$

where $P(O|\lambda)$ is the forward (or backward) probability of the complete utterance and $\alpha_T(N)$ is the observation probability of the complete utterance.

In the maximisation step, the HMM parameters a_{ij} and b_j are recalculated, i.e. maximised, as follows:

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} \xi_t(i,j)}, \tag{2.29}$$

$$b_j(v_k) = \frac{\sum_{t=1 \text{s.t.} O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, \tag{2.30}$$

where "$t = 1\text{s.t.}O_t = v_k$" means sum over all t for which the observation at time t was v_k.

The EM steps are repeated until the model converges.

The applied use of the EM algorithm is explained in the context of HMMs and ASR in Section 2.2.5 below.

2.2.5 HMM acoustic models

The power of speech recognition using HMMs (that is GMM-HMMs) lies in their ability to model sentences, words or phones. HMM-based speech recognition started with whole word small vocabulary applications such as digit recognition. Increasing amounts of training data and computing power today allow continuous speech recognition on broadcast news programmes (BN), spontaneous telephone conversations (CTS) or meeting recordings. The performance of ASR on these tasks will be reviewed in Section 2.2.8.

As noted above, the aim of GMM-HMM-based speech recognition is to find the likelihood of a sequence of acoustic observations given an acoustic model. Assuming a ASR which recognises 60,000 words, the number of possible sentences would be near infinite while the number of words is 60,000. Working with word models would require the presence of many examples of every word to train the models, which in practise is not feasible. A feasible solution is to build GMM-HMMs of the individual phones. For English that would require approximately 40 models, 25 vowel and 15 consonant models.

A complicating factor in spoken language is that the pronunciation of a phone changes in the context of a word, such as 'oo' in mood versus cool. The realisation of 'oo' is different depending on the preceding and following consonants.

The solution to the problem is to use a context-dependent phone model, i.e. a triphone model, as illustrated in Figure 2.6.

In this example the words 'stop that' are first split into phones and then triphones, with x-q+y denoting a triphone with q being the target phone, x the preceding phone and

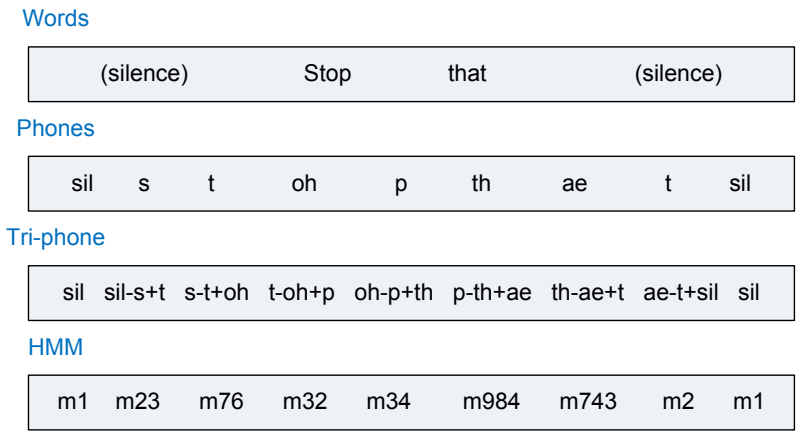


Figure 2.6: Context-dependent phone modelling

y the following phone. Note that triphone context can (and should) span across word boundaries, therefore allowing modelling of e.g. phone suppression or deletion such as ‘p’ which is unreleased by the following consonant ‘t’ in ‘stop that’.

If the English language requires 40 phone models for good recognition, then N^3 (= 64,000) triphones would need to be trained, again leading to the problem of data sparsity, i.e. many examples of each triphone are required to train their GMM-HMMs.

The solution is the formation of tied-state phone models, as illustrated in Figure 2.7.

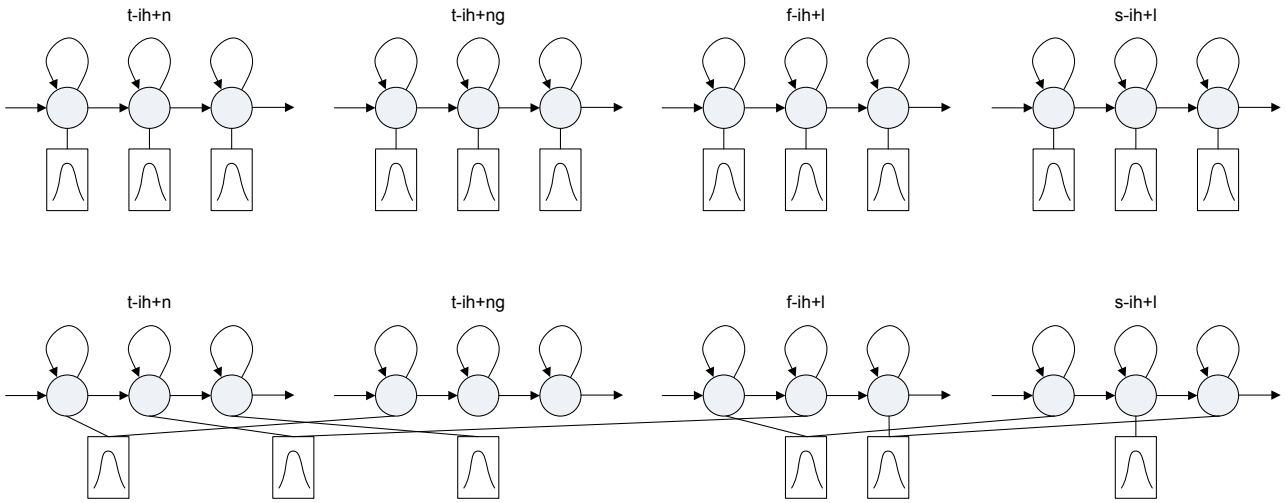


Figure 2.7: Formation of tied-state phone model

Similar states – as frequently occur for the biphone of each triphone – can share the same GMM-HMM, as is for example possible with t-ih in t-ih+n (tin) and t-ih-ng

(ting)[4]. In practise the total number of triphone models can be significantly reduced for ASR from tens of thousands of triphone sharing models to a few thousand tied-state models.

The partitioning of the triphone models to tied-state models is usually carried out using decision tree clustering, as shown in Figure 2.8.

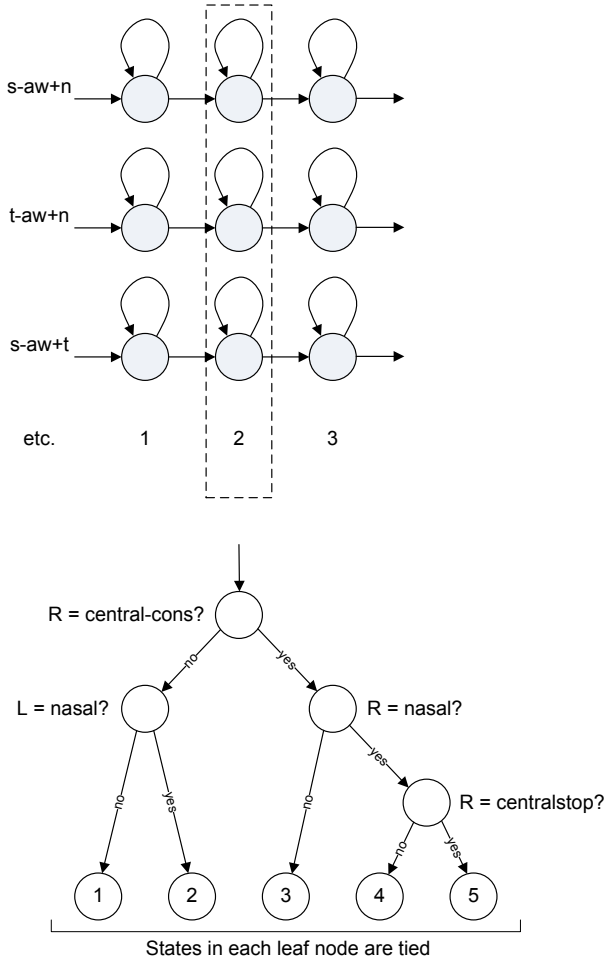


Figure 2.8: Decision tree clustering

Tied-state models are well suited to handling data sparsity when training the triphone GMM-HMMs, while decision-tree-based clustering established itself as the most efficient means to determine which triphones to tie in similar models [Woodland et al., 1994].

In the proposed method, triphones are grouped for the same base phone and, using a

[4]See `http://www.speech.cs.cmu.edu/cgi-bin/cmudict` to look up the pronunciation of a word or a sentence as defined in the CMUdict

predefined set of phonetically driven decisions, the tree is grown, i.e. split from the root by maximising the log likelihood of the left and right side data. The aim is to have the best balanced tree and enough data in each leaf to train the GMM-HMMs. The best possible tree is grown from the top down using every question left in the pool (from which previously used questions are removed).

2.2.6 Language model

As stated in Equation 2.21, the probability of a sequence of words, given a sequence of acoustic observations, is calculated from the acoustic model and the language model. The language model takes the output of the decoder, i.e. the recognised lattice, and calculates the probability of a word W being in that position of the lattice given the preceding words in the lattice. The prior probability of a word sequence $W = w_1, ..., w_k$ is given by:

$$P(W) = \prod_{k=1}^{K} p(w_k | w_{k-1}, w_{k-2}, ..., w_1) \tag{2.31}$$

Equation 2.31 implies that the probability $P(W)$ is calculated depending on all previous words $W = w_1, ..., w_k$. For practical purposes the word history is truncated to $N-1$ words, i.e.

$$P(W) = \prod_{k=1}^{K} p(w_k | w_{k-1}, w_{k-2}, ..., w_{k-N-1}),, \tag{2.32}$$

where N is typically in the range 2–4. The set of $P(W)$ is called a language model (LM). N-grams as defined in Equation 2.32 are one of the most important tools in speech and language processing and are not only used for speech recognition but also for machine translation and spelling checkers for example[5].

Three main problems for language modelling will now be looked at in more detail. The main problem of LM is the same as for acoustic modelling, i.e. data sparsity. Training a 3-gram LM for a ASR system with a vocabulary of 60,000 words requires calculating $P(W)$ for $216 \cdot 10^{12}$ 3-grams. However large the training set for modelling

[5] 3 or 4-grams are typically used in state-of-the-art ASR, while higher-order N-grams are employed in state-of-the-art machine translation

is, most 3-grams will not be seen in a training set and therefore acquire a probability of zero, thereby obliterating $P(W)$ in the recognition process.

In order to avoid null probabilities, smoothing is required. For smoothing, 'unseen' N-grams are assigned some discounted probability mass taken from the 'seen' N-grams. Turing-Good and Kneser-Ney smoothing are two well-known and established smoothing methods used in state-of-the-art ASR systems (for details see Young [2008] and references therein).

Another problem of LMs, as per Rosenfeld [2000], is

> "... brittleness across domains: Current language models are extremely sensitive to changes in the style, topic or genre of the text on which they are trained. For example, to model casual phone conversations, one is much better off using 2 million words of transcripts from such conversations than using 140 million words of transcripts from TV and radio news broadcasts. This effect is quite strong even for changes that seem trivial to a human: a language model trained on Dow-Jones newswire text will see its perplexity doubled when applied to the very similar Associated Press newswire text from the same time period."

The same applies to meeting domain data, where unfortunately only little annotated data is available to train a language model.

The third problem for acoustic and language model training is out-of-vocabulary (OOV) words. If a recogniser does not contain a word in its dictionary and acoustic model, then it cannot decode it, but will come up with the most likely alternative. Adding new words to a recogniser can be quite complex. First, the N-gram probabilities for the new words need to be calculated. Then, depending on the decoder architecture, these new words need to be made available for the dynamic decoding or to regenerate the FSTs. In the case of FSTs this requires restructuring and optimising the complete FST, a time-consuming task that is not feasible for online ASR. One possible solution to this problem is on-the-fly composition of the FST [Hori et al., 2007]. Modifying a language model efficiently for online processing is an open research problem.

2.2.7 Adaptation

Speech recognition using GMM-HMMs is formulated as a statistical pattern recognition task. Speech models are generated from a training set and tested on a test set. For good performance it is necessary (but often impractical) that the training and test

set should be well matched. In practice there will always be a change in the acoustic environment or a new speaker who is poorly matched to the training data leading to a degradation in the recognition performance.

It is therefore essential for good recognition to be able to easily adapt the existing models to a new input with a small amount of data from a new speaker. This is called adaptation.

Two principal different adaptation techniques are looked at in detail here, these being maximum a posteriori (MAP) adaptation and maximum-likelihood linear regression (MLLR).

For **MAP** adaptation [Gauvain and Lee, 1994], given new input speech features and the corresponding triphone alignment, the original GMMs are gradually modified to increasingly match the new input data. The major drawback of MAP adaptation is that each GMM is modified individually, requiring sufficient amounts of data to do so with confidence. Using MAP adaptation, the initial speaker-independent (SI) model estimates gradually converge to the maximum-likelihood (ML) estimate.

In contrast, **MLLR** adaptation [Leggetter and Woodland, 1995] is a method well suited to limited new data. The basic idea of MLLR adaptation is to estimate transforms that are applied to the means and variances of multiple acoustic models rather than directly adapting each single model parameter. Using MLLR, linear transforms are applied to the parameters of a set of Gaussians and shared across multiple GMMs.

MLLR addresses the locality problem, i.e. the data sparsity of MAP adaptation, since there are relatively few adaptation parameters to be estimated. Each adaptation transform can affect many GMM means and variances, making the estimation robust.

The number of different transforms (i.e. regression classes) for applying MLLR is first defined, a frame-state alignment is then performed and the GMM means $\boldsymbol{\mu}_{jm}$ and variances $\boldsymbol{\Sigma}_{jm}$ are updated as per Equations 2.33 and 2.34.

$$\hat{\boldsymbol{\mu}}_{jm} = \boldsymbol{G}\boldsymbol{\mu}_{jm} + \boldsymbol{b} \tag{2.33}$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{H}\boldsymbol{\Sigma}_{jm}\boldsymbol{H}^{T} \tag{2.34}$$

The GMM means are modified using the transformation matrix $\boldsymbol{G}$ and the offset $\boldsymbol{b}$ and the GMM variances are modified using $\boldsymbol{H}$. The more adaptation data is available,

the more transformation matrices can be inserted leading to improved matching of the GMMs to the new input data.

In practice, few regression classes and well-estimated transforms are chosen for best performance. MLLR adaptation is very often carried out in two steps: (1) two transformation matrices for two classes – silence and speech – are trained given very little adaptation data and (2) multiple classes – e.g. one silence and 31 acoustic classes – are generated once more adaptation data is present. The regression classes are automatically built from the training data available using hierarchical clustering (see e.g. Gales [1996]).

MLLR is the best known linear transform approach to speaker adaptation and has been extensively used for the research presented in this thesis.

There are two main variants of MLLR, constrained and unconstrained [Gales, 1998]. For constrained MLLR (i.e. cMLLR) the same transforms are applied to modify the GMM means and variances, that is $\boldsymbol{G} = \boldsymbol{H}$. Constrained MLLR can therefore be viewed as a feature space transformation.

The power of GMM adaptation lies in performing discriminative training, i.e. if we want to generate models for the individual speakers (in e.g. a meeting). In order to overcome the data sparsity problem, initially speaker-independent GMM-HMMs are trained, after which speaker adaptive training (SAT) is carried out to generate speaker-dependent models by using only transformation matrices for MLLR adaptation. cMLLR has proved to be more efficient in practice then MLLR for SAT. For details see Young [2008] and references therein.

2.2.8 State-of-the-art ASR performance

This section presents the performance of state-of-the-art speech recognition systems. It is first necessary to define an error metric to measure speech recognition accuracy, which is the word error rate (WER). After this, state-of-the-art ASR and multipass recognition architectures are reviewed and their recognition performance presented.

2.2.8.1 Word error rate (WER)

In order to calculate the WER of the decoded text output of a speech recognition system compared to a reference, the output of the system and the reference output are first aligned using dynamic string alignment. After this, the word error rate is computed as

$$WER = \frac{S+D+I}{N} \tag{2.35}$$

with N being the number of words in the reference output, S the number of substitutions, D the number of deletions and I the number of insertions. Consequently, the accuracy (ACC) of an ASR system is defined as

$$ACC = 1 - WER = \frac{N-S-D-I}{N} \tag{2.36}$$

Note that WER and ACC results are usually reported in % figures.

The error metrics defined above were used for the NIST RT evaluations [Fiscus et al., 2008]. In addition to these error metrics NIST also defines an *RTTM file format specification* that can be used to run automatic scripts to calculate the WER. A definition of the RTTM file format and the scripts are available on the NIST website[6].

2.2.8.2 State-of-the-art ASR

Although speech recognition is an active area of research it is already found in the latest consumer devices as a commercial product. Performance figures for state-of-the-art speech recognition are available from published work but not commercial products. The latter usually keep their performance figures (and architecture and implementation details) confidential as a trade secret – therefore leaving very little information for researchers to measure their performance. If any data is available then it is usually not comparable with published research results. Two studies of speech recognition running on desktop and mobile devices are presented in Liu et al. [2011] and Darre and Yussupov [2011]. The authors have investigated commercially available speech recognition software in terms of its suitability in health care and reported that enhanced speech recognition accuracy would be desirable. However, no concrete experimental evidence was presented on the performance of the software under test.

[6] http://www.itl.nist.gov/iad/mig/tests/rt/

The reminder of this section therefore presents performance results for published systems. These state-of-the-art speech recognition systems bear little resemblance to the one presented in Figure 2.3, despite using the basic building blocks as described above.

Modern speech recognition systems are based on **multipass recognition architectures** (see e.g. Young [2008], Hain et al. [2012]). As implied by the name multipass, recognition is carried out in several stages. Typically the initial pass is carried out using an unadapted conventional acoustic model and low-order language model. After this, feature adaptation (cepstral means and variance normalisation and VTLN), acoustic model adaptation (cMLLR/ MLLR), speaker adaptive training (SAT) and a more complex language model are inserted over several iterations.

A very good indication of the WER achievable with state-of-the art speech recognition systems was presented at the NIST Rich Transcription 2009 Meeting Recognition Evaluation Workshop[7]. It is reproduced here in Figure 2.9.

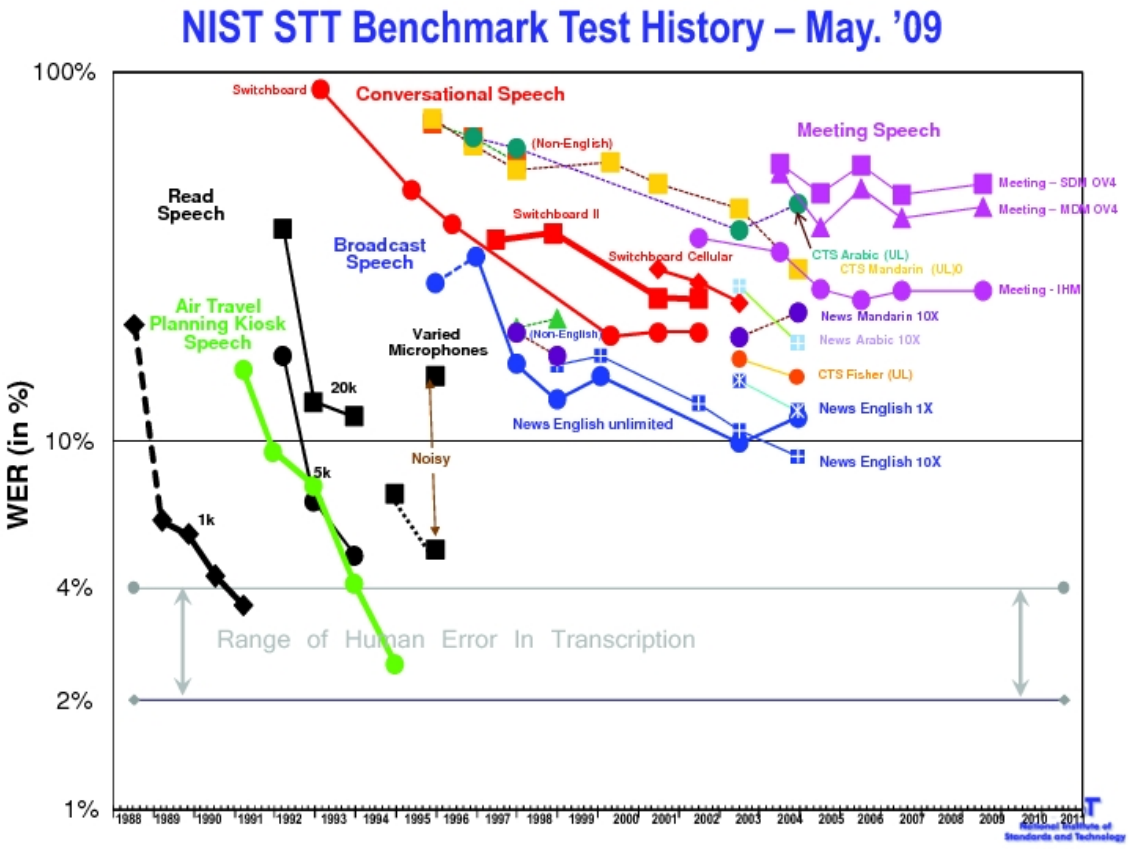


Figure 2.9: NIST RT09 STT and SASTT performance (with kind permission of Jonathan Fiscus - NIST)

[7] http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/RT09-Agenda.htm

Table 2.1: [%] WER performance of HTK and KALDI on the RM and WSJ corpora

Corpus	**Resource Management** (RM)	**WSJ** (20k open)
HTK	4.10	14.5
Kaldi	4.06	15.0

Figure 2.9 shows the performance of speech recognition on the speech to text (STT) and speaker attributed (SA) speech to text (SASTT) tasks. This clearly illustrates the way in which speech recognition improved over the last two decades and how the task was gradually made more difficult. Initial speech recognition was performed on read speech of very limited content, such as digit recognition, the 1000 word resource management (RM) corpus [Price et al., 1988] or the Wall Street Journal (WSJ) corpus [Paul and Baker, 1992]. Recognition performance on these tasks matches the range of human error in transcription. After this the research efforts moved to conversational speech, such as broadcast news (BN) [Graff, 2002], conversational telephone speech (CTS) such as the Fisher corpus [Cieri et al., 2004] and the Switchboard corpus (SWB) [Godfrey et al., 1992]. The best performing recognition systems achieve as little as 10% WER on read and conversational speech but are unable to match human performance (ranging from 2–4%). These figures were accomplished using clean audio data from close talking microphones. More recently speech recognition has also been aimed at distant speech from e.g. meetings [McCowan et al., 2005]. Here state-of-the-art recognition performance is currently around 30% WER.

Only a handful of AST tools are looked at in detail in the remainder of this section. These are the HTK [Cambridge University Engineering Department (CUED), 2012], Kaldi [Povey et al., 2011] and IBM Attila [Soltau et al., 2010] speech recognition toolkits. This is a very limited selection but one which nonetheless clearly shows the state-of-the-art.

Comparing the **state-of-the-art speech recognition performance** of the systems selected is actually difficult, mainly because their performance has been measured on many different tasks and they are therefore optimised for that specific task.

Table 2.1 summarises the performance of state-of-the-art speech recognisers on read speech, i.e. the 1000 word resource management (RM) corpus and the 20,000 word open vocabulary Wall Street Journal (WSJ) corpus, using the HTK and Kaldi toolkits.

As already indicated in Figure 2.9, read speech recognition can match human trans-

Table 2.2: [%] WER performance of the IBM Attila GMM and DBN speech recogniser on the Fisher and SWB corpora

Corpus	**Fisher** (CTS)	**SWB** (03)	**SWB** (05)
GMM	17.6	26.3	15.1
DBN	16.4	25.5	13.3

cription accuracy. This is the case for both HTK and Kaldi on the RM task, where they achieve 4% WER, but not on the WSJ task, where they achieve 15% WER.

The performance of the IBM Attila speech recogniser has been measured on conversational speech recognition and is shown in Table 2.2.

Researchers from IBM published the performance of the **IBM Attila** speech recognition toolkit in Soltau et al. [2010]. Using a GMM-HMM system they achieved 17.6% WER on the Fisher corpus (CTS), 26.3% WER on the 2003 Switchboard (SWB) task and 15.1% on the 2005 SWB task[8]. The IBM Attila toolkit is also able to perform speech recognition using deep belief networks (DBN) [Kingsbury et al., 2012]. The DBN recogniser achieved 16.4% WER on CTS, 25.5% on the 2003 SWB corpus and 13.3% WER on the 2005 SWB corpus, therefore performing marginally better than the GMM-HMM-based recogniser.

The IBM Attila toolkit was also tested using speech from the English Broadcast News (BN) corpus, achieving 15.5% WER [Soltau et al., 2010] using the GMM-HMM-based recogniser.

Note that the IEEE Signal Processing magazine has published a dedicated special issue (Volume 29, Issue 6 [Furui et al., 2012a]) on the subject “Fundamental Technologies in Modern Speech Recognition” [Furui et al., 2012b]. The results presented in the 10 articles in this issue match the results presented here.

[8]Note that both the Fisher and SWB corpus are CTS.

2.3 Distant speech recognition

Speech recognition, as looked at so far, was mostly based on speech from head-mounted or close-talking single microphones. The impressive advances in speech recognition over the last two decades have led to increased requirements, mostly for **robust** and **distant** speech recognition.

Robust speech recognition is speech recognition in adverse environments such as noise, reverberation or overlap. Distant speech recognition is speech recognition using distant speech capturing devices, i.e. the microphone is moved away from the mouth of the speaker, as is typically the case for hands-free applications or when recording a meeting with one or more table-top microphones.

This section looks at robust and distant speech recognition (DSR) and presents a review of their performance.

2.3.1 Robust speech recognition

The problem of speech recognition in adverse environments is not new. A review of the field was presented as early as 1991 in Juang [1991]. The authors looked at speech feature enhancement, noise reduction and model adaptation, and reported significant improvements in WER on digit and isolated word recognition. The methods investigated then still achieve significant improvements in speech recognition systems today, as demonstrated above and reviewed in Droppo and Acero [2008]. Obviously feature enhancement, noise reduction and model adaptation have progressed enormously since then in step with the progress in the recognition, as shown by Droppo and Acero [2008].

2.3.2 Audio-visual (AV) speech processing

A more recent approach to improve speech recognition accuracy is **audio-visual (AV) speech processing**. Today's increased computing power allows for combining visual features with audio features, thus allowing AV voice activity detection, AV sound source localisation and AV speech recognition (AVSR), as reviewed in Chin et al. [2012] for example. Until recently the research in AVSR has concentrated on audio and video information from close talking speakers [Chin et al., 2012].

This trend is now shifting, in part due to the availability of Microsoft's Kinect™ [9], an affordable sound, video and motion capturing device that

> " gives computers eyes, ears, and the capacity to use them ... allowing [people] to interact naturally with computers by simply gesturing and speaking."

The Kinect™ and its software development kit (SDK) enabled researchers at IBM [Galatas et al., 2012] to incorporate facial depth in addition to the audio and video features for carrying out speech recognition experiments. The authors achieved absolute WER improvement of up to 10% (from 50% using audio features only) at 0 dB SNR and 30% WER improvement at -10 dB SNR (from 15% using audio features only) on a connected digit recognition task.

2.3.3 Distant speech recognition system architecture

Figure 2.10 shows a typical distant speech recognition system architecture using audio and video signals to carry out audio-visual speech recognition (AVSR), as per Wölfel and McDonough [2009].

First, in the so-called front-end stage, the multiple audio and video channels are compressed and enhanced to reduce the data rate at the highest possible quality. For the audio signal, speaker localisation and tracking is carried out and the best audio channel selected and/or the audio quality improved by means of noise reduction, acoustic beamforming and postfiltering. For the video signal, areas of interest (such as movement) are detected and the video is compressed without loss of quality in important regions.

After this, voice activity detection and segmentation and clustering are carried out to detect "who spoke when", i.e. to perform speaker diarisation. VAD and diarisation are reviewed in detail in Section 2.4 below. Automatic speech recognition (reviewed above) follows downstream from the diarisation process.

McDonough et al. [2008b] reviewed the single components of a complete DSR system and found that:

> " while it is tempting to isolate and optimize each component individually, experience has proven that such an approach cannot lead to optimal performance."

[9] http://www.microsoft.com/en-us/kinectforwindows

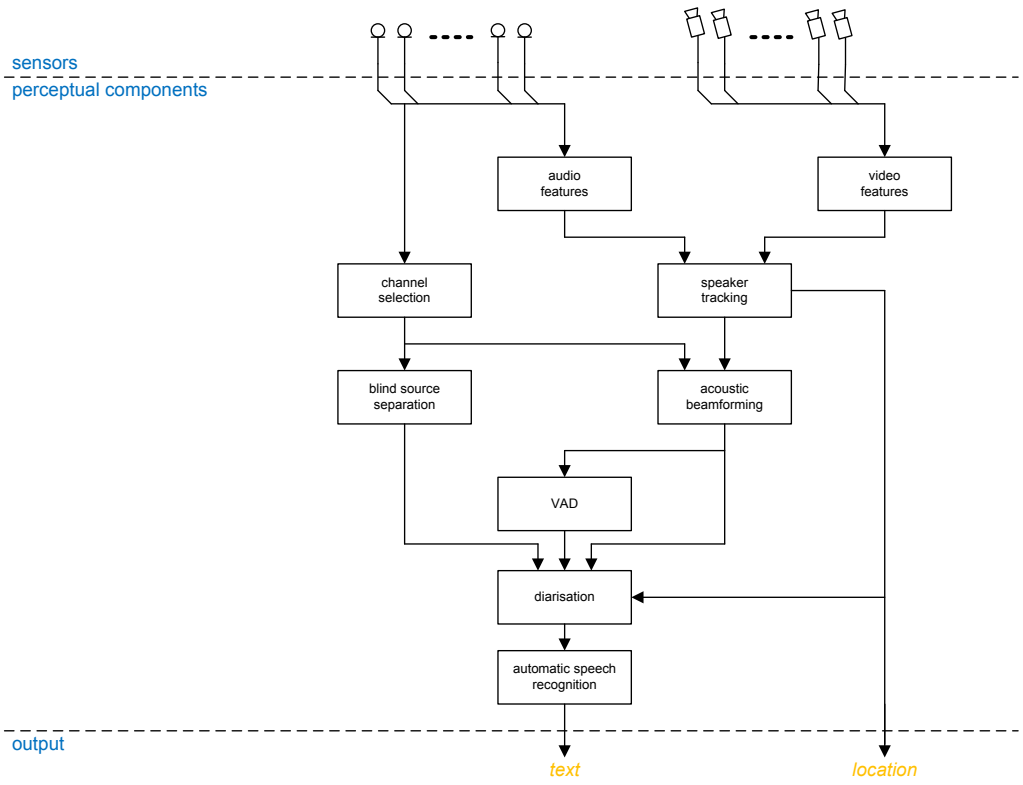


Figure 2.10: Architecture of an audio-visual DSR system (with kind permission of John Wiley & Sons, Inc., modified from Wölfel and McDonough [2009], Figure 1.9)

The success of a good DSR system lies in the architecture and combination of its components as well as the interaction of their input and output features, as demonstrated next in Section 2.3.4. Please note that the reminder of this thesis looks at the audio components of DSR only.

2.3.4 State-of-the-art DSR performance

Like ASR, distant speech recognition performance is measured by word error rate (WER). While ASR performance is commonly measured using close talking microphones (ctm), DSR performance is usually reported on a single distant microphone (sdm) and multiple distant microphones (mdm) – as per the NIST RT evaluations [Fiscus et al., 2008].

A typical distant speech recognition task is the transcription of meetings, as demonstrated by Hain et al. [2012] for the AMI/DA system. The authors implemented a 10-pass ASR system which is used for meeting transcription. Speech recognition in meetings

(from multiple distant microphones) is considered one of the most difficult ASR tasks today.

The AMI/DA system contains the speech pre-processing, that is the speech front-end, consisting of noise reduction and acoustic beamforming; speech segmentation and clustering, or diarisation; speech feature generation (PLP) and enhancement; and a multiple-pass ASR system containing state-of-the-art adaptation techniques such as VTLN, cMLLR and MLLR.

Hain et al. [2012] achieved 29.3% WER on the NIST RT07 and 33.2% on the RT09 SASTT tasks. Only two teams, AMI/DA and ICSI/SRI, participated in the STT and speaker attributed STT (SASTT) challenge of the RT09 workshop[10]. Both teams achieved similar results as presented above for the AMI/DA system (see the RT09 workshop homepage for details).

Two years earlier, five teams participated in the STT and SASTT challenge of the NIST RT07 workshop[11], that is AMI/DA and ICSI/SRI on the meeting or conference room data and ICSI/SRI, IBM and UKA on the lecture room data. WERs for the meeting room data were between 30% and 50% on the mdm condition and between 40% and 70% for the lecture data (see the RT07 workshop homepage for details).

Two years earlier, five teams participated in the STT and SASTT challenge of the NIST RT07 workshop[12], that is AMI/DA and ICSI/SRI on the meeting or conference room data and ICSI/SRI, IBM and UKA on the lecture room data. WERs for the meeting room data were between 30% and 50% on the mdm condition and between 40% and 70% for the lecture data (see the RT07 workshop homepage for details).

Meeting recognition, as stated above, is quite a complex task, dealing with distant speech signals degraded by noise, reverberation and overlapping speech. While it is important to deal with the complete system for a realistic WER and performance (as per McDonough et al. [2008b]), it is also equally important to be able to isolate the individual problems to find good solutions (and algorithms).

One open research problem for good DSR is speech recognition of overlapping speech. A vehicle to carry out research in recognising overlapping speech is the PASCAL speech separation challenge 2 (SSC2) [Lincoln et al., 2005] which produced the MC-

[10] http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/RT09-Agenda.htm
[11] http://www.itl.nist.gov/iad/mig/tests/rt/2007/workshop/RT07-Agenda.htm
[12] http://www.itl.nist.gov/iad/mig//tests/rt/2007/workshop/RT07-Agenda.htm

WSJ-AV corpus of overlapping speech recorded with multiple distant microphones. The MC-WSJ-AV corpus and the DSR systems that were developed and evaluated for the PASCAL speech separation challenge will be presented and reviewed in Section 2.5.1, after the review of speaker diarisation in the next section.

2.4 Speaker diarisation

Speaker diarisation is the process of determining *who spoke when* in a multi-party conversation such as a TV or talk show, a lecture or a meeting. Initially, the main application for diarisation was upstream processing for automatic speech recognition, that is speaker attributed text to speech (SASTT) processing. In recent years diarisation has meanwhile become a key technology not only for audio transcription but also for audio classification and content retrieval, audio segmentation for archiving and sound indexing and search.

The principal diarisation flow is shown in Figure 2.11.

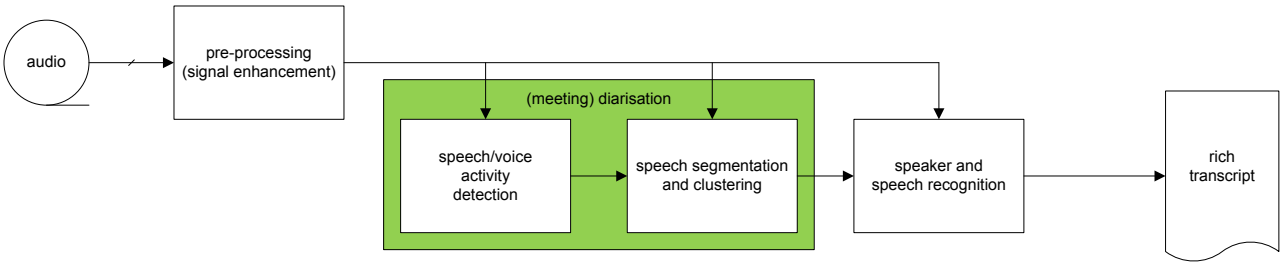


Figure 2.11: Typical processing flow for rich transcription

The first step in diarisation is to determine whether an incoming audio segment[13] is speech or not, i.e. voice activity detection (VAD). Silence (and music) segments are not of interest in diarisation and are therefore discarded for the downstream processing. The individual speech segments detected by the VAD process are analysed for a change of speaker. If this is the case, the speech block is segmented. In the third and final step, speech segments are analysed to ascertain whether they belong to the

[13] An audio segment in the context of VAD and diarisation is a continuous block of audio containing no elements of silence longer than 200 ms.

same (unidentified) speaker and segments believed to come from the same speaker are clustered.

The audio input to speaker diarisation can either be from a recording or fed in continuously in an online system. The work presented in this thesis is not restricted to recorded audio data. All the algorithms analysed and developed are assessed for their suitability to run online.

It is common to merge the segmentation and clustering steps and execute them in multiple iterations, as shown in Figure 2.11. A very good review of the recent research in speaker diarisation is presented in Anguera Miro et al. [2012].

2.4.1 Voice activity detection (VAD)

The following section is a review of voice (or speech) activity detection. Voice activity detection (VAD) is one of the first processes that is applied to an incoming audio stream before it is used for further processing. Subsequent processes rely on the correct labelling of the incoming audio stream(s) for optimal performance, i.e. to reduce false alarms. Problems with the incoming audio stream are:

- noise, both stationary and intermittent
- reverberation
- overlapping speech/speakers
- non-speech segments (such as music)

Disturbing signals that interfere with speech processing techniques are not only noise (both stationary such as from an equipment fan or intermittent like clapping) but also reverberation, echo and music. Stationary noise removal and echo and reverberation cancelling may be dealt with in the speech processing front-end. Intermittent noise or music (possibly mixed with speech), on the other hand, are usually dealt with during speech enhancement or acoustic modelling. Alternatively, the diarisation (and recognition) models may be trained with noisy speech and the output ought to then be correctly decoded. Algorithms developed for speech activity detection need to distinguish between music and speech and also need to be able to cope with overlapping speech. There are four main techniques used for speech activity detection (also known as accurate endpointing). These are:

- signal (or energy) threshold detection
- hidden Markov models (HMM)
- likelihood ratio tests (LRT)
- neural networks

Mobile devices rely mostly on energy-based VAD methods such as the ones defined in the ITU-T P.56 "Objective measurement of active speech level" standard [ITU-T, 2011] or the ETSI ES 202 050 "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms" standard [ETSI, 2007].

Assuming that speech is recorded in a quiet room, the most basic method of detecting speech is to measure the energy level of the recorded signal, track the power envelope and use a threshold level to determine whether speech is present or not, as outlined in Equation 2.37

$$VAD = (E_{current} - E_{average}) > E_{threshold}, \tag{2.37}$$

where $E_{current}$ is the current energy of the input audio signal, $E_{average}$ is the average of the tracked power envelope and $E_{threshold}$ is a predefined threshold which is obtained through experiments and adapted to the recording environment. Note that these thresholds may use a forget-me factor, i.e. a predefined time over which the power envelope is tracked and averaged in order to compensate for a change in the quasi-stationary noise.

For meeting recordings, using the latest technology allows recording equipment to be built that is highly flexible, mobile and lightweight [Hori et al., 2010]. This enables recording of conversations in many different environments, therefore making the process of speech activity detection increasingly difficult. These problems are overcome using multiple channel recording (e.g. microphone arrays) and enhanced signal processing, utilising the ever increasing computer power available.

VAD methods utilised in meeting recordings are for example LRT-based [Sohn et al., 1999], neural-network-based [Adami et al., 2002b] or GMM-HMM-based [Huijbregts, 2006]. These and additional algorithms and methods will be reviewed in detail in Chapter 5.

2.4.2 Speech segmentation and speaker clustering

The next section serves as a review of speech segmentation and speaker clustering. Segmentation and clustering is the task of partitioning audio segments into homogeneous regions which are believed to belong to one speaker only. Speech segmentation is the process of taking audio and splitting it into segments speaker change points. Speaker clustering is performed to assign the speech segments to the individual speakers. Segmentation and clustering is traditionally carried out in combination over multiple iterations.

2.4.2.1 The Bayesian information criterion

Speech segments identified by voice activity detection algorithms may contain speech from more than one speaker. The Bayesian information criterion (BIC [Chen and Gopalakrishnan, 1998b]) has been found to be a reliable measure to determine whether a segment contains one or more speakers, i.e. to avoid the entire segment being incorrectly assigned to a single speaker during diarisation. It is used extensively in most of the state-of-the-art diarisation systems such as the ICSI [Friedland et al., 2012]) and SHoUT [Huijbregts, 2008] tools.

BIC is a likelihood criterion penalised by the model complexity, i.e. the number of parameters in the model. If $\mathcal{C}$ is the audio data and M are the candidate models, the task at hand is now to maximise the likelihood for each model $L(\mathcal{C},M)$ penalised by the number of parameters of M. For this the audio data $\mathcal{C}$ is modelled as a multivariate Gaussian distribution $N(\mu,\Sigma)$. The Bayesian information criterion for an audio cluster $\mathcal{C}_k$ is now defined as

$$BIC(\mathcal{C}_k) = \sum_{i=1}^{k} \Big\{ -\frac{1}{2} n_i \log|\Sigma_i| \Big\} - \lambda P, \tag{2.38}$$

where n_i is the number of samples in the cluster and Σ_i is the sample covariance matrix. The penalty P is defined as

$$P = \frac{1}{2}(d + \frac{1}{2} d\,(d+1)) \log N, \tag{2.39}$$

where $N = \sum_i n_i$ is the total sample size and d the number of parameters per cluster. Note that λ, the penalty weight, is usually set to 1.

The Bayesian information criterion can now be used to calculate whether a speech segment contains one or more different speakers (i.e. segmentation) and to determine whether two speech segments are from the same speaker (i.e. clustering). Using the BIC for segmentation and clustering can best be explained for the latter. The increase in the ΔBIC value for merging two segments s_1 and s_2 can be shown to be:

$$\Delta BIC = BIC(s_1 + s_2) - (BIC(s_1) + BIC(s_2)) \tag{2.40}$$

$$= n \log \Sigma - n_1 \log \Sigma_1 - n_2 \log \Sigma_2 - \lambda P, \tag{2.41}$$

where n_i is the number of samples in the cluster $i = 1, 2$ and Σ_i is the sample covariance matrix of each segment, $n = n_1 + n_2$ and Σ is the covariance matrix of the combined input segments.

If the ΔBIC value is greater than zero then the information content of the merged segments is higher than the individual segments and the two segments are likely to belong to the same speaker and should be merged. Similarly, a speaker change is indicated by a positive peak of the ΔBIC value when calculating a series of ΔBIC values for a sliding split point of a speech segment.

Chen and Gopalakrishnan [1998b] carried out experiments to detect the change point in a speech segment, comparing the first cepstral coefficient and the Gish (i.e. log-likelihood) and KL2 (Kullback Leibler) distance with the ΔBIC, and found that the ΔBIC is the best performing measure. Using the ΔBIC allows the search for speaker changes in audio segments of variable length and provides a termination criterion for speech segment clustering. Details and results on using the BIC in speaker change detection can be found in Chen and Gopalakrishnan [1998b] and details on how to use the BIC for speech clustering are presented in Chen and Gopalakrishnan [1998a].

Using the ΔBIC to detect speaker changes in a speech segment (and deciding whether to merge two audio segments) can be carried out by means of the speech features (e.g. MFCCs) as defined in Equation 2.41. This requires application of the penalty λP to compensate for the difference in length of the two speech segments to be compared. Ajmera and Wooters [2003] devised a method that does not require this penalty. If Gaussian mixture models (GMM) are trained on the two speech segments and if the

combined GMM contains the sum of mixtures of the two separate models, then the penalty λP cancels out.

2.4.3 Diarisation

This section reviews typical diarisation methods and algorithms. Diarisation typically follows VAD as presented in Figure 2.11. Over recent years speaker diarisation has gradually become more difficult by progressing from working with telephony data (CTS) to broadcast news (BN) and then meeting data, similar to speech recognition.

Speaker diarisation can be carried out in two principally different ways: top-down or bottom-up. Top-down methods naturally require the presence of the complete data and are therefore not suitable for online processing. Both methods employ the same principal algorithms and achieve similar results, as presented in Evans et al. [2012].

One of the best known and successful diarisation systems has been developed at the ICSI (International Computer Science Institute [Friedland et al., 2012]). Its performance has been consistently at the top and most other systems have been designed in a similar manner. Its review therefore serves as the basis for the review of state-of-the-art diarisation.

The ICSI diarisation engine follows the basic flow diagram presented in Figure 2.11. First the single or multiple channel audio input signal is enhanced, i.e. a dynamic range compression is performed, before the noise is reduced using Wiener filters and, in the case of multiple input channels, the audio is compressed and enhanced to a single channel using acoustic beamforming. Next VAD is carried out using preliminary bootstrapping of the speech and silence regions which are then used to train the VAD models (GMM-HMM). Over multiple iterations a speech model, an audible non-speech model and a silence model are trained using the EM algorithm. Speech segmentation and clustering are performed iteratively following the steps outlined below and illustrated in Figure 2.12.

1. initialise basic models
2. resegment audio data and retrain models using the EM algorithm
3. merge models using ΔBIC
4. repeat 2 & 3 as long as model purity improves
5. perform final segmentation and write output

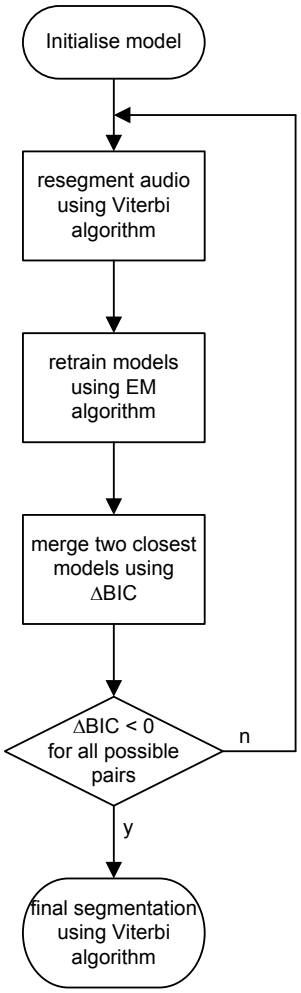


Figure 2.12: Schematic overview of the ICSI diarisation system

First, the basic models are initialised by splitting the complete audio recording into $2N$ segments. After this N models are trained by taking separated audio segments. Viterbi alignment is used to generate a diarisation sequence with N speakers and the N models are retrained with their new speech segments. Step 2, resegmentation and retraining, is usually repeated a few times. The N models are then tested to determine whether they might contain the same speaker and top-scoring models are merged using the Bayesian

information criterion (ΔBIC), thus decreasing N. Steps 2 and 3 are then repeated as long as the model purity improves until a stopping criterion is reached. The diarisation process is completed by writing the output. Note that there is no segmentation step as such, but speaker changes are detected using Viterbi alignment and the minimum speech segment length is defined by the length of the HMM.

Diarisation performed using the bottom-up flow principle has the following weaknesses:

- Using Viterbi alignment on GMM-HMMs and the ΔBIC to merge speech segments, the minimal speech segment duration needs to be rather long, i.e. 3 or more seconds, which does not fit with the average speech segment length in meetings of approximately 1.5 s (see Figure 6.11 on Page 128).
- Initial segmentation is carried out after removing non-speech (and non-audible) segments and may occur at places which are not speaker boundaries. Recovering from these artificial cuts is difficult.
- The number of active speakers is not known at any stage during merging and it is therefore difficult to define the merge stopping criterion.

Bottom-up approaches capture comparatively purer models then top-down methods which provide less discriminative and potentially better normalised speaker models [Evans et al., 2012]. The top-down approach starts with a single general speaker model from which it constructs new speaker models one-by-one. First a general speaker is trained with all the available acoustic data. A new speaker model is then introduced and trained with a subset of the total data. Selecting a good subset is crucial and – in the implementation presented by Evans et al. [2012] – the longest speech segment detected by the VAD process has proved to give consistently good performance.

A major advantage of top-down diarisation is potentially more reliable new speaker models as they are drawn from a well-normalised background model which has been trained from the complete speech data.

Top-down diarisation achieves similar performance but, like bottom-up approaches, suffers from an undefined segmentation and clustering stopping criterion, while also being unable to run in an online matter, as mentioned above.

Open issues in diarisation are looked at in detail in the next few sections.

2.4.3.1 Cluster initialisation

In segmentation and clustering, cluster initialisation is an open research problem for both top-down and bottom-up approaches. This problem has been addressed in multiple ways such as using the overall meeting length [Friedland et al., 2012] or using K-means initial assignments instead of bootstrapping [Ben-Harush et al., 2012], producing minor improvements at the expense of increased system complexity and computational demand.

2.4.3.2 Cluster merge (and segmentation) stop

Precise change point detection and merge stopping is also an open research issue. Investigations undertaken by Chen and Gopalakrishnan [1998b,a] have been extended using the

- generalized likelihood ratio (GLR) and penalised GLR,
- (symmetric) Kullback-Leibler (KL) and asymmetric KL2 divergence [Siegler et al., 1997] and
- information change rate (ICR), or entropy.

Experiments carried out using these algorithms on different data sets led to no significant improvement over the ΔBIC [Anguera Miro et al., 2012, Barras et al., 2006, Han and Narayanan, 2007, 2008, Kotti et al., 2008]. Segmentation and clustering using the ΔBIC, however unsuited to short speech segments, remains the first choice for speaker diarisation [Anguera Miro et al., 2012].

2.4.3.3 Cluster purification

Segmentation and clustering very often lead to impure clusters, such as clusters created from speech segments containing multiple speakers (overlapping or not) or when a speech segment includes silence, non-speech (e.g. music) or a mixture of silence, music and speech. These clusters are 'impure' and a cause of major degradation in the diarisation task [Sinclair and King, 2013].

Anguera et al. [2006] suggest both segment and frame level purification for improved diarisation. For segment level purification, the authors remove the speech segments

which are most dissimilar from each speaker model after each iteration of segmentation and clustering. For frame level purification, the statistics of the individual speech segments are analysed as to whether they contain silence, and speech segments believed to do so are subsequently removed from training. The proposed scheme achieves purer speaker models and an improved diarisation output.

Inspired by the re-sampling technique known from pattern classification, Nwe et al. [2012] performed clustering several hundred times (with different numbers of GMMs) on the assumption that pure segments will always cluster to the same model, while impure ones will not. This process creates a consensus matrix, allowing impure clusters to be removed from the model. The proposed method is part of the I2R diarisation system, the best-performing diarisation system in the NIST RT09 challenge.

Bozonnet et al. [2010] applied the purification method suggested by researchers from the I2R to their top-down diarisation system, also achieving purer speaker models and improved diarisation.

2.4.3.4 Multiple acoustic features

Acoustic features such as MFCCs or PLPs are optimised to reduce the speaker dependency of the incoming audio signal for best speech recognition. This, unfortunately, is not optimal for speaker diarisation which can be significantly improved by combining traditional short-term features (MFCCs) with prosodic and other long-term features.

Friedland et al. [2009] analysed 70 different long-term features (such as pitch, energy, formants, harmonics-to-noise ratio and long-term average spectrum) with respect to improved diarisation on the sdm condition, assuming that multiple microphone signals would not always be available.

They found that ...

> "... the median and average fundamental frequency are the best features, followed by high formants (F4, F5). Also, the mean harmonics-to-noise ratio and the variance of the long-term average spectrum achieved a high score. Although pitch median and pitch mean are likely to be highly correlated, we decided to keep them both since their scores are outstanding."

Given audio data from multiple microphones, localisation data can be extracted using sound source localisation and the audio signal can be enhanced using beamforming

techniques. Unfortunately, if neither the microphone positions nor the time alignment of the audio channels are defined (as is the case for the NIST RT evaluations), then sound source localisation can only be performed by means of TDOA estimation and only delay-sum beamforming is possible [Bitzer and Simmer, 2001].

Ellis and Liu [2004] successfully demonstrated the use of microphone channel cross-correlation to detect speaker turns in meetings. Pardo et al. [2006] applied this method to carry out diarisation using the TDOA features only, and later combined the acoustic (MFCC) and location features [Pardo et al., 2007].

In today's state-of-the-art diarisation systems, using multiple acoustic features gives a consistent 30% relative improvement in diarisation error rate (DER) [Friedland et al., 2012].

Pardo et al. [2012] found that improved performance can be achieved by adding the intensity channel contribution and interpolated fundamental frequency to the feature stream. The intensity channel contribution is the normalised energy of the signal arriving at the different microphones and, in addition to the TDOAs, another way to determine the direction of arrival of the sound. The second new feature is the F0 frequency, or pitch, of the incoming audio. By adding these two features, the authors managed to improve the DER of their system. The additional features require weighting and careful alignment as they are produced in different time intervals.

The major problem of combining multiple features is that TDOA features are long term features (only stable over several seconds) while MFCC are short term features (typically stable over a vowel length of a few 100 ms). The combination of short and long term features makes implementation difficult, and segmentation and clustering of short speech segments is hindered by these constraints. Pitch periods of speech are only available for voiced speech, therefore making their integration difficult, despite them being short term features.

Pardo et al. [2012], Ishiguro et al. [2012], Nwe et al. [2012], Zelenak et al. [2012] and many other researchers realised that the TDOAs are crucial for good diarisation, but also very difficult to integrate. Pardo et al. [2012] (and Friedland et al. [2012] as well as Huijbregts et al. [2012]) integrate the TDOA values as a (weighted) parallel feature stream.

Nwe et al. [2012] treat the TDOAs as multi-dimensional features and use consensus-based cluster purification to reduce the dimension of the TDOAs to two only, thereby

determining the number of speakers in the recording. Ishiguro et al. [2012] use the direction of arrival (DOA) information calculated from a microphone array and the 'bag-of-words' model[14] (applied to these DOAs) to determine the number of (active) speakers. ΔBIC-based segmentation and clustering then leads to improved diarisation results as the cluster stopping criterion is known.

Video recordings of meetings are available for the ICSI and AMI/DA meetings, including those used for the NIST RT evaluations. Friedland et al. [2012] looked at combining the acoustic, prosodic and video features into the diarisation task to perform audio-visual (AV) diarisation. Using the close-up video information from EDInburgh and IDIap meetings, the authors calculate average motion vector magnitudes over estimated skin blocks and add these as an additional feature stream. Unfortunately, the proposed AV diarisation system performs marginally worse than the baseline system (using sdm audio features only).

Combining audio and visual features is computationally expensive and still an open research area.

2.4.3.5 Overlapping speech

Overlapping speech accounts for 5–10% of all speech in meetings, leading to a significant degradation in speaker diarisation if not detected or handled. Zelenak et al. [2012] developed a diarisation system that takes account of simultanous (i.e. overlapping) speech. The system is based on agglomerative clustering like e.g. the ICSI system presented above. While the systems presented so far ignore overlapping speech, Zelenak et al. [2012] developed a method to detect simultaneous speech and attribute these segments to the correct speakers.

In the proposed method the authors implemented a two-stream GMM-HMM diarisation system with an overlap detection component combining spectral (audio) and spatial (TDOA) features which correctly detected 20% of the overlapping speech on the AMI data set and 5% on the RT09 data set, leading to a marginally improved DER.

This diarisation system and overlap detection method will be reviewed in detail in

[14] Bag-of-words (BoW) approaches are well known histographical representation method initially used in natural language processing and information retrieval research. BoW representation describes a document by a histogram of words which are appearing in it. For diarisation, the authors use the BoW representation to represent the frame-wise observations of the speaker localisations.

Section 7.2.

2.4.3.6 Non-parametric diarisation

A novel method for diarisation using a non-parametric approach is the information theoretic framework presented by Vijayasenan [2010]. The information bottleneck (IB) principle is a non-parametric clustering method aimed to find the relevance variables of a cluster where, in diarisation, each remaining cluster represents a speaker upon completion. The critical component of IB-based diarisation is the stopping criterion, i.e. when to stop the merging of clusters.

The system presented achieves similar speaker error rates[15] (16.8%) as a baseline GMM-HMM system (17.0%) while performing the diarisation six times faster than real-time, compared to the GMM-HMM-based system which is slower than real-time. The IB principle is also much better suited to integrating different features as it combines the feature streams in a normalised space of relevance variables compared to GMM-HMM-based systems which make use of log-likelihood combination.

The IB-based system achieved 5% absolute improvement over the baseline for the two feature combination (MFCCs and TDOAs) and 7% for the four feature combination (MFCC, TDOA, modulation spectrum and frequency domain linear prediction) while still running in real-time and ten times faster than the GMM-HMM-based system.

2.4.3.7 Online processing

Diarisation as reviewed so far is not able to run in an online manner but requires the complete recording to be available in order to process the data iteratively many dozens of times. Online diarisation requires that the data is processed as it arrives and that the latency, i.e. the time from arrival to completed processing of a data segment, is within acceptable limits of e.g. a few seconds.

A real-time online diarisation, speech recognition and speech analysis system has been developed by researchers at NTT [Ishiguro et al., 2012, Hori et al., 2010, Ishiguro et al., 2012].

[15] DER results are not reported

The researchers at NTT use dereverberation, noise reduction and acoustic beamforming to enhance the audio signal. Their system performs combined noise suppression and VAD by applying a likelihood-ratio test (LRT) to the GMMs of noisy speech and silence models. In the proposed method (called DIVIDE), clean pre-trained speech and silence models are combined with noise models into GMMs on which the LRT test is performed. The noise models are constantly updated using Kalman filtering under the assumption of non-stationary noise. Segmentation and clustering are carried out using the speech that is passed on from the VAD by performing the 'who speaks when' process in two steps: (1) DOA features generated using independent component analysis are used to determine the speaker position after which (2) speech separation is carried out. The DOA, i.e. the speaker position, and the VAD output are combined to generate the diarisation output. Speech recognition is then carried out using the results from the diarisation process. In the final stage speech analysis, that is speaker activity detection (e.g. speaking, laughter, watching someone) and meeting statistics (e.g. topic, activeness, casualness) are produced. When testing their system on the RT data the researchers at NTT unfortunately failed to achieve results matching the ICSI or SHoUT systems due to the NTT system working in online mode compared to the batch mode processing used by the ICSI and SHoUT systems.

2.4.4 State-of-the-art diarisation

Information on current and state-of-the-art diarisation systems can be obtained from the NIST RT challenges. These challenges are very useful for researchers as they define standard experimental tools and databases which enable researchers to compare their algorithms and systems.

The next section serves to present the diarisation error metric (DER), a measure to assess and compare the performance of a diarisation algorithm or system, defined by the NIST. This will be followed by a discussion of the performance of different diarisation systems.

2.4.4.1 Diarisation error metrics

Two metrics are used to verify the performance of speaker diarisation systems: these are the VER (VAD error rate) and DER (diarisation error rate). The VER is calculated

from the missed speech (MS) and false alarms (FA). Missed speech are recorded audio segments that are not detected as speech and therefore lost. False alarms are segments passed to the next processing step that are actually not speech. The DER is calculated from the missed speakers, false alarms and the speaker error, i.e. audio segments that are assigned to the wrong speaker.

These individual errors are calculated as follows:

E_{miss}: percentage of missed speaker time

$$E_{miss} = \frac{\sum_{s=1}^{S} dur(s) \cdot (N_{ref}(s) - N_{hyp}(s))}{T_{score}} \qquad \forall\, (N_{ref}(s) - N_{hyp}(s)) > 0 \tag{2.42}$$

with $T_{score} = \sum_{s=1}^{S} dur(s) \cdot N_{ref}$.

S is the total number of segments of the recording in the reference, $dur(s)$ is the length of a segment and N_x is the speaker ID for the individual segment with $N_x = 0$ indicating a non-speech segment. Note that segment boundaries are defined as speaker change points in the reference transcript and the system output.

E_{FA}: percentage of false alarm time

$$E_{FA} = \frac{\sum_{s=1}^{S} dur(s) \cdot (N_{hyp}(s) - N_{ref}(s))}{T_{score}} \qquad \forall\, (N_{hyp}(s) - N_{ref}(s)) > 0 \tag{2.43}$$

E_{spkr}: percentage of speaker error time assigned to the wrong speaker

$$E_{spkr} = \frac{\sum_{s=1}^{S} dur(s) \cdot (min(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))}{T_{score}} \tag{2.44}$$

The DER is then calculated from Equations 2.42, 2.43 and 2.44 as follows:

$$DER = E_{miss} + E_{FA} + E_{spkr} \tag{2.45}$$

Calculating the DER as defined in Equation 2.45 includes overlap errors, i.e. regions of speech where multiple speakers are talking and where the incorrect number of speakers is found. Note that overlap errors fuse into E_{miss} and E_{FA} if missed speech and false alarms are reported per speaker. Note also that the scoring script supplied by NIST[16] allows a flag to be set which defines whether overlapping speech is ignored or not.

The VER and DER results presented in this thesis do not include overlapping speech, unless stated otherwise, and are calculated as per Equations 2.42, 2.43 and 2.44 with

[16]The latest version of SCTK tools are obtainable from `ftp://jaguar.ncsl.nist.gov/pub/`

the modification that $dur(s)$ is calculated as $(dur(s) \cdot X)$ with X being either 0 or 1, i.e. $(N_{ref}(s) - N_{hyp}(s))$ in Equation 2.42, $(N_{hyp}(s) - N_{ref}(s))$ in Equation 2.43 and $(min(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))$ in Equation 2.43 are limited to a maximum of 1.

Note that a collar (or hang-over) of ± 250 ms is defined at the edges of speech segments. This allows for a tolerance band for the automatic processing of the audio files and for the error due to the human labelling of the references.

2.4.4.2 State-of-the-art diarisation performance

The performance of state-of-the-art diarisation systems is reviewed in what follows, including the analysis of top-down vs. bottom-up diarisation as well as the improvements of multiple-acoustic-feature diarisation. This is achieved by comparing the performance of all the contestants of the NIST RT09 evaluation, that is the diarisation systems implemented by Friedland et al. [2012], Evans et al. [2012], Huijbregts et al. [2012], Pardo et al. [2012], Zelenak et al. [2012] and Nwe et al. [2012]. Note that all results presented here apply to the mdm condition.

A good starting point is the ICSI diarisation system [Friedland et al., 2012]. It uses multiple acoustic features for the model training, i.e. audio (MFCCs), localisation (TDOAs) and other speech features such as pitch, formants and harmonics. It achieves 4.9% VER and 17.2% DER on the RT09 evaluation data set. Combining audio and visual features does not show any improvements.

Another well known system is the SHoUT diarisation and speech recognition system [Huijbregts et al., 2012, Huijbregts and van Leeuwen, 2012]. It follows the same basic steps as the ICSI system and achieves 26.6% DER on the RT09 data set. The SHoUT Speech Recognition Toolkit is freely downloadable for research purposes (from Huijbregts [2006]).

Pardo et al. [2012] found that improved performance can be achieved by adding the intensity channel contribution[17] and interpolated fundamental frequency to the feature stream. Adding these two features the authors managed to improve the DER of their system by 16.7% to 21.4% overall for the RT09 evaluation data set. The additional features require weighting and careful alignment as they are produced in different time intervals. The system was optimised using RT05 and RT06 data and achieved the best

[17] The intensity channel is the channel with the relatively highest energy, updated on a frame-base.

Table 2.3: State-of-the-art diarisation error DER [%] on the NIST RT data

System	ICSI	SHoUT	LIA/Eurecom	UPM	SUT	I2R	IB
RT07	n.a.	n.a.	18.6 21.9[a]	n.a.	n.a.	7.5	11.9
RT09	17.2	26.6	21.3 26.8[a]	21.4	42.2[a]	8.8	13.2

[a] DER calculated by adding overlap errors

results with a weight of 0.9 on the acoustic features (MFCCs) and a weight of 0.1 equally distributed on all the other features.

Nwe et al. [2012] achieved the best DER on the NIST challenges at 7.5% for the RT07 and 8.8% for the RT09 data. The researchers at the I2R implemented a 2-stage diarisation system, where the first stage was similar to the ICSI and SHoUT systems. In the second stage, the authors implemented a consensus-based cluster purification method that removed inpure speech segments, leading to better speaker models and the overall best diarisation system presented at the NIST RT09.

Zelenak et al. [2012] developed a method to detect simultaneous speech and attribute these segments to the correct speakers. The diarisation rate of the baseline system measured on the RT09 data set is 42.2%, significantly higher than previously reported results because overlapping speech is not ignored. The authors achieved an absolute DER improvement of 1.2% and a relative improvement of 2.7% with the proposed overlap detection method.

Please note that all diarisation systems contain many (model and hyper-) parameters which are finely tuned using data from e.g. the RT05 and RT06 evaluations and tested on the RT07 and RT09 evaluations. The variation in the DER of the individual meetings is very large, as shown in Pardo et al. [2012], for example.

Table 2.3 summarises the results presented at the NIST RT09 workshop. The codes for the different research groups are: ICSI [Friedland et al., 2012], SHoUT [Huijbregts et al., 2012], LIA/Eurecom [Evans et al., 2012], SUT (Slovak University of Technology) [Zelenak et al., 2012], UPM (Universidad Politécnica de Madrid) [Pardo et al., 2012] and Institute for Infocomm Research (I2R) [Nwe et al., 2012]. Results for the information bottleneck (IB) [Vijayasenan, 2010] principle are added for completeness.

2.5 Speech corpora and (open source) software

This section reviews speech corpora and open source software. Almost all the results presented so far have been produced from experiments carried out on publicly available speech corpora. These corpora and the metrics and scripts to measure the performance of a novel algorithm or method define their success or failure.

While it is impossible to present every single corpus available to speech researchers, the next section will give an overview and provide details of the two most relevant ones for the research presented in this thesis. These are multi-party conversation and overlapping speech corpora. After this follows a list of the open source software used along with a brief explanation of the function of the components.

2.5.1 Multi-party conversation corpora

State-of-the-art performance of diarisation systems is usually measured using a multi-party conversation corpus. This started as early as 2002 when experiments were carried out using broadcast news and conversational telephone speech for the NIST RT02 challenge on the speech to text and meta data extraction (MDE, including speaker diarisation) tasks.

From then on the NIST organised the RT challenges every year (or even twice-yearly in the early years) until 2009. English meeting domain data was first released for the 2005 challenge. Different research institutes had by then created extensive corpora from which the NIST data was taken. These include the corpora from:

- ICSI [Wooters, 2009]
- M4 [Renals, 2004]
- AMI(DA) [Renals, 2010]
- CHIL [Mostefa, 2008]

The first known corpus of meeting recordings is the ICSI corpus created between 1999 and 2002. It contains many meetings recorded with headset and tabletop microphones. The corpus unfortunately lacks video data or information on the precise speaker positions. Nevertheless, inspired by the work at ICSI, CMU and Microsoft each created a corpus of meetings containing video data and a meeting browser. The Microsoft

corpus included distributed meetings with video and audio broadcasting capabilities – note though it is not publically available.

The next milestone in meeting recordings was the M4 corpus, a collection of short meetings run using scripts and created by the Idiap research institute in Switzerland from 2002 to 2004. The M4 corpus contains audio and video data of meetings, white-board screenshots and, for the first time, audio data from distant microphone arrays. Care was also taken to ensure that the individual audio and video streams were synchronised and aligned.

Following on from the M4 corpus, the AMI, CHIL and later AMIDA corpora were created. Like the M4 corpus, audio and video data were carefully synchronised. The AMI/AMIDA corpora contains scenario and non-scenario, i.e. real meetings. In the AMI and AMIDA scenario meeting recordings 4 participants — a project manager, a marketing expert, an interface designer and an industrial designer — were given the brief of designing a remote control, and information and tasks were fed to them at defined points in the process to influence the scenarios to a certain degree (see Hain et al. [2012] for details).

Confidentiality in recordings is a major issue and can cause problems in meeting recordings, while scenario recordings can considered to be unnatural. The 2012_MMA corpus was created to address these issues and details will be presented in Chapter 4.

2.5.2 Overlapping speech corpora

Overlapping speech causes major degradation in the performance of diarisation (and speech recognition) systems. Shriberg et al. [2001] observed as much as 9–17% overlap in meetings while I have measured 1.6–36.2% overlap in the NIST RT meetings (see Figure 2.13) and 1.9–6.4% overlap in lively discussions in my own meeting recordings (see Table 3.2 in Section 3.2.2.1).

The overlap for two speakers calculated from the results in Figure 2.13 averages out at 23% for the RT06, 11% for the RT07 and 14% for the RT09 meetings. Overlap figures for three or more speakers are 4.7% for the RT06, 1.2% for the RT07 and 2% for the RT09 meetings.

All RT data sets contain mixtures of meetings from the same set of recordings, e.g. CMU, EDI and NIST. These sets of meetings have been recorded within the same

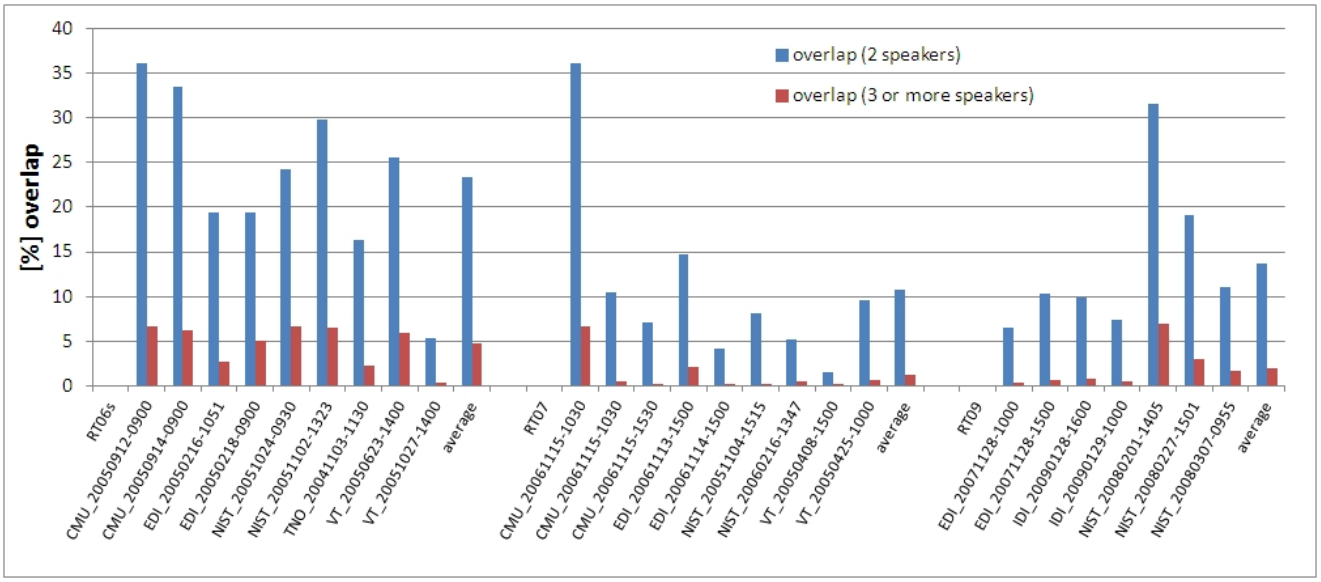


Figure 2.13: Overlap analysis on the NIST RT meetings

framework and are of similar nature so that comparable overlap ratios might be expected. The only known distinction between the sets is that different people worked on the transcription (Steve Renals and Mike Lincoln, personal communication, 19 April 2013).

Comparing these figures with the results from my own recordings of meetings, I conclude that the overlap measured for the RT07 and RT09 meetings indicates a 'normal' meeting while figures for the RT06 meetings are quite extreme and the results from the RT06 data should therefore be accorded less attention.

Overlapping speech can nevertheless deteriorate the performance of diarisation and speech recognition systems and it is important to be able to handle it properly.

While a final solution to the problem needs to be able to deal with overlapping speech in real meetings, it is also important to have data produced in a more controlled environment to develop algorithms that can process overlapping speech.

I am only aware of the existence of one microphone array based corpus of natural overlapping speech dedicated to speech separation experiments, which is the MS-WSJ-AV corpus, presented in Lincoln et al. [2005]. Separating overlapping speech is a research problem that combines acoustic array processing and automatic speech recognition (ASR). Kumatani et al. [2012] states that researchers working in these two areas unfortunately

> "... have failed to adopt each other's best practices. For instance, the array processing community tends to ignore speaker adaptation techniques, which can compensate for mismatches between acoustic conditions during training and testing. Moreover, this community has largely preferred to work on controlled, synthetic recordings, obtained by convolving noise- and reverberation-free speech with measured, static room impulse responses, with subsequent artificial addition of noise, as in the recent Pattern Analysis, Statistical Modeling, and Computational Learning (PASCAL) Computational Hearing in Multisource Environments (CHiME) Speech Separation Challenge (see e.g. Christensen et al. [2010], Barker et al. [2012] and references therein). A notable exception was the PASCAL Speech Separation Challenge 2 [McDonough et al., 2008a, Himawan et al., 2008, Kumatani et al., 2008] which featured actual array recordings of real speakers; this task, however, has fallen out of favor, to the extent that it is currently not even mentioned on the PASCAL CHiME Challenge Web site, nor in any of the concomitant publications. This is unfortunate because improvements obtained with novel speech enhancement techniques tend to diminish, or even disappear, after speaker adaptation; similarly, techniques that work well on artificially convolved data with artificially added noise tend to fail on data captured in real acoustic environments with real human speakers. Mainstream speech recognition researchers, on the other hand, are often unaware of advanced signal and array processing techniques. They are equally unaware of the dramatic reductions in error rate that such techniques can provide in DSR tasks."

The MC-WSJ-AV corpus offers researchers an intermediate task between simple digit recognition and ASR. It consists of sentences read from the Wall Street Journal (WSJ) taken from the test set of the WSJCAM0 database [Robinson et al., 1995]. A total of about 45 speakers, male and female, are recorded in three different scenarios, these are:

- single (stationary) speaker
- two (stationary) overlapping speakers
- single moving speaker

Speakers reading WSJ sentences from prompts are recorded using a headset and lapel microphone and an eight-channel microphone array. In the single speaker scenario and in the overlapping speaker scenario participants are assigned a fixed position for the entire recording while for the moving speaker scenario participants are asked to read from six different, fixed positions.

Fifteen participants were recorded for the single scenario, nine pairs for the overlap-

ping scenario and nine for the moving scenario. Each read approximately 90 sentences which are available for speech separation and recognition experiments. These 90 sentences consist of 17 adaptation sentences, approximately 40 sentences from the 5k-word WSJ corpus and another approximately 40 sentences from the 20k-word WSJ corpus, identical to the WSJCAM0 corpus.

Only two teams entered the PASCAL Speech Separation Challenge 2 (SSC2) and the results are presented in McDonough et al. [2008a] and Himawan et al. [2008].

Himawan et al. [2008] implemented a microphone array beamforming approach to blind speech separation. This method estimates the position of the microphones from the noise field model in order to use these locations to carry out SDB (which requires knowledge of the microphone position) and subsequent blind speech separation. Speech separation is carried out by estimating the speaker location using SRP-PHAT followed by superdirective (MVDR) beamforming and frequency domain masking postfiltering. Himawan et al. [2008] achieved a WER of 54.8% for both speakers and 35.1% for the better speaker on the development set of the SSC2 corpus and a WER of 58.0% for both and 38.3% for the better speaker on the test set.

The system presented by McDonough et al. [2008a] is made up of four principal components: speaker localisation, beamforming, postfiltering and automatic speech recognition. In the proposed system, speaker localisation is carried out by extended Kalman filtering of the TDOA values derived using GCC-PHAT. With the a-priori knowledge of two active speakers, the system then uses the speaker locations to steer two GSC beamformers in the direction of the speakers, while the blocking matrices of the beamformers are optimised according to a minimum mutual information (MMI) criterion. The outputs of the beamformers are further enhanced using postfiltering and binary masking. Finally, an ASR engine based on a weighted FST is used to generate the word hypotheses in four decoding passes where the HMMs of the recogniser are adapted to the speakers.

McDonough et al. [2008a] achieve a WER of 39.6% with the proposed system.

All the results presented so far have been obtained using signal enhancement through acoustic beamforming. Kumatani et al. [2008, 2012] compare the performance of the system presented by McDonough et al. [2008a] on the single speaker task of the MC-WSJ-AV corpus using data from the close talking microphone channel, a single audio channel from the distant microphone array and all array channels. The authors

achieved WERs of 28% on a single array channel; 12.2% with their best acoustic beamforming and speaker adaptation technique on all eight array channels; and 6.5% on the speech from the close talking microphone channel. No results were presented for the overlapping speaker task.

A single distant microphone could be considered as an array of 0 m diameter, although having numerous microphones at the exact same location would still allow for advanced noise reduction but without the capability of extracting TDOA information.

Note that topics such as noise reduction, speech separation and dereverberation are not the main focus of the work presented in this thesis and are therefore not reviewed here. State-of-the-art research in these areas is presented in the relevant places in this thesis with the overall aim of best readability.

2.5.3 Open source software

The work and results presented in this thesis would have not been possible without the use of (open source) software which researchers have made available to the public and in some cases to me individually. This section lists this software along with a brief explanation of the function of the components used if this has not been explained above.

The **QIO-FE** [Adami et al., 2002a] is a collection of tools for robust feature extraction from an audio signal for distant speech recognition (DSR). I used the noise reduction and VAD components of the QIO-FE for the research presented in this thesis. The QIO-FE noise reduction employs VAD to detect silence regions and Wiener-filtering to reduce the noise.

The **BeamformIt** [Anguera, 2006] tool is an acoustic beamforming tool. It is optimised for best diarisation output on the NIST RT challenges. BeamformIt uses GCC-PHAT for TDOA estimation followed by TDOA smoothing after which it carries out delay-sum beamforming.

TDOA estimation for a microphone array requires setting a reference channel. As mentioned above, the microphone with the highest energy level is generally used as the reference. Selecting the reference channel using the signal energy can be difficult (particularly for small array geometries) if there is little measurable energy difference. In addition, for large geometries the microphone closest to the active speaker will show

the highest energy level, therefore requiring constant change of the reference during a multi-party conversation.

BeamformIt therefore uses a different method to determine the reference channel. First, it calculates the cross-correlation of every possible microphone pair for the entire recording after which it selects the microphone with the highest time-averaged value as the reference. The reference microphone is then kept for the entire recording.

The major strength of BeamformIt is its two-phase Viterbi TDOA smoothing scheme, leading to significantly improved DER if the TDOAs are used as an additional (localisation) feature stream for diarisation.

TDOA smoothing in the first phase is carried out on the single channels using the best TDOA from GCC-PHAT as per Equation 2.12 and not just the maximum as per Equation 2.13. First, TDOA values smaller then a threshold are discarded after which the two best TDOA values are chosen from the N-best values using Viterbi decoding.

In the second phase, Viterbi decoding is used again to smooth the two best TDOA values of all channels by finding the smoothest path through the TDOA-HMM. The TDOA-HMM emission probabilities are derived from $log(\hat{G}_{PHAT}(f))$ and the transition weights from the delay distance of two adjacent states. Please refer to Anguera et al. [2007] for details.

The **mdm tools** [McCowan, 2005] are, like BeamformIt, a front-end system for ASR, providing noise reduction as well as delay-sum and superdirective (MVDR) beamforming. TDOA estimation is carried out using GCC-PHAT. The mdm tools select the microphone channel with the highest energy over the entire audio input as the reference for TDOA estimation. This reference microphone is determined once and retained throughout the recording.

Details on GCC-PHAT, delay-sum and MVDR beamforming have been reviewed above (cf. Section 2.1).

SHoUT [Huijbregts, 2006] is a ASR system which also provides a VAD and diarisation component.

The **ICSI** [Friedland et al., 2012] and **LIUM** [Meignier and Merlin, 2010] toolkits are software dedicated to speaker diarisation. The SHoUT and ICSI systems are very similar, performing VAD and diarisation using GMM-HMMs in a bottom-up manner, while the LIUM follows the same principle but processes the data top-down.

The SHoUT, ICSI and LIUM diarisation tools have been reviewed in detail above (cf. Section 2.4).

The Hidden Markov Model Toolkit **HTK** provides sophisticated tools for HMM training, testing and results analysis [Cambridge University Engineering Department (CUED), 2012]. The principal working of its components and its performance have been presented above (cf. Section 2.2).

Chapter 3

MEMS microphones and microphone arrays

The advent of MEMS (micro electro-mechanical systems) technology which combines silicon- and nanotechnology to build microphones has caused a major shift in the electronics market, particularly for consumer electronics. MEMS microphones with either analogue or digital output are not yet good enough for high-fidelity applications, but their advantages outweigh their disadvantages for consumer devices such as mobile telephones or tablets.

In order to research the use of MEMS microphones for meeting capturing I have developed a sequence of digital MEMS microphone arrays, the DMMA.1, DMMA.2 and DMMA.3.

A microphone array is a collection of three or more microphones that work in unison. Although not strictly necessary, it is recommended that the microphone signals are sample-synchronous, i.e. all microphones are sampled at exactly the same time. In addition, superdirective beamforming is only possible for sample-synchronous audio signals and if the geometry of the array, i.e. the relative position of each microphone, is known.

The digital MEMS microphone arrays satisfy these requirements.

3.1 Digital MEMS microphones

MEMS technology is a miniaturisation technology defined as the technique of designing structures well below one micron[1]. Typical MEMS systems today are sensors or actuators. It is common to classify MEMS into their fields of application, with accelerometers, gyroscopes, microphones and pressure sensors being typical sensors and inkjet printheads or fluid accelerators being typical actuators.

The basic building blocks used for MEMS components are not only silicon but also polymers, metals and ceramics. Processing these materials is usually carried out in three basic steps: deposition, patterning and etching. These processing steps are very similar to photographic processes, with the major distinction that for microchip and MEMS manufacturing both chemical as well as mechanical changes are applied to the basic material. Examples of such processes are wet or dry etching of the silicon to corrode a microphone membrane, as shown in Figure 3.1.

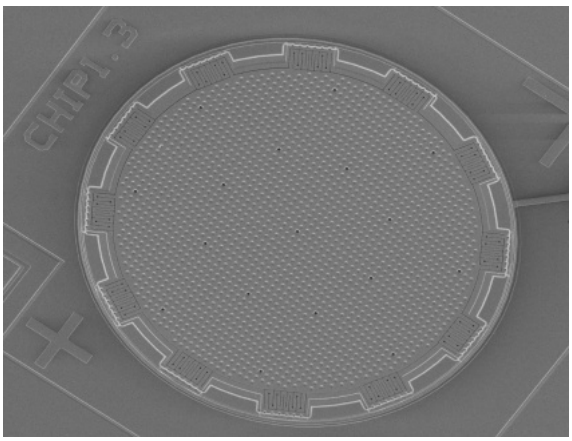
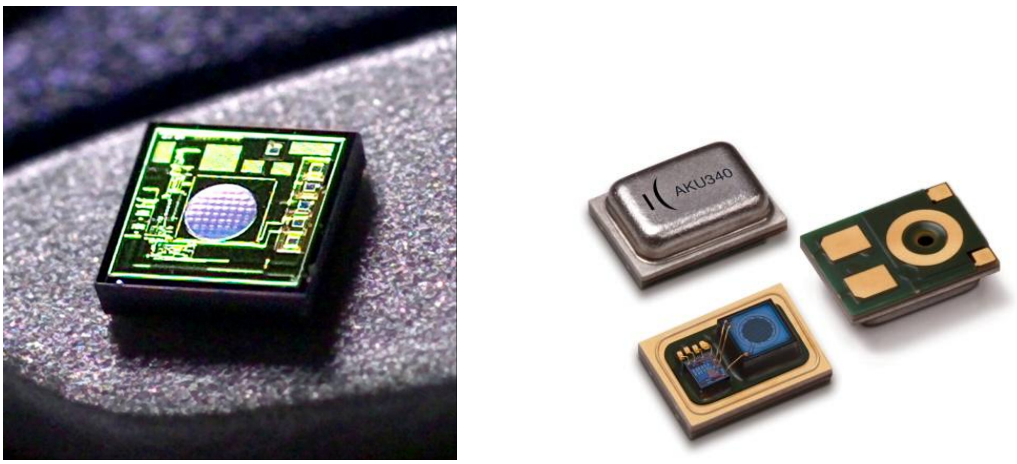


Figure 3.1: MEMS microphone membrane (courtesy of Analog Devices Inc.)

The chemical processes, i.e. the substances, temperatures and pressures necessary to process the silicon for microchip transistor structures are not identical to the ones used to etch a sensor or actuator. Many different methods have been developed allowing manufacturing of single chip MEMS devices at increased process costs. Alternatively, a MEMS system may contain two microchips, both optimised for their individual process requirements at cheaper process costs but higher component count. One-chip and two-chip MEMS microphones are shown in Figure 3.2.

Looking at the basic building blocks of MEMS microphones, many different imple-

[1] http://www.memsnet.org/mems/what_is.html



(a) One chip MEMS microphone (b) Two-chip MEMS microphone

Figure 3.2: MEMS microphone (courtesy of Akustica Inc.)

mentations are possible. The first MEMS microphones available on the market had an analogue output, therefore omitting the analogue to digital conversion (ADC). Later the first digital output MEMS microphones became available. These contain a 1-bit ADC and the output is a pulse-density modulated (PDM) bitstream, typically at 64 times the signal sample rate 64 Fs. Recently Analog Devices Inc. released a MEMS microphone with built-in signal processing, providing the user with a down-sampled digital audio signal at the sample rate using the industry standard I^2S interface. The block diagram of this microphone, the ADMP441, is presented in Figure 3.3.

The building blocks of the ADMP441 digital MEMS microphone are: the microphone membrane, a 1-bit ADC, a digital downsampling filter and the I^2S interface. In addition, a power management block (to provide the microphone with an up-converted and ultra-clean supply voltage between 10 and 15 V) is required plus a hardware control block to configure the microphone.

With most MEMS microphones the microphone membrane is used as a capacitor at a constant charge. Any changes in the capacity caused by air pressure changes on the loose capacitor plate in relation to the second fixed capacitor plate will result in a voltage change which is amplified using a JFET transistor. The power management circuit keeps the charge constant. Figure 3.4 shows this principal schematic and a cross section of it on the microchip.

The sales of MEMS microphones is growing to impressive new levels year by year. In 2012 2.05 billion units were sold, up 57% from 2011, while forecasts for 2016 are for

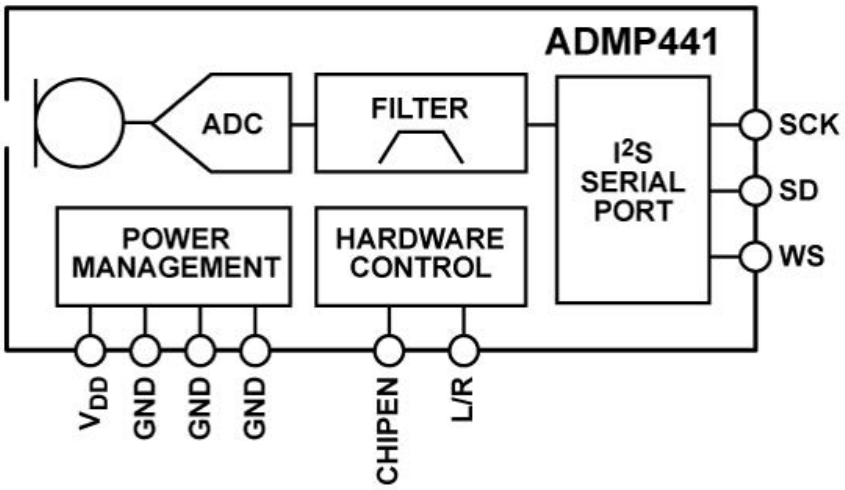


Figure 3.3: ADMP441 digital MEMS microphone block diagram (courtesy of Analog Devices Inc.)

4.65 billion MEMS microphones shipped, predicting three more years of double digit growth[2].

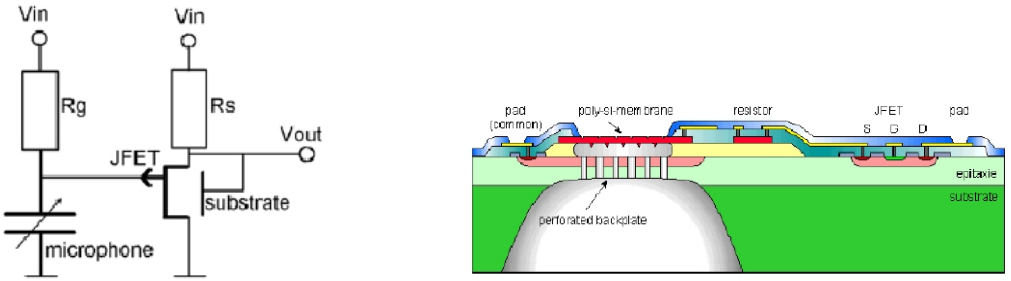
MEMS microphones can be found in smartphones (e.g. Apple iPhone, HTC, Samsung, LG), tablets and ultrabooks (e.g. Apple iPad, Amazon Kindle, Microsoft Surface), laptops (e.g. HP, Dell, Lenovo and Asus), headsets, gaming (e.g. Nintendo and Sony), cameras, televisions and hearing aids.[3]

3.2 Digital MEMS microphone array

Microphone arrays are a key element for data capture in meetings. They allow hands-free sound signal acquisition and are becoming more prevalent in modern consumer devices. The advent of MEMS microphones now enables the construction of cheap commodity microphone arrays [Zwyssig et al., 2010]. MEMS microphones (and microphone arrays) have recently attracted a great deal of interest and might well be the future of sound signal acquisition.

[2]http://www.digikey.com/supply-chain-hq/us/en/articles/semiconductors/mems-microphone-market-revenues-soar-42-in-2012/1497

[3]http://www.electronicsweekly.com/news/business/apple-determines-mems-microphone-market-2013-05



(a) MEMS microphone FE

(b) MEMS microphone FE chip

Figure 3.4: MEMS microphone front-end (FE) and chip cross section (from Brauer et al. [2001], with kind permission of Infineon Technologies AG and IOP publishing)

I designed and built the first digital MEMS microphone array as part of an MSc dissertation during the summer of 2009 using Knowles Acoustics SPM0205HD4 digital MEMS microphones, a Xilinx Spartan 3A FPGA and the Texas Instruments TUSB3200A USB streaming controller[Zwyssig, 2009]. This implementation was very much a prototype and only supported sampling the audio signal at 16000 Hz while requiring post processing of the output signal to extract the raw audio. Speech recognition experiments carried out with the prototype demonstrated the viability of MEMS microphones for speech processing [Zwyssig et al., 2010].

When the USBPAL from Rigisystems and the Analog MEMS microphone ADMP441 became available I redesigned the digital MEMS microphone array as part of my PhD in 2011. The second version of the MEMS microphone array is a USB device supported by Windows PC and MAC OS X and allows recording of eight channels of audio at 8000, 9600, 11025, 12000, 16000, 19200, 22050, 24000, 32000, 38400, 44100 and 48000 Hz. Experiments in speaker diarisation carried out with the DMMA.2 again demonstrated the value of the new MEMS technology. Unfortunately, due to the prototype build of the DMMA.2, the clocking was not as clean as required and the audio recorded with the DMMA.2 contains non-white noise. Standard Wiener-based noise filtering is required to clean up the signal for speech processing.

Built for capturing the audio signal from meetings, the DMMA.1 and DMMA.2 are circular arrays of eight microphones with a diameter of 20 cm, therefore allowing superdirective beamforming due to knowledge of the relative microphone positions and sample-synchronous audio channels. Microphone array beamforming and blind source separation for improved speech processing are also a requirement for hand-held

devices. This requirement and the need to address the noise issues of the DMMA.2 led me to design the third generation digital MEMS microphone array, the DMMA.3, in 2012.

For the DMMA.3 I customised the USBPAL for my requirements and designed an array that would fit onto the back of most mobile devices. The DMMA.3 contains eight microphones on a circle of diameter 4 cm. In order to overcome the problems of spatial aliasing the audio signal sampling rate had to be increased in inverse proportion to the shrinking of the array geometry. Preliminary experiments showed that the ADMP441 microphones work perfectly well at 96 kHz, despite being specified for a maximum sample frequency of 48 kHz. The experiments showed a minimal decrease of the SNR.

3.2.1 DMMA.1

As mentioned above, experiments with the DMMA.1 indicated the suitability of MEMS microphones for speech processing and the results are presented in Table 3.1. These experiments were performed on recordings made in a typical meeting room at the University of Edinburgh using the DMMA.1 and an analogue array of the same dimensions built using Sennheiser MKE 2 microphones. The experiments are identical to the ones used for the MC-WSJ-AV corpus (cf. Section 2.5.1 and Lincoln et al. [2005]).

Six male and six female speakers were recorded reading sentences from the WSJCAM0 test and development sets [Robinson et al., 1995]. All participants were native British English speakers. The set of prompts for each speaker was selected from one of the sets used in WSJCAM0 and typically contained 17 TIMIT style sentences (for adaptation), 40 sentences from the 5,000 word (closed vocabulary) sub corpus of WSJCAM0 and 40 sentences from the 20,000 word (open vocabulary) sub corpus. Each audio channel was recorded as a single wav file, and the files were manually split into individual sentences for recognition.

First, noise reduction and acoustic beamforming was carried out using the mdm tools after which speech recognition was performed using HTK, identical to the setup used for the MC-WSJ-AV corpus [Lincoln et al., 2005].

The results show that the digital array recordings give a substantially increased WER compared with those obtained from the analogue array. The SNR of the digital micro-

Table 3.1: [%] WER on 5k-word MC-WSJ-AV single speaker task for 6 male and 6 female speakers performed on recordings of WSJ sentences using the digital MEMS microphone array DMMA.1 and an equivalent analogue array [Zwyssig et al., 2010]

	Male			Female			Average		
Adaptation Technique	Analogue	Digital	Δ	Analogue	Digital	Δ	Analogue	Digital	Δ
None	30.2	40.7	10.5	36.9	55.1	18.2	33.6	47.9	14.3
MLLR Channel	22.6	27.4	4.7	22.2	32.2	10.0	22.4	29.8	7.4
cMLLR Channel	21.3	26.3	5.0	20.7	29.7	9.0	21.0	28.0	7.0
MLLR Speaker and Channel	18.2	20.7	2.5	19.4	25.9	6.6	18.8	23.3	4.5

phones is lower than that of the analogue microphones, meaning that the audio from the digital array is less well matched to the recognition models which are trained on speech from high quality analogue headset microphones. This is likely to be the reason for the observed decrease in accuracy.

In order to address the mismatch between training and test data, three experiments were conducted in which the recognition models were adapted to the acoustic properties of the recordings using MLLR and cMLLR adaptation (cf. Section 2.2.7).

First, the acoustic models were adapted to the channel by pooling the 17 adaptation sentences recorded for each speaker to produce transforms specific to the digital and analogue arrays. Recognition was then performed on the 5k-word data from the matched array and the results are shown as 'Channel' in Table 3.1. As expected, the adaptation gives decreases in WER for both analogue and digital arrays. More importantly, the absolute difference in WER between the analogue and digital arrays is reduced by nearly 50%, from 14.3% to 7.4%. This suggests that, although the quality of the output from the digital array is lower than that of the analogue array and therefore not as closely matched to the close talking models, it still contains much of the speech information required to perform recognition, providing the models are matched to the microphones. Performing cMLLR channel adaptation in a second experiment resulted in further decreases in the WER.

Third, experiments were performed in which the models were adapted to the speaker and to the channel by defining the adaptation sets as those sentences recorded from the same speaker on the same array. In this case the absolute difference in the WER between the analogue and digital arrays was further reduced by about 40% to 4.5%.

Looking at these results I conclude that the most probable cause for the decreased

WER performance of the digital MEMS microphone array is the increased SNR of the microphones and that simple MLLR and cMLLR adaptation can be used to achieve almost identical speech recognition performance from the digital MEMS and analogue microphones.

See Zwyssig [2009] for a full description of the work and Zwyssig et al. [2010] for a detailed summary.

3.2.2 DMMA.2

Initial research using MEMS microphones for signal acquisition in meeting rooms produced very promising results as demonstrated above in Section 3.2.1. The DMMA was therefore redesigned and improved, and the second version, DMMA.2, allows recording of eight microphone channels at sample rates from 8 to 48 kHz.

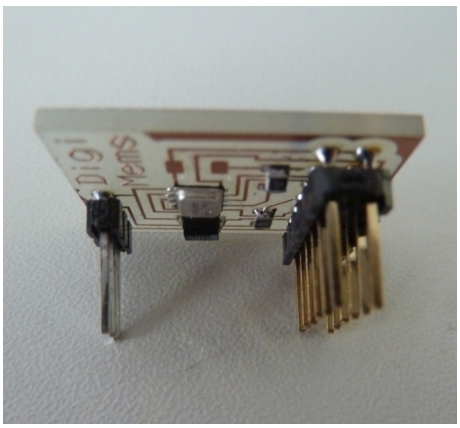
The DMMA.2, shown in Figure 3.5, is built using ADI ADMP441 omnidirectional MEMS microphones[4] with bottom ports and I^2S outputs and the Rigisystems USB-PAL[5], a USB 2.0 multi-channel audio interface for Windows PC and MAC OS X.

The digital MEMS microphones are mounted on daughterboards which themselves are mounted on a disk-shaped motherboard, both of which I designed as part of my PhD. The daughterboards contain the microphones, a de-coupling capacitor for improved power supply and a resistor for clock signal termination. Eight daughterboards are mounted on the motherboard so as to be placed equidistant on a circle of diameter 20 cm, identical to the analogue array used for the AMI/DA recordings [McCowan et al., 2005]. The motherboard itself is then plugged onto the USBPAL, resulting in an 8-channel microphone array USB device ready for recording, as shown in Figure 3.5.

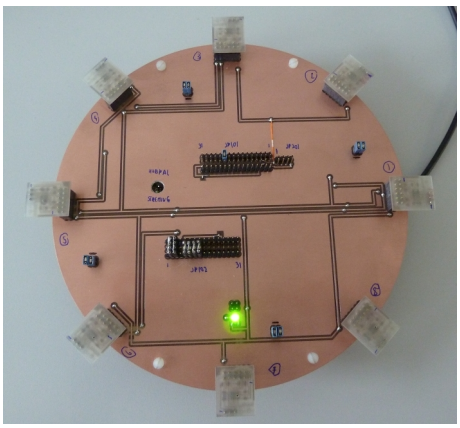
Digital MEMS microphones have significantly lower intrinsic SNRs compared to analogue microphones. Tests on the microphones used in the DMMA.2 show that this sensor noise is not white as would be expected. While SNR and THD (total harmonic distortion) measurements show the microphones to be within specification, the MEMS microphones output a non-white chirping noise which originates from poor PCB layout due to using prototype technology, i.e. lack of clock shielding, lack of ground planes and undefined clock and signal routing impedances.

[4] http://www.analog.com/en/mems-sensors/microphones/admp441/products/product.html

[5] http://www.rigisystems.net/



(a) Microphones on daughter boards



(b) Complete microphone array

Figure 3.5: The digital MEMS microphone array DMMA.2

3.2.2.1 AD_IMR corpus

The DMMA.2 and an array with identical geometry constructed using high SNR analogue microphones were used to simultaneously record six research meetings of around one hour in length. The recordings were made in a typical meeting room at the University of Edinburgh. The analogue array is identical to that used in the AMI meeting corpus recordings and is fully documented in Section 4.3 and in Hain et al. [2010].

From each of the recordings, a continuous ten to fifteen minute segment containing lively discussion was selected, creating a total of approximately 78 minutes of recordings. These extracts were annotated to show speech/non-speech events and for each speech segment the speaker ID was marked. Both overlapping speech (where more than one speaker is talking simultaneously) and back channels (short interjections from listeners, typically indicating agreement or disagreement with the main speaker) were included in the annotations. The annotations was formatted using the RTTM specification, as defined by NIST[6], allowing scoring of automatically generated diarisation annotations using the standard NIST evaluation tools. Details of the meeting recordings contained in the corpus, named AD_IMR, are listed in Table 3.2.

3.2.2.2 Methods

Experiments were conducted to investigate the effect on the diarisation task of using the digital array and superdirective beamforming. Two state-of-the-art diarisation systems

[6] http://www.itl.nist.gov/iad/mig/tests/rt/

Table 3.2: Summary of AD_IMR corpus meeting recordings

Recording	**Length** [s]	**# of speakers**	**min/avg/max speech segment length** [s]	**# of segments**	**Overlap** [%]
rec14june2011	825	5	0.14 / 1.94 / 8.5	351	3.0
rec15june2011	804	7	0.23 / 1.92 / 18.1	314	4.2
rec21june2011	630	4	0.21 / 1.71 / 10.4	286	1.0
rec22june2011	856	4	0.15 / 1.50 / 10.0	313	2.4
rec28june2011	607	4	0.19 / 1.74 / 8.0	245	1.9
rec29june2011	914	6	0.21 / 1.47 / 13.8	501	6.4

were employed to compare the error rates achieved by the low SNR recordings from the DMMA.2 with recordings of the same meeting from the analogue array. Using both smoothed and unsmoothed delay estimates, I then compared diarisation errors from the MVDR beamformer and the currently used delay-sum beamformer.

Figure 3.6 shows the data flow for the experiments. Initially, Wiener-filter-based noise reduction using the QIO-FE was applied to the analogue and digital microphone signals and both smoothed and unsmoothed TDOA values for each of the channels calculated using BeamformIt and the mdm tools. Enhanced signals were then generated using three techniques: (1) delay-sum beamforming with smoothed delay estimates, (2) superdirective beamforming with unsmoothed delay estimates and (3) superdirective beamforming with smoothed delay estimates.

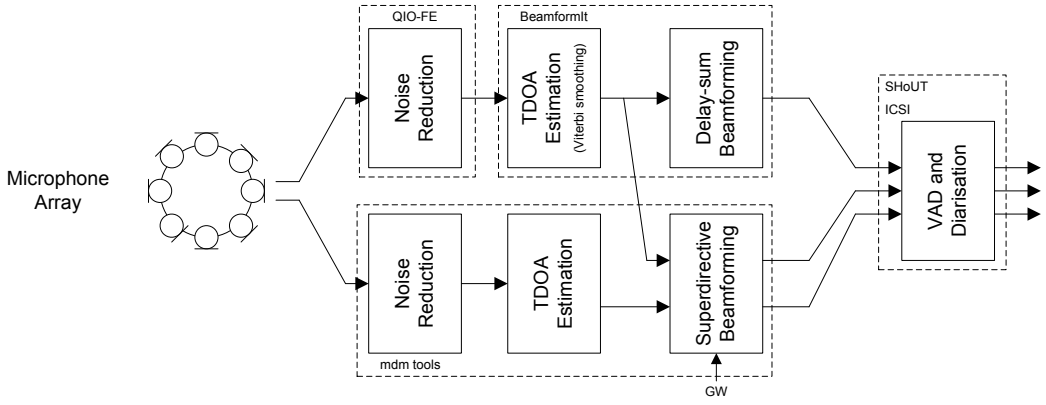


Figure 3.6: Flow diagram to investigate the effect of using the digital array and superdirective beamforming on the diarisation task

Details of the TDOA smoothing method are given in Section 2.5.3 and Anguera et al. [2007].

Speaker diarisation was then performed on the three enhanced signals using two diarisation systems, the SHoUT speech recognition toolkit and the ICSI speaker diarisation system[7].

The ICSI system was made available to me thanks to Gerald Friedland.

3.2.2.3 Results

The missed speech (MS), false alarms (FA), voice activity detection error rate (VER) and diarisation error rate (DER) results for the six meetings in the AD_IMR corpus are given in Table 3.3.

Table 3.3: [%] DER, DER, FA and MS for delay-sum (DSB) and superdirective (SDB) beamforming for analogue and digital arrays using the ICSI and SHoUT diarisation systems on recordings from the AD_IMR corpus

		SHoUT				**ICSI**			
		DER	VER	FA	MS	DER	VER	FA	MS
DSB	analogue	20.5	2.3	1.3	1.0	22.5	2.2	1.3	0.9
(TDOA smoothing)	digital	21.9	3.0	1.5	1.5	22.8	2.9	1.5	1.4
SDB	analogue	29.2	4.8	3.5	1.3	28.2	4.7	3.5	1.2
GW=0.6	digital	35.2	4.9	3.0	1.9	30.3	4.8	3.1	1.7
modified SDB	analogue	23.1	3.6	1.9	1.7	21.6	3.5	1.9	1.6
GW=0.6 (TDOA smoothing)	digital	25.5	3.7	1.6	2.1	28.8	3.7	1.7	2.0

The results show that, for diarisation, the new digital microphone array compares well with the analogue array despite the reduced SNR, producing only marginally increased error rates. This result suggests that MEMS microphone technology provides a viable alternative to analogue devices for speech data capture.

Table 3.3 also shows that superdirective beamforming results in a decrease of the diarisation performance compared to Viterbi smoothing of the TDOA coefficients and delay-sum beamforming. Using Viterbi smoothing of the TDOA coefficients and superdirective beamforming leads to an improved DER, though this does not match the results from simple delay-sum beamforming.

[7]The implementation of the ICSI system evaluated here only uses acoustic features, in contrast to the system used in the ICSI submission to the NIST RT09 evaluation which incorporates TDOA features directly as an input to the diarisation system.

The main problem of using GCC-PHAT-based TDOA estimation for beamforming on the diarisation task is that the values of the TDOA and therefore the direction of the acoustic beam is undetermined, i.e. neither the direction of the beam nor the positions of the active speakers are known at any particular time. Viterbi smoothing produces more stable TDOA values and interrupting speakers are therefore ignored. This results in improved diarisation performance.

Both the SHoUT and ICSI diarisation tools have been optimised for GCC-PHAT-based TDOA estimation and delay-sum acoustic beamforming. Using superdirective beamforming for diarisation will result in a different acoustic speech signal, particularly in regions of overlapping speech and speaker changes. Superdirective beamformers will also remove acoustic information from the sidelobes which may lead to an increased DER as the diarisation tools are tuned to acoustic output from a delay-sum beamformer.

Note also that the TDOA smoothing method was optimised for diarisation performance using a delay-sum beamformer. An alternative TDOA coefficient optimisation may well lead to improved diarisation when using superdirective beamforming.

Analysing the effect of the superdirective beamformer white noise gain constraint GW on the diarisation error rate, I found that tuning GW does not have an effect on the DER of the digital MEMS microphone but decreasing GW increases the DER of the analogue array as shown in Figure 3.7.

Note that setting $GW = 8$ for an eight-channel array results in the superdirective beamformer being a simple delay-sum beamformer.

Figure 3.7 shows that the white noise gain GW has little effect on the digital MEMS microphone array which might be due to the non-white noise of the DMMA.2.

Increasing the directionality of the beam decreases the performance of the analogue array, indicating again that removing the sidelobes of the acoustic beam decreases the diarisation performance.

I conclude that SDB beamforming is not desirable for best diarisation performance. See [Zwyssig et al., 2012b] for details.

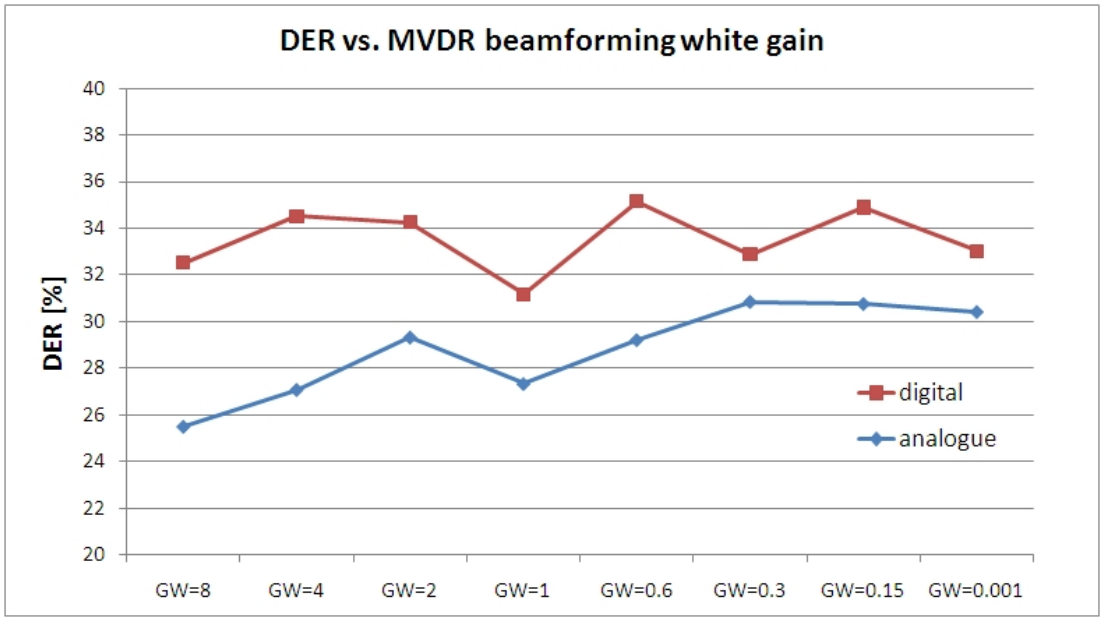


Figure 3.7: Effect of white noise gain constraint GW on [%] DER from diarisation experiments on the AD_IMR corpus

3.2.3 DMMA.3

As mentioned above, almost all mobile consumer devices nowadays contain one, or possibly more microphones. Research has shown that multiple mobile devices can be combined to build an ad-hoc microphone array [Hennecke and Fink, 2011]. Unfortunately, using this configuration of mobile devices the individual signals are not aligned and their clocks and signals will drift. In addition, the position of the devices (and therefore microphones) is not known, making superdirective beamforming impossible [Elko and Meyer, 2008]. Any beamforming algorithm apart from delay-sum beamforming requires knowledge of the microphone position and synchronised audio samples. If only one of these requirements is given, then algorithms exist that allow the other to be estimated (see e.g. McCowan et al. [2008]) but these approaches are often not practical.

I have therefore designed the third generation digital MEMS microphone array DMMA.3 which would fit onto most current smart mobile devices. The digital MEMS microphone array DMMA.3 is shown in Figure 3.8.

The DMMA.3 has been designed by myself and 10 units have been manufactured and tested by Rigisystems[8].

[8] http://www.rigisystems.net/

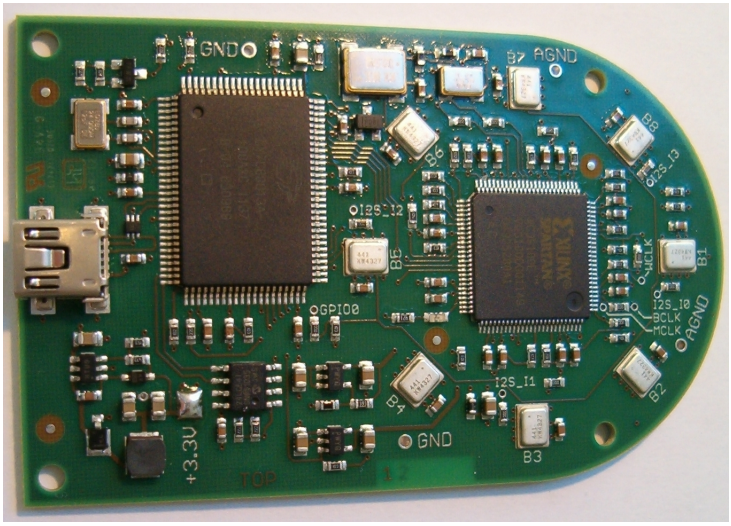


Figure 3.8: DMMA.3 (underside) – the upper side is empty to allow unobstructed sound wave propagation

Using the DMMA.3 and DMMA.2 and equivalent analogue arrays of identical sizes, the research questions to be answered are: what are the effects of the MEMS microphone performance, the array size and the sample rate on state-of-the-art speech processing algorithms.

3.2.4 Verifying the microphone SNR

The performance of a microphone is usually measured using acoustic parameters such as microphone self-noise, sensitivity, frequency response, directionality and maximum input sound pressure as well as electrical parameters such as power consumption and impedance.

If combined with an ADC the microphone system performance is also defined using additional parameters such as the THD (total harmonic distortion), SNR or power supply rejection ratio (PSRR).

MEMS microphones are defined as a system and their performance is given in a datasheet, provided by the supplier. Analogue microphones, on the other hand, require an external amplifier, ADC and interface. These components are specified separately. Comparing the performance of MEMS microphone systems with analogue microphone systems is therefore not straightforward.

Nevertheless, the MEMS industry is close to producing MEMS microphones that

achieve comparable performance to analogue microphones. The main performance degradation of MEMS microphones used for the experiments presented in this thesis is the significantly reduced SNR[9].

This section explains the meaning of the 'significantly reduced SNR' of digital MEMS microphones compared to analogue microphones and how the SNR is measured. Please note that the only performance degradation perceivable using the digital MEMS microphone arrays is the non-white noise of the DMMA.2.

Analog Devices Inc. give a good explanation of the SNR as

> "... the ratio of a reference signal to the noise level of the microphone output. This measurement includes noise contributed by both the microphone element and the ASIC incorporated into the MEMS microphone package. The SNR is the difference in decibels between the noise level and a standard 1 kHz, 94 dB(SPL) reference signal.
>
> SNR is calculated by measuring the noise output of the microphone in a quiet, anechoic environment. This specification is typically presented over a 20 kHz bandwidth as an A-weighted value (dBA), which means that it includes a correction factor that corresponds to the human ear's sensitivity to sound at different frequencies. When comparing SNR measurements of different microphones, it is important to make sure that the specifications are presented using the same weighting and bandwidth; a reduced bandwidth measurement makes the SNR specification better than it is with a full 20 kHz bandwidth measurement."

The two microphones used for the research presented in this thesis are the Sennheiser MKE 2 sub-miniature clip-on lavalier microphone and the Analog Devices Inc. ADMP441 omnidirectional digital MEMS microphone.

Sennheiser's technical datasheet for the MKE 2 specifies:

- sensitivity in free field, no load (1 kHz) of 5 mV/Pa $\pm$ 3 dB
- equivalent noise level of 26 dB

Analog Devices specifications for the ADMP441 are:

- high SNR of 61 dB(A)
- high sensitivity of -26 dB(FS)

Lewis [2012] explains that the sensitivity for a digital microphone defines when the maximum possible output or maximum digital number is reached, i.e. the maximum

[9] http://www.eetimes.com/document.asp?doc_id=1280170

output at 94 dB(SPL) with a sensitivity of -26 dB translates into the maximum possible output of 94 dB + 26 db = 120 dB(SPL).

The noise level is calculated as the reference level (which is 94 dB(SPL)) minus the specified SNR, that is 94 dB - 61 dB = 33 dB(SPL).

This compares to the 26 dB noise level of the Sennheiser MKE 2 microphone. The maximum possible output of the MKE 2 is not specified as it is defined by the amplifier. The amplifier used for the research presented in this thesis is the MOTU 8pre which provides eight microphone pre-amplifiers and analog to digital converters, allowing 40 dB of gain on the inputs.

The input gain is set to approximately 15 dB so as to match the output of the digital MEMS microphones within $\pm$6 dB, as shown in Table 3.4.

In order to verify the specifications given by Sennheiser and Analog Devices Inc., I analysed the output of the analogue and MEMS microphones under two conditions, (1) when stimulated with the Brüel & Kjær (B&K) sound calibrator type 4231 which produces a calibrated sine wave of 1000 Hz at a sound pressure level of 94 dB(SPL) and an accuracy of $\pm$0.2 dB and (2) with no signal, i.e. self-noise.

The resulting output levels and frequency responses are shown in Table 3.4 and Figures 3.9 and 3.10.

Table 3.4: Analogue (Sennheiser MKE 2) and digital (ADMP441) microphone V_{rms} measurements

Microphone	**Diameter** [cm]	**Fs** [kHz]	V_{rms}
Analogue	20	48	0.089
Analogue	4	96	0.057
Digital	20	48	0.037
Digital	4	96	0.039

I recorded about five seconds of audio with the four different microphone arrays at two different samples rates. The V_{rms} output measured and presented in Table 3.4 shows that the gain of the four arrays are matched within one bit of their outputs, i.e. 6 dB or 0.5.

It is now of interest to see whether the increased noise level of the MEMS microphones shows in the frequency responses given in Figures 3.9 and 3.10.

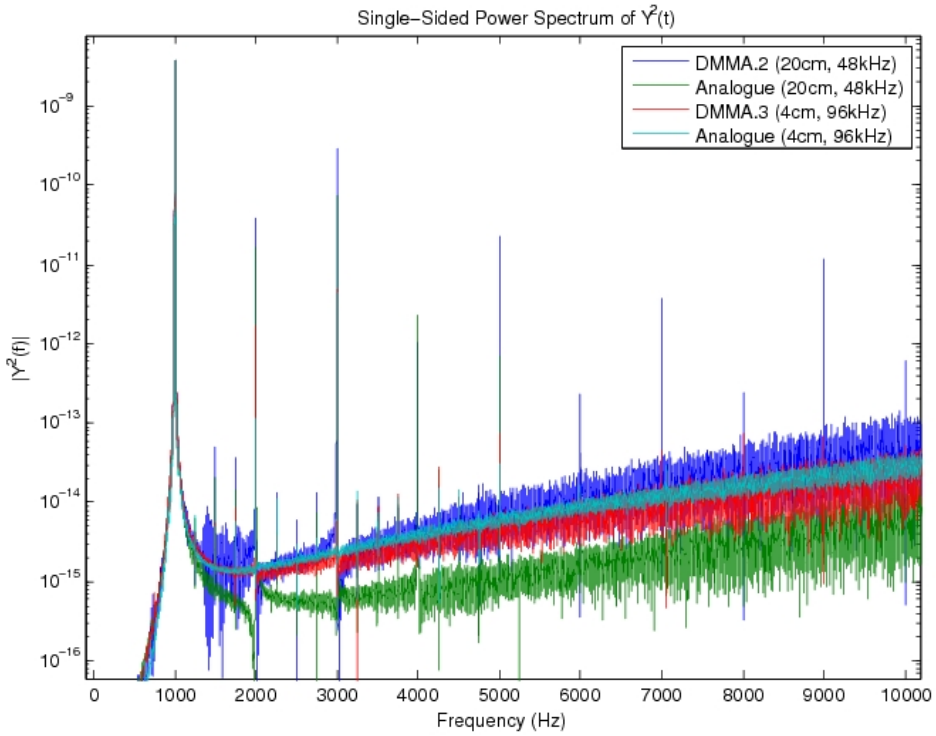


Figure 3.9: Microphone calibration signals (with B&K reference signal)

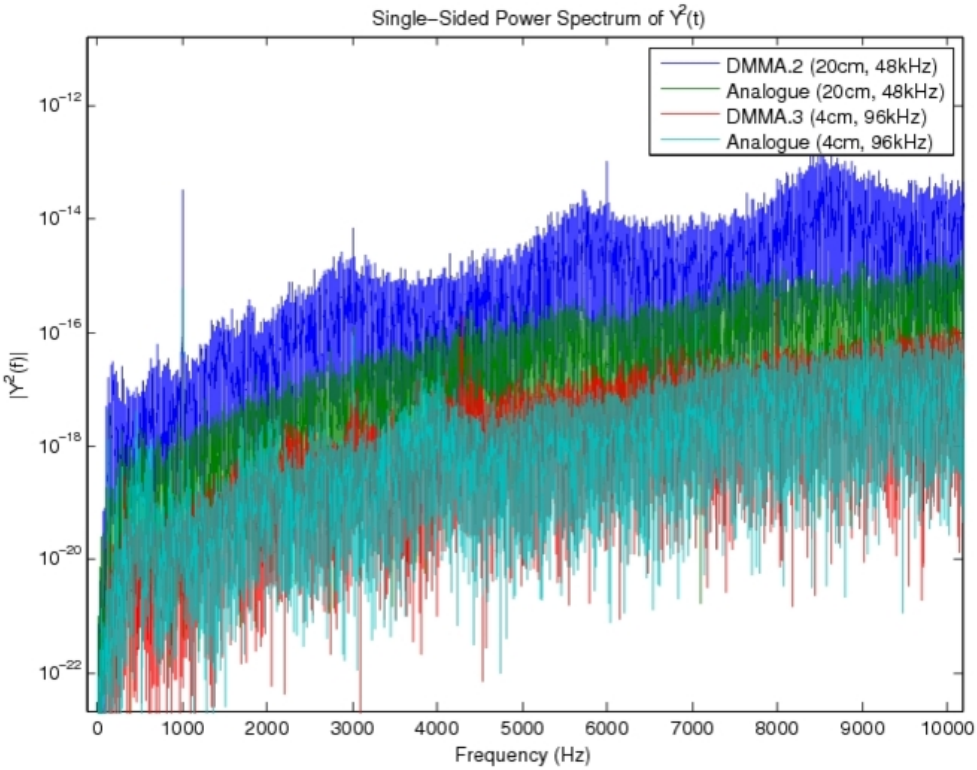


Figure 3.10: Microphone calibration signals (with self-noise)

For the frequency responses as shown in Figures 3.9 and 3.10, I calculated the FFT using Blackman windowing of one second of the recorded output signal from the Brüel & Kjær sound calibrator type 4231.

The best SNR can be observed for the Sennheiser MKE 2 (analogue) microphone, sampling at 48 kHz. This SNR degrades slightly when recording at 96 kHz. The digital MEMS microphone performs slightly less well (DMMA.3) and the 'design-issue' of the DMMA.2 shows clearly in the significantly higher noise floor and heavily distorted frequency response, particularly between 1500 and 2000 Hz.

Without the DMMA.2 'design-issue', its noise level ought to be better than that of the DMMA.3.

Please note that the SNR is measured as the accumulated noise over the complete input signal frequency range. The noise levels of the individual FFT frequency bins displayed for higher sample rates will therefore naturally be much lower than for lower sample rates, nevertheless resulting in a comparable SNR performance. Looking at Figures 3.9 and 3.10, this means that the SNR for Fs = 96 kHz is not better than for Fs = 48 kHz but worse. Figures 3.9 and 3.10 can only serve to compare the analogue vs. digital microphones but not the different sampling rates of 48 kHz vs. 96 kHz.

Figures 3.9 and 3.10 only show the frequency responses from 0 to 10,000 Hz in order to maximise readability around 1000 Hz. The frequency responses above 10,000 Hz are as expected for a 1-bit SDM (sigma-delta modulator) ADC and do not add any further information.

As described in this chapter, the digital MEMS microphones have shown their great potential, encouraging me to proceed with the analysis of speech processing algorithms using digital MEMS microphones. For this I have recorded a corpus using four microphone arrays to carry out further experiments. This corpus and the experiments performed using it are presented in the next chapter.

Chapter 4

2012_MMA corpus

This chapter presents the 2012_MMA (multi microphone array) audio-visual corpus of read and conversational speech recorded with microphone arrays built using digital MEMS and analogue microphones.

Speech research is inevitably driven by the data that is available. The 2012_MMA corpus bridges the gap between existing corpora recorded with conventional analogue microphone arrays and speech processing algorithms used on mobile devices which increasingly use MEMS microphones.

This corpus of MEMS microphone recordings makes audio data available for researchers to guarantee that their research results are fit for this changeover. It is designed to support research in:

- noise reduction
- speaker localisation
- acoustic beamforming
- speech separation
- blind source separation (BSS)
- voice activity detection (VAD)
- speaker diarisation
- speech recognition
- discourse analysis
  - agent goals/beliefs/desires/interaction (prioritisation/goals/conflict/game theory)

The different components of the 2012_MMA corpus are presented in Table 4.1.

Table 4.1: Overview and brief description of the 2012_MMA corpus

Data set	Environment	# of spkr	Brief description	Reference
WSJ	IMR hemi-anechoic	1 1	Wall Street Journal	text (fixed) speaker position
MSWSJ	IMR hemi-anechoic	2 2	Multiple Speaker Wall Street Journal	text (fixed) speaker position
Settlers	IMR hemi-anechoic	6 4	Settlers of Catan	VAD/diarisation/(text) (fixed) speaker position
Wargames	IMR	4 (+)	Warhammer 40000	VAD/diarisation/(text) speaker position

The 2012_MMA corpus contains four different data sets, the WSJ, MSWSJ, Settlers and Wargames recordings. Note that all participants were native British English speakers.

4.1 WSJ and MSWSJ data sets

In the WSJ recordings 12 participants, 6 males and 6 females, were recorded reading Wall Street Journal sentences prompted on a screen in both an instrumented meeting room (IMR) and a hemi-anechoic chamber[1]. This allows research in algorithms and methods for noise reduction, dereverberation and echo cancelling (by e.g. convolving the clean speech with any room impulse response or noise), where the output performance can be evaluated using ASR software.

For the WSJ and MSWSJ data sets of the 2012_MMA corpus, participants recorded 17 adaptation sentences, approximately 40 sentences from the 5k-word WSJ corpus and another approximately 40 sentences from the 20k-word WSJ corpus, resulting in approximately one minute of adaptation data and two times seven minutes of test data per speaker, providing about one hour of audio data for each data set. This is summarised in Table 4.2.

[1] An anechoic chamber, that is a non-echoing or echo-free chamber, is a room designed to absorb reflections of sound (or radio waves) in order to conduct experiments in nominally "free field" conditions. Full anechoic chambers aim to absorb energy in all directions and the device-under-test and necessary test equipment are placed on a wire-mesh in the centre of the room. Semi-anechoic chambers are built with energy absorbing walls and ceiling but have a solid floor onto which the device-under-test and any equipment are placed.

Table 4.2: Overview and brief description of the WSJ and MSWSJ data subsets from the 2012_MMA corpus

Data subset	Number of sentences	Description
adap	approximately 17	TIMIT style, for adaptation [Garofolo, 1993]
5k	approximately 40	5,000 word (closed vocabulary) sub corpus of WSJCAM0 [Robinson et al., 1995]
20k	approximately 40	20,000 word (open vocabulary) sub corpus of WSJCAM0

Participants were recorded using five different microphone arrays plus a panoramic video camera, as shown in Figure 4.1. This basic recording setup was identical for the entire 2012_MMA corpus.

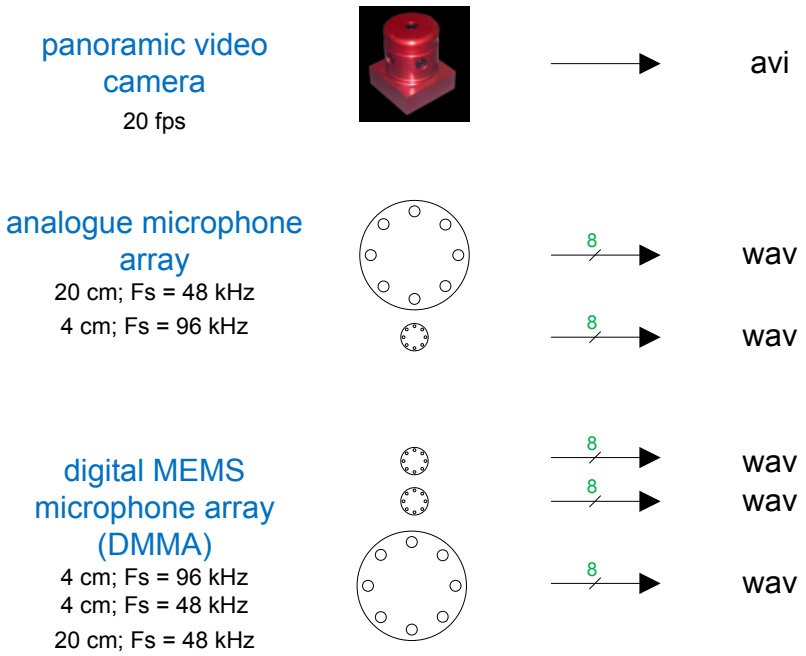


Figure 4.1: Basic recording setup for 2012_MMA corpus

For the MSWSJ data set, six pairs of speakers, male+male or female+female, were recorded simultaneously reading different Wall Street Journal sentences prompted on a screen. As with the WSJ recordings, adaptation, 5k-word and 20k-word sentences were recorded in both an IMR and a hemi-anechoic chamber.

The MSWSJ data allows research in blind source and speech separation as well as noise reduction, dereverberation and echo cancelling. As with the WSJ data, the output performance of any algorithm working on this data can be measured in WER using an ASR system.

The setups for the WSJ and MSWSJ recordings are shown in Figures 4.2, 4.3 and 4.4.

Figure 4.2 shows the setup and measurements of the recordings for the WSJ data sets for both the IMR and hemi-anechoic chamber. Participants were recorded using five different microphone arrays where, for each array, eight microphones are placed counter-clockwise on a circle and channel one is marked with a red triangle.

Figure 4.3 shows the setup and measurements of the recordings for the MSWSJ data sets and Figure 4.4 gives an idea of the room layout.

4.2 Settlers and Wargames data sets

As mentioned above (cf. Section 2.5), speech research corpora try to balance the difficulty of the task and the naturalness of the data. A good corpus therefore provides simple tasks such as sentences read from scripts as well as natural conversational speech, such as occurs in meetings. Meeting conversations unfortunately contain confidential information which might need to be removed by manual labour which is expensive and time consuming.

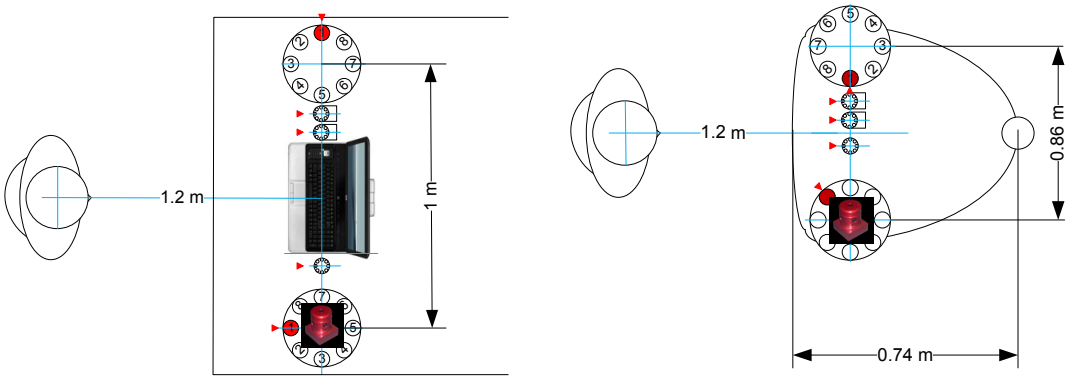
Two more data sets have been added to the 2012_MMA corpus in order to address these problems. These are the Settlers and Wargames recordings.

The **Settlers** recordings contain four to six players, male and female, playing *Settlers of Catan*, a strategic board game, in both an IMR and an hemi-anechoic chamber. The Settlers data allows research in blind source separation, speaker localisation and speech separation as well as noise reduction, dereverberation and echo cancelling. Using the output of an ASR system on the Settlers dialogue (i.e. transcripts) allows discourse analysis and research on prioritisation, goal, conflict and game theory, also in the context of Software agents.

Finally, the **Wargames** recordings contain four players playing *Warhammer 40,000* in an IMR at the Speech and Hearing Group, Department of Computer Science, University of Sheffield[2].

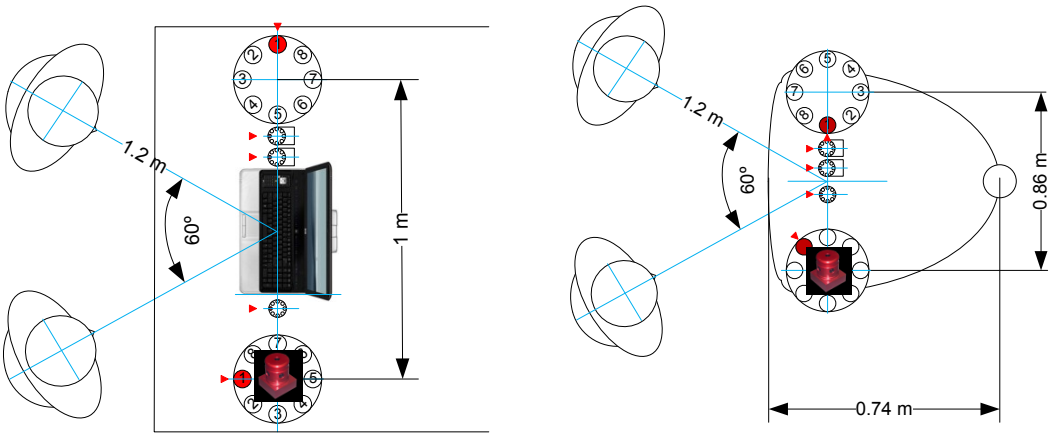
Given the reference transcript of the players and their location, the Wargames are the first recordings of moving speakers interacting in a lively manner, recorded with close

[2]Thomas Hain and his group kindly invited me to help carry out the experiments which were recorded with Edinburgh and Sheffield University and EADS IW equipment.



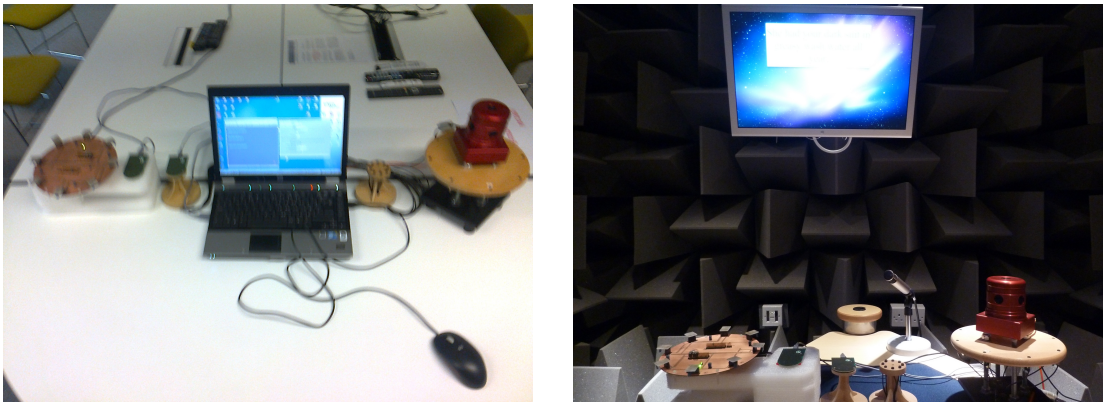
(a) WSJ (b) WSJ_anechoic

Figure 4.2: Single speaker recording setup for the WSJ and WSJ_anechoic data sets



(b) MSWSJ (b) MSWSJ_anechoic

Figure 4.3: Overlapping speakers recording setup for MSWSJ and MSWSJ_anechoic data sets



(a) IMR (b) Hemi-anechoic chamber

Figure 4.4: Room layout for the IMR and hemi-anechoic chamber

and distant microphones. This allows research in blind source separation, speaker localisation and speech separation, noise reduction, dereverberation and echo cancelling, voice activity detection, speaker diarisation and automatic speech recognition as well as complex discourse analysis beyond the possibilities of the Settlers games. Most discussions and disputes in the *Warhammer 40,000* game, for example, are on the rules of the armies, and an intelligent system could find the correct entry in the rulebook when such a dispute is taking place.

4.3 Recording equipment

The WSJ, MSWSJ and Settlers data sets were recorded with five microphone arrays and a panoramic video camera[3]. The microphone arrays used were

- eight channel, 20 cm analogue array sampling at 48 kHz
- eight channel, 4 cm analogue array sampling at 96 kHz
- eight channel, 20 cm digital array sampling at 48 kHz (DMMA.2)
- eight channel, 4 cm digital array sampling at 96 kHz (DMMA.3)
- eight channel, 4 cm digital array sampling at 48 kHz (DMMA.3)

This is shown in detail in Figure 4.5.

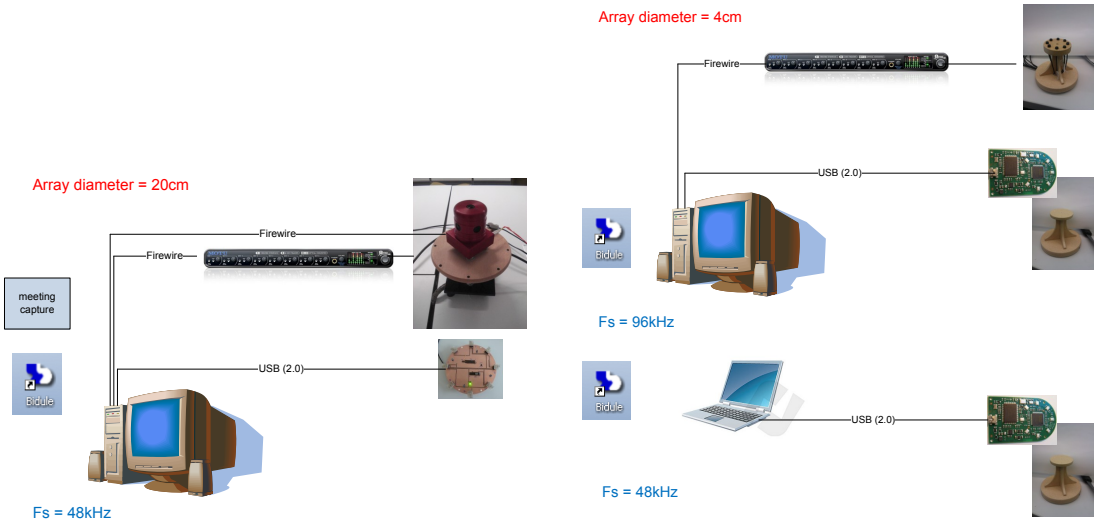


Figure 4.5: Detailed recording setup for the 2012_MMA corpus

[3]The video recordings for the WSJ and MSWSJ data sets are not part of the released corpus as the participants do not move and therefore do not add any information to the audio.

The analogue microphone arrays require external analogue to digital conversion, while this is integrated in the digital ones.

The analogue microphone array setup comprises:

- (8 x) Sennheiser MKE 2-P-C microphone
- Motu 8pre Firewire audio interface[4]
- Firewire interface on PC running Microsoft XP
- AMI/DA meeting recording software for 20 cm array
- Bidule recording software for 4 cm array[5]

The digital microphone array setup consists of either the DMMA.2 or the DMMA.3 and comprises:

- (8 x) ADI ADMP441 digital MEMS microphone
- Rigisystem's USBPAL high speed multi-channel high performance audio interface
- USB interface on PC running Windows XP
- Bidule recording software

The panoramic video recording equipment comprises:

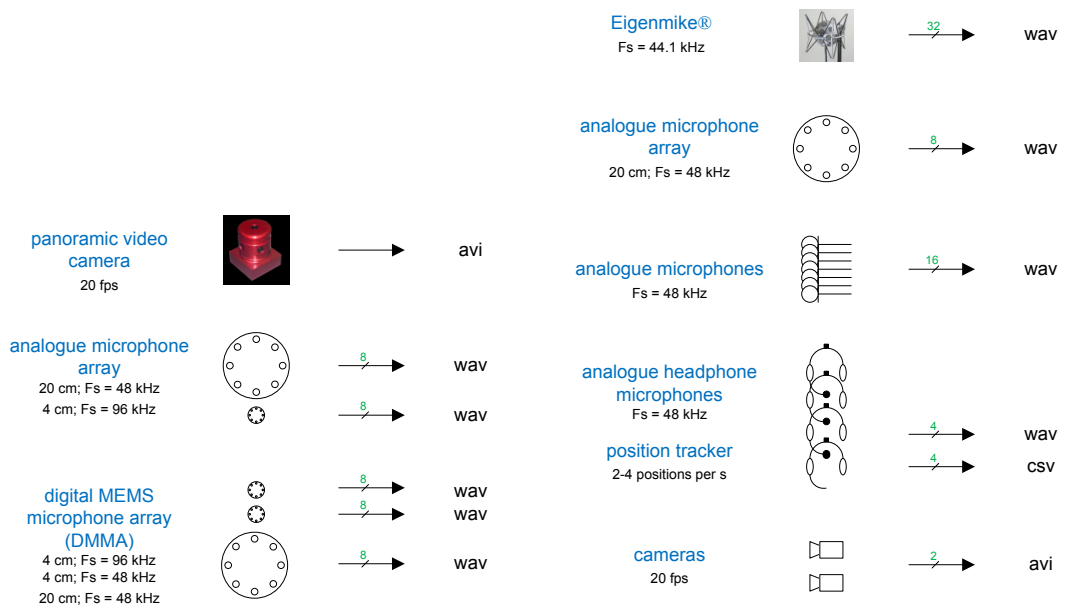
- Point Grey Research's Ladybug2 (1394b) spherical vision camera[6]
- Firewire interface on PC running Windows XP
- AMI/DA meeting recording software

In addition, for the Wargames recordings, several fixed microphones were placed in the room and two cameras recorded the games from opposite corners of the room. This was supplemented by the spherical Eigenmike®[7] placed onto the table, recording 32 analogue channels at 44.1 kHz. Most importantly, each speaker wore a headset and a localisation tracker[8], allowing research in speaker localisation, speaker tracking, etc.

The complete recording suite for the Wargames data set is shown in Figure 4.6.

The WSJ and MSWSJ recordings naturally contain word level transcription as they are sentences read from script. For the Settlers and Wargames recordings I plan to provide diarisation references which can be used for evaluation of VAD and diarisation algorithms verified by VER and DER measurements. Thomas Hain, Charles Fox and

[4] http://www.motu.com/products/motuaudio/8pre
[5] http://www.plogue.com/products/bidule/
[6] http://www.ptgrey.com/products/ladybug2/ladybug2_360_video_camera.asp
[7] http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html
[8] http://www.ubisense.net/en/



(a) 2012_MMA basic recording setup

(b) Wargames recording setup

Figure 4.6: Complete recording setup for Wargames data set from the 2012_MMA corpus

Yulan Liu from the University of Sheffield are working on full word-level transcription of selected sequences of the Wargames recordings.

Each participant was asked to fill in a form confirming that their recordings can be used for research. For simplicity I first used the form provided for the AMI/DA meeting recordings. The AMI/DA form was modified for the Wargames recordings and later a new form was used for Settlers recordings. All forms are attached in Appendix A. Please note that participants were paid for their contribution.

4.4 Data preparation

The recorded wav files for the WSJ and MSWSJ corpus had to be pre-processed before being used for speech recognition. This required the following steps:

1. mark beginning and end of each utterance for one microphone array
2. convert utterance labels to contain sentence reference
3. measure offset (and drift) of all the other microphone array recordings
4. generate utterance labels for all microphone arrays
5. split recording into utterances and store separately

The process was automated as much as possible and manual work was carried out on mono audio files which were re-sampled at 16 kHz so as to reduce the file size and therefore enable manual processing. The original wav files are impossible to handle as their size exceeds 1 GB (or even 10 GB for the Eigenmike® data).

Details are available on the 2012_MMA corpus homepage at `http://www.cstr.ed.ac.uk/corpora/2012_MMA/`.

4.5 Baseline results

Baseline results for the WSJ, WSJ_anechoic, MSWSJ and MSWSJ_anechoic data sets are presented in Chapter 7.

Chapter 5

Voice activity detection

Voice activity detection (VAD) is a well established research area and an integral part of different speech communication systems such as teleconferencing, speech recognition or hands-free telephony. Standards for VAD were defined as early as 1989 [Freeman et al., 1989] in the wake of the advent of mobile telephony. Over the last two decades many algorithms and methods for VAD have been presented. These have been tested in various environments using a range of data sets.

VAD, i.e. speech activity detection, was one of four evaluation tasks of the spring NIST RT challenge in 2005[1] where participants had the task of detecting when someone in a meeting was speaking.

VAD performance is therefore one critical parameter in research in speaker diarisation and it is important to look at the most promising algorithms and evaluate them in the context of speaker diarisation in meetings. For this I selected a number of methods and analysed them with respect to their performance on the NIST RT meeting data.

In the following section, five VAD schemes will be presented and their performance measured and analysed. This is followed by the conclusion and discussion of the results and an evaluation of the suitability of these algorithms for online diarisation and handling overlapping speech.

Voice activity detection is very important as it is one of the earliest steps in any speech processing system and errors at this stage propagate directly into downstream processes. A missed speech segment, for example, cannot be recovered. A false alarm

[1] http://www.itl.nist.gov/iad/mig/tests/rt/

segment, i.e. claiming the presence of speech during silence, will also cause problems downstream as an ASR system will generate words for the silence received if its silence model does not recognise it as such. Downstream processes can be designed to handle false alarm segments, most VAD algorithms are therefore tuned for lowest missed speech at an increased false alarm rate.

5.1 VAD algorithms

Most of the VAD algorithms presented in the literature are compared to one or both of two well-known methods: the ITU-T P.56 VAD standard [ITU-T, 2011] and Sohn et al.'s LRT (likelihood-ratio-test) VAD algorithm [Sohn et al., 1999]. It is therefore essential to have these methods available for comparison and for reporting results. In addition, the best performing VAD methods for speaker diarisation in meetings are analysed. This is followed by recently published VAD algorithms and my own modifications and improvements of them.

In summary, the following VAD algorithms have been considered for speaker diarisation in meetings:

- ITU-T P.56 [ITU-T, 2011],
- Sohn [Sohn et al., 1999],
- QIO-FE VAD [Adami et al., 2002b],
- SHoUT [Huijbregts, 2006],
- AZR [Ghaemmaghami et al., 2010] and
- all ones/zeros

5.1.1 All ones (on) and zeros (off)

To date, no results have been presented as to what would happen if the VAD step for speaker diarisation in meetings was left out. For the VAD algorithms the easiest test is to compare the output of the voice activity detector with a reference, where either the complete audio waveform is speech or non-speech, i.e. the VAD output is all ones (always on) or all zeros (always off). This test is also useful as a reference and to

check whether an existing or new VAD algorithm performs better then omitting the VAD process entirely. The all ones and zeros test therefore provides floor results.

5.1.2 P.56

The International Telecommunication Union (ITU) defines an *Objective Measurement of Active Speech Level* standard ITU-T P.56 to ensure that different parties measure speech levels the same way and results can be compared. The P.56 standard defines a software voltmeter that measures the speech signal power, ignoring segments of silence longer than 200 ms. Speech is considered to be present if the current speech activity level exceeds the long-term speech activity level minus 15.9 dB. Freely available C++ code implementing the P.56 standard is part of the ITU-T recommendation G.191. All details are presented in P.56 [ITU-T, 2011] and G.191 [ITU-T, 2010]. The basic building block of the P.56-based VAD scheme is the software voltmeter *sv56*.

According to ITU-T P.56, VAD implementation has a warm-up period, i.e. speech (or increased input signal power) has to be present for some time before non-speech is detected. This warm-up time is defined by the smoothing time of the software voltmeter (smoothing coefficient $\tau_{envelop}$) of 30 ms. Two-stage exponential averaging is carried out on the rectified input signal which outputs the signal envelope. The threshold of 15.9 dB is then applied to determine whether speech is present or not. I added a simple IIR filter to the input of the software voltmeter to remove any DC signal as recommended by the P.56 standard.

The ITU-T P.56 standard also defines a hang-over time of 200 ms, i.e. speech is only considered absent if the signal power level is 15.9 dB less than the long-term activity for 200 ms or longer. Any periods of silence with a length of 200 ms or less are therefore ignored.

5.1.3 Sohn

A second well established benchmark VAD algorithm is the LRT-based VAD scheme presented by Sohn et al. [Sohn and Sung, 1998, Sohn et al., 1999]. Decision rule-based speech detection is composite hypothesis testing, i.e. the output of a VAD detection system is a flag indicating either speech or non-speech. Assuming that the speech is degraded with noise, two hypotheses are possible:

$$H_0 : \text{speech absent} : \mathbf{X} = \mathbf{N}, \tag{5.1}$$

$$H_1 : \text{speech present} : \mathbf{X} = \mathbf{N} + \mathbf{S}, \tag{5.2}$$

where **N** is a noise feature vector and **S** is a speech feature vector.

The principle of decision rule-based LRT is to classify an audio input frame as either speech (incl. noise) or non-speech,

$$\Lambda(k) = \frac{p(X(k)|H_1)}{p(X(k))|H_0} \underset{H_0}{\overset{H_1}{\gtrless}} v, \tag{5.3}$$

where $X(k)$ is the discrete Fourier transform (DFT) of the input signal, H_0 and H_1 are Gaussian distributions of the noise alone and noisy speech respectively, $p(X(k)|H_i)$ is the probability density function (pdf) of the received signal assuming the hypothesis H_i, and v is a threshold which is defined for a specific environment.

Λ, i.e. $\log \Lambda$, then defines whether an incoming audio frame is noise or noisy speech and is calculated from the geometric mean of the likelihood ratios of the individual frequency bins of the DFT, defined as

$$log\Lambda = \frac{1}{L} \sum_{k=0}^{L-1} log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} v, \tag{5.4}$$

where L is the number of DFT frequency bins.

In their work, Sohn et al. [1999] demonstrate that the dependency of Λ on the statistics of the noise alone and noisy speech can be reduced to the noise alone, i.e.

$$\frac{1}{L} \sum_{k=0}^{L-1} \frac{|X_k|^2}{\Lambda_N(k)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 + \alpha, \tag{5.5}$$

where $\Lambda_N(k)$ is the variance of the noise in the individual DFT frequency bins.

The decision statistics defined in Equation 5.5 are therefore an average of L subband SNR ratios, i.e. speech is present if the mean activity (i.e. variance) of all subbands exceeds a pre-defined threshold. This matches the general assumption that noise is static (within a limited time frame) and speech dynamic.

The authors estimate and track the noise statistics using a secondary LRT-based VAD process which keeps the noise model, i.e. the noise power spectra, up to date.

In practice, the VAD output of the proposed scheme needs smoothing. Sohn et al. [1999] use a HMM-based model which tracks the current and previous VAD outputs and delays the transition from H_1 to H_0, therefore reducing missed speech at the cost of an increased false alarm rate.

Sohn et al.'s algorithm is available as a Matlab™ implementation and is part of Voicebox [Brookes, 2011].

5.1.4 QIO-FE VAD

VAD is used in different domains, one of which is front-end processing for telecommunication applications such as mobile phones. The aim is to detect speech and only transmit a signal if a party is talking. These algorithms are optimised for lowest possible missed speech. A set of such algorithms, called QIO-FE, has been designed by a team from Qualcomm, ICSI and OGI and is available online [Adami et al., 2002a]. The QIO-FE contains a VAD algorithm based on a multilayer perceptron (MLP) and a noise reduction Wiener-filter. The QIO-FE program `silence_flags` has been used for VAD.

QIO-FE VAD is implemented using a single hidden-layer feed-forward MLP. The MLP therefore contains three layers. The input layer contains nine frames of six cepstral coefficients computed from low-pass filtered log-energies. These cepstral coefficients are calculated from the speech input from 23 Mel filters. The middle layer of the MLP is hidden. The output layer of the trained MLP contains two elements, $p_n(sil)$ and $p_n(nosil)$, giving an estimate of the posterior probability of the current frame being speech or non-speech. Training of the MLP is done offline using a noisy database and the well known back-propagation algorithm.

The QIO-FE VAD system is shown in Figure 5.1.

Taking the six cepstral coefficients $c_i(n)$ computed from the 23 Mel filters, the inputs $c'_i(n)$ to the MLP are calculated from $c_i(n)$ and the previous values $c'_i(n-1)$ using low

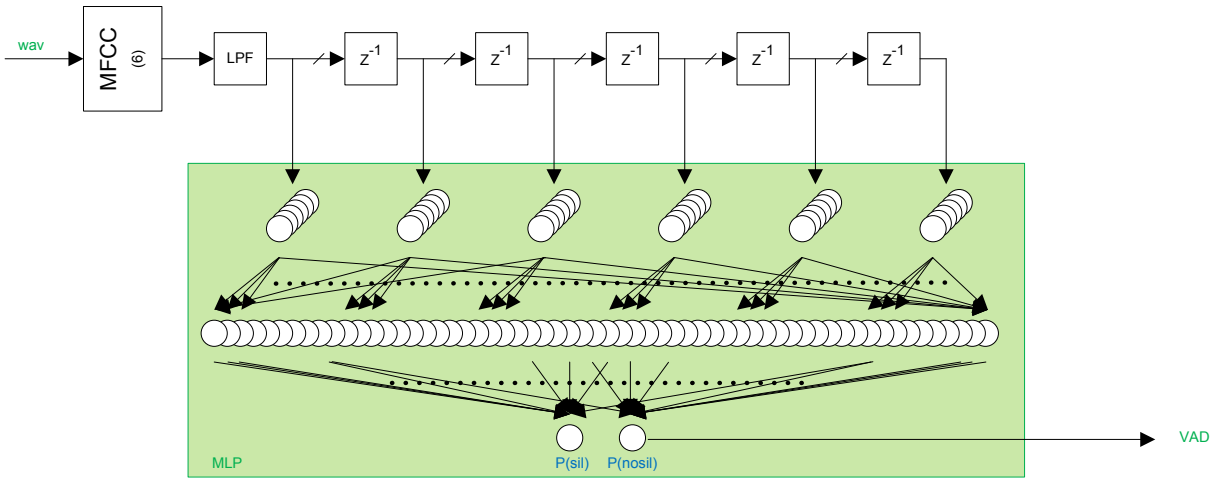


Figure 5.1: VAD using the QIO-FE

pass filtering (LPF) as

$$c_i'(n) = 0.5 * c_i(n) + 0.5 * c_i'(n-1). \tag{5.6}$$

Equation 5.6 represents a first order IIR filter where coefficient a is equal to coefficient b which is equal to 0.5.

The output $p_n(sil)$ of the MLP is now determined as

$$p_n(sil) = \frac{e^{y_n(sil)}}{e^{y_n(sil)} + e^{y_n(nosil)}} \tag{5.7}$$

with

$$y_n(sil) = \sum_{k=0}^{49} w_{k,sil}^2 sigm\left(\sum_{j=0}^{5}\sum_{i=-4}^{4} w_{i,j,k}^1 c_j(n+i)\right) \tag{5.8}$$

and

$$y_n(nosil) = \sum_{k=0}^{49} w_{k,nosil}^2 sigm\left(\sum_{j=0}^{5}\sum_{i=-4}^{4} w_{i,j,k}^1 c_j(n+i)\right). \tag{5.9}$$

$p_n(sil)$ is the probability of a frame being silence and $p_n(nosil)$ the probability of a frame not being silence, i.e. speech or music. $c_j(n)$ is the j^{th} order cepstral coefficient

of frame n, $w^q_{i,j,k}$ is the weight of the first MLP layer associated with the hidden unit k, cepstral coefficient j and frame $n+1$. $w^2_{k,sil}$ and $w^2_{k,nosil}$ are the weights between the hidden unit and the outputs $y_n(sil)$ and $y_n(nosil)$. The function *sigm* is defined as

$$sigm(x) = \frac{1}{1+e^x} \tag{5.10}$$

To obtain the final VAD output a median filter of length 21 is applied. Details are available in Adami et al. [2002a] and the references given therein.

5.1.5 SHoUT

Marijn Huijbregt's SHoUT speech recognition toolkit is another freely available software packet that contains VAD and speaker diarisation programs. The SHoUT toolkit is based on GMMs and HMMs and optimised for the NIST RT challenges. The VAD program `shout_segment` and the diarisation program `shout_cluster` are used for the purpose of this research.

VAD using GMMs and HMMs follows a well established sequence using unsupervised learning techniques and the SHoUT tool is no exception to that. VAD using GMMs and HMMs works on MFCC coefficients provided in a single channel data stream. The incoming audio is therefore usually enhanced using noise reduction and acoustic beamforming techniques to reduce the multi-channel audio stream to a mono channel. After this the speech is split into chunks of variable length, ten minutes in the case of SHoUT. These audio chunks then require pre-processing, i.e. some coarse method to extract speech and silence segments. At this stage it is important to use only the segments with the highest confidence of being pure speech or silence. The SHoUT toolkit uses bootstrapping to determine these, the criteria for finding high confidence segments being segment energy and zero-crossing rate. Each ten minute chunk of audio is therefore split further into equal segments of arbitrary length, typically a few seconds (e.g. 5 s). The energy and zero-crossing rate of each segment is then measured, and segments with high energy and zero-crossing rate are believed to be pure speech, while segments with low energy and zero-crossing rate are considered to be silence.

GMMs are trained using these audio segments and the EM algorithm. This process is repeated iteratively while the number of Gaussian mixtures is increased and the thresholds for speech or silence are relaxed. The iteration is aborted after a fixed

number of repetitions or when a threshold is exceeded and the speech model is trained using all speech segments in one last iteration.

In the second stage the algorithm is repeated for the audible non-speech model, and again in the third stage for the silence model.

Training of the speech, silence and audible non-speech model is completed if the ΔBIC score confirms that the models differ significantly, otherwise the process is repeated from the start with two models only, speech and silence.

Finally, VAD is carried out by a single VAD alignment over the complete audio recording using the Viterbi algorithm. The winning model determines whether a segment is speech, silence or audible non-speech.

To summarise:

- create initial segmentation using a speech/silence acoustic model
- iteratively train three new GMMS (speech, silence and audible non-speech) using the high confidence fragments of the initial segmentation
- check if the speech and audible non-speech models are different using the Bayesian information criterion (ΔBIC); discard all models and retrain only two GMMs (speech and silence) if they are identical
- run a final Viterbi decode to determine the regions of speech, silence and (possibly) audible non-speech

A detailed documentation is available in Huijbregts [2008] and Huijbregts [2006].

5.1.6 AZR

Ghaemmaghami et al. [2010] presented a noise-robust VAD algorithm based on the fusion of two systems. In the proposed method the authors looked at typical characteristics of vowels and their cross-correlation, i.e. they used two measures, the maximum peak of the normalised autocorrelation (MaxPeaks) and the zero-crossing rate of the autocorrelation (CrossCorr) of an audio signal to determine whether the incoming signal is speech or non-speech. The proposed method has two advantages: it runs in the time-domain and it is suitable for online processing.

A detailed analysis of the AZR algorithm showed that the MaxPeaks component of the algorithm is ineffective and that only the CrossCorr algorithm achieves an improved VER compared to the all ones algorithm (see [Zwyssig, 2011] and [Zwyssig et al., 2012a] for details). Henceforth any reference to the AZR algorithm implies that only the CrossCorr part of the algorithm is used.

In the proposed method an input signal $s[i] = s_1[i], s_2[i], ..., s_k[i]$, where i represents the sample number, is segmented into 50 ms frames. To obtain the CrossCorr, the autocorrelation $R_k[z]$ is first calculated as

$$R_k[z] = \frac{\sum_{i=1}^{n-z} x_k[i]x_k[i+z]}{\sum_{i=1}^{n} x_k^2[i]}. \tag{5.11}$$

The authors use only the $x_k[i]$ values for pitch periods of 2 to 20 ms. For Fs = 16 kHz this implies $minF = 50$ Hz and $maxF = 500$ Hz and therefore only values of $x_k[i]$ for i from 32 to 320 are evaluated. The periods of the autocorrelation $\hat{R}_k[z]$ are then extracted, i.e. a list of the distances between every two zero-crossing points P_y is generated. Next the cross-correlation $\hat{R}_y[z']$ of two sets of periods P_y is calculated as

$$\hat{R}_y[z'] = \sum_{j=1}^{n'-z'} P_y[j]P_{y+1}[j+z']. \tag{5.12}$$

Again, this is only done for pitch periods of 2 to 20 ms. Finally, the measure for the periodicity of an incoming audio signal is determined as the maximum of the cross-correlation of two adjacent sets of the periods of the autocorrelation, i.e.

$$C[k] = \sum_{y=1}^{m-1} max(\hat{R}_y[z']). \tag{5.13}$$

Ghaemmaghami et al. [2010] then applied fusion and smoothing to determine whether an incoming signal is speech or not.

The actual implementation requires a few further steps. First, it is essential to remove any DC offset from the audio input. The authors added a de-emphasis filter before the VAD process, i.e. a first order IIR high pass filter defined as

$$y_k[i] = (x_k[i] - \mu_k) - \alpha(x_k[i-1] - \mu_k), \tag{5.14}$$

where x_k is the input signal and i the sample number, μ_k the mean of the input signal x_k, and α the de-emphasis constant, here set to 0.96 (as per Ghaemmaghami et al. [2010]).

Next, the periodicity (CrossCorr) $C[k]$ is measured according to Equations 5.11, 5.12 and 5.13. A voiced speech segment is considered to be present if the CrossCorr value exceeds a certain threshold. I tuned this threshold using the NIST RT06 data and tested it on the RT07 and RT09 data. Finally, it is essential to smooth the CrossCorr value as it is only able to detect voiced sounds. Ghaemmaghami et al. [2010] (and references therein) suggest applying a smoothing window of 200 ms before and 500 ms after detecting a vowel (i.e. CrossCorr exceeds the predefined threshold), i.e. VAD is set to be active for a period of 800 ms if a vowel is detected.

5.2 NIST RT corpus

Many hours of annotated meetings have been made available for the NIST evaluation of meeting diarisation. The data from the RT06 (R106), RT07 (R123) and RT09 (R123) evaluations has been used to test the VAD algorithms presented above. Table 5.1 below lists the specific meeting recordings used.

Table 5.1: Complete list of the NIST RT meetings

RT06-R106	**RT07-R123**	**RT09-R123**
CMU_20050912-0900	CMU_20061115-1030	EDI_20071128-1000
CMU_20050914-0900	CMU_20061115-1530	EDI_20071128-1500
EDI_20050216-1051	EDI_20061113-1500	IDI_20090128-1600
EDI_20050218-0900	EDI_20061114-1500	IDI_20090129-1000
NIST_20051024-0930	NIST_20051104-1515	NIST_20080201-1405
NIST_20051102-1323	NIST_20060216-1347	NIST_20080227-1501
TNO_20041103-1130	VT_20050408-1500	NIST_20080307-0955
VT_20050623-1400	VT_20050425-1000	
VT_20051027-1400		

Please note that only EDI[2], TNO[3] and IDI[4] meetings were recorded with a circular eight-channel microphone array of 20 cm diameter. All other meetings were recorded

[2]CSTR - Centre for Speech Technology Research - The University of Edinburgh

[3]Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek - Dutch Organization for Applied Scientific Research

[4]Institute Dalle Molle d'Intelligence Artificielle Perceptive - Dalle Molle Institute for Perceptual Artificial Intelligence

with an ad-hoc setup of two to seven microphones.

5.3 Evaluation of VAD algorithms

In the experiments presented here, Wiener-filter-based noise reduction was first applied to the individual microphone signals. Delay-sum beamforming was then performed on the signals after which VAD was carried out. Scoring was performed using the NIST scoring tools.

Two different noise reduction and beamforming tools were used. This allowed running the VAD experiments with two completely different pre-processing methods for improved verification of the algorithms used. Noise reduction was carried out using the QIO-FE `nr` tool and the in-house mdm tools. The mdm tools also provide delay-sum and superdirective beamforming. In these experiments, delay-sum beamforming was used in order to compare the results with the BeamformIt delay-sum beamformer which was combined with the noise reduction of the QIO-FE, as shown in Figure 5.2.

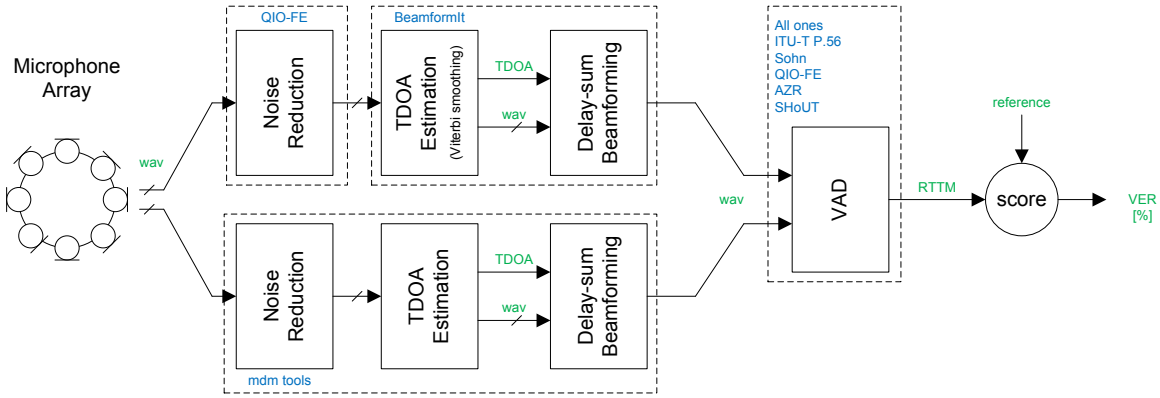


Figure 5.2: Flow diagram for verification of VAD algorithms

Tables 5.2 and 5.3 show the voice activity detection error rate (VER) for each of the algorithms when tested on the complete NIST RT06, RT07 and RT09 data sets. These tables show mean and standard deviation (SD) for the VER. Note that the % VER figures for the mean and SD are scaled in proportion to the individual meeting lengths.

On the RT06 test set, perhaps surprisingly, classifying all segments as speech outperforms all the other algorithms with 6.8% VER. This implies that, for this particular set of meetings, there are very few non-speech intervals leading to few false alarm errors for the all ones algorithm. For RT07 and RT09, which contain more non-speech

Table 5.2: Voice activity detection error rate VER [% mean (SD)] for all algorithms using the BeamformIt and QIO-FE nr tools on the NIST RT meeting data

VER [%]	All ones	ITU-T P.56	Sohn	QIO-FE VAD	AZR	SHoUT
RT06	6.8 (4.9)	14.3 (8.3)	16.7 (3.0)	9.0 (3.2)	15.6 (12.5)	15.6 (2.6)
RT07	13.7 (3.6)	12.2 (6.1)	11.9 (1.0)	4.7 (1.3)	13.5 (11.3)	4.1 (1.5)
RT09	11.3 (9.7)	10.3 (3.1)	10.8 (2.1)	3.4 (1.5)	7.7 (2.4)	7.1 (3.9)

Table 5.3: Voice activity detection error rate VER [% mean (SD)] for all algorithms using the mdm tools on the NIST RT meeting data

VER [%]	All ones	ITU-T P.56	Sohn	QIO-FE VAD	AZR	SHoUT
RT06	6.8 (4.9)	27.5 (8.9)	14.6 (3.5)	10.3 (3.4)	12.3 (13.6)	15.5 (2.5)
RT07	13.7 (3.6)	14.9 (8.4)	20.5 (26.4)	4.4 (1.2)	12.5 (9.2)	4.8 (2.0)
RT09	11.3 (9.7)	10.8 (3.8)	10.0 (2.2)	3.2 (3.2)	6.3 (4.4)	7.8 (4.0)

segments, the all ones, ITU, Sohn and AZR algorithms have similar results, and are consistently outperformed by QIO-FE VAD and SHoUT, with QIO-FE VAD having the lowest overall error when averaged over all three test sets.

The VAD results of the RT06, RT07 and RT09 meetings show a great variability of the error for the different meetings. This is something that has been noted and documented before (see e.g. Vijayasenan [2010]). Figure 5.3 shows the variability of the VER on the RT corpora with the QIO-FE and SHoUT VAD algorithms using the QIO-FE nr and BeamformIt tools.

I carried out a detailed analysis of all meetings looking at overlap (two or more active speakers) and minimal, average and maximum speech segment length of the individual meetings compared to the VAD error rate. Most RT06 meeting recordings contain a very high percentage of overlap (15-35%) compared to the RT07 and RT09 meeting recordings (4–15%), but each meeting set contains at least one outlier.

The VER displays a similar behaviour, i.e. a high variance in VER, but no correlation could be observed between % overlap or minimal/average/maximal speech segment and the VER[5].

Minimum and average speech length are almost double for the RT06 meeting re-

[5]Percent overlapping speech and maximum speech segment length for the RT06 data are twice as high as as for the RT07 and RT09 data. Considering that all data are 'normal' meetings this indicates problems with the transcription, i.e. reference. More weight is therefore applied to results from the RT07 and RT09 data.

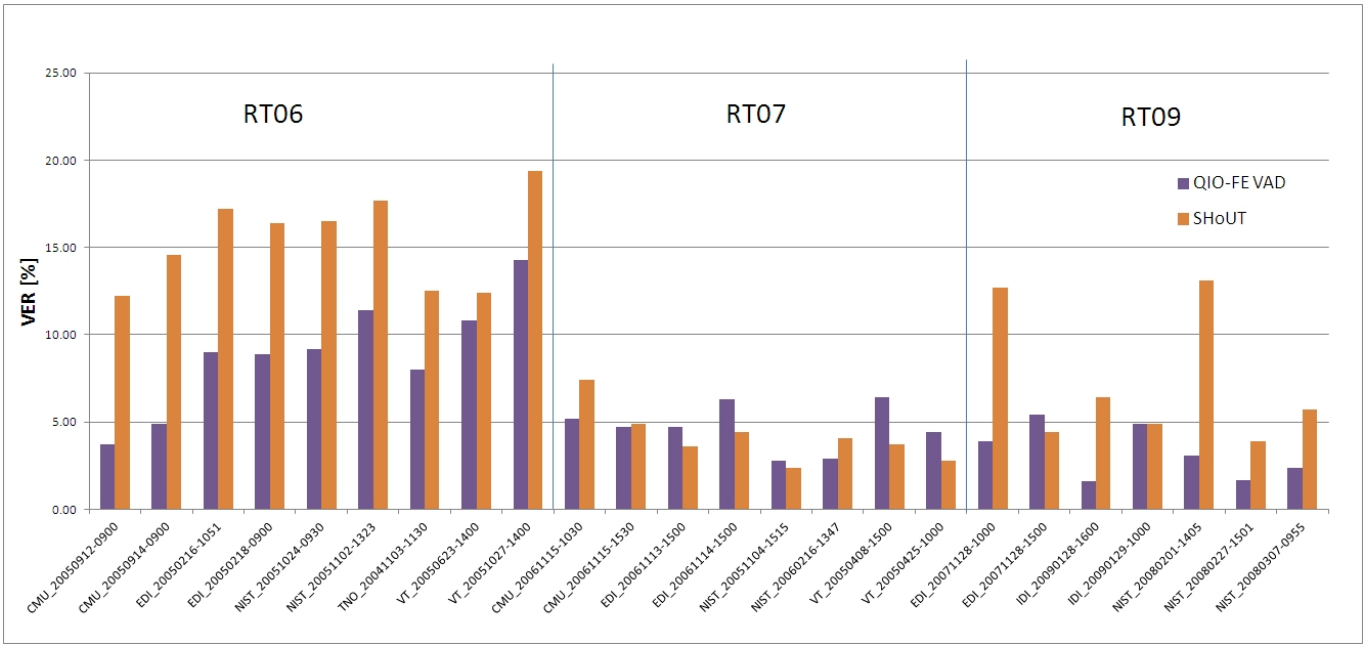


Figure 5.3: Detailed VAD results (QIO-FE VAD) on the NIST RT meeting data

cordings (0.2–0.4s/3–4s) compared to RT07 and RT09 (0.05–01s/1.5–2s), perhaps indicating that the VAD algorithms tested are better suited for shorter speech segments. Again, for the RT09 test set, the SHoUT VAD algorithm shows two large outliers, the EDI_20071128-1500 and NIST_20080201-1405 meetings. Looking at the overlap (2/3+ speakers) of these two meetings (EDI_20071128-1500: 6.5%/0.3%; NIST_20080201-1405: 31.6%/7%) no conclusion can be drawn as to the correlation of the meeting statistics and the performance of a particular VAD scheme.

It is therefore interesting to see how the VAD algorithms deviate over the individual test sets and meetings. To give a better picture of the nature of the VER over the different data sets the box plot results for the algorithms tested are shown in Figures 5.4, 5.5 and 5.6. The median is shown as the interface between the brown and red boxes, the lower quartile as a brown box, the upper quartile as a blue box and the minimum and maximum VER as whiskers.

The median, lower and upper quartile and minimum and maximum VER of the different algorithms and data sets shows that the QIO-FE and SHoUT VAD algorithms have the least variability over all meetings and are best suited for VAD in meetings. I have therefore decided to use the QIO-FE VAD tool for my research and further experiments.

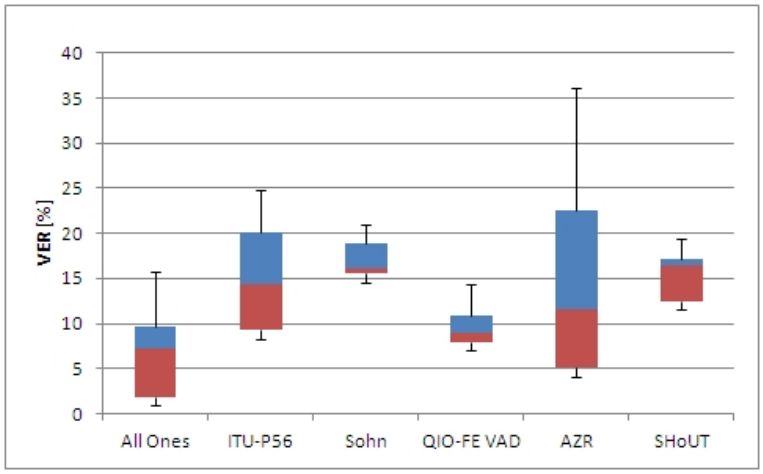


Figure 5.4: Detailed VAD results (NIST RT06 meeting data)

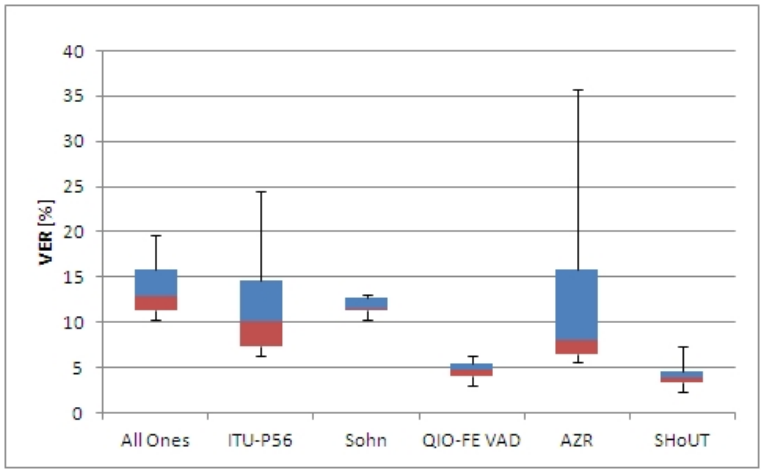


Figure 5.5: Detailed VAD results (NIST RT07 meeting data)

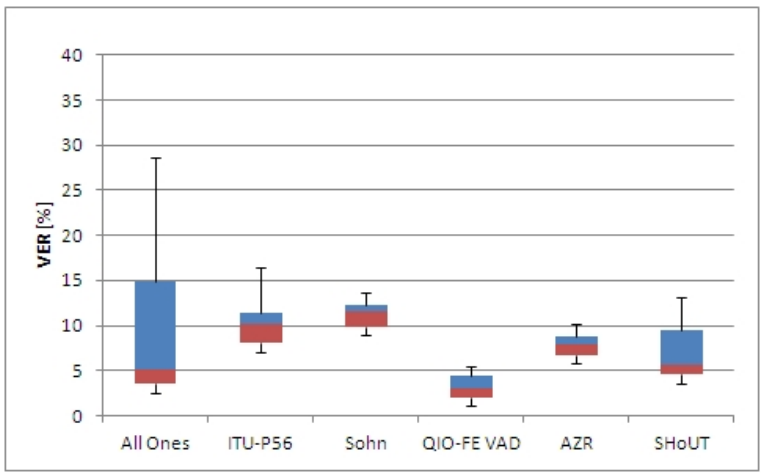


Figure 5.6: Detailed VAD results (NIST RT09 meeting data)

5.4 Re-training the QIO-FE VAD MLP

Little information is available as to the nature and quantity of the data used for training the MLP weights of the QIO-FE VAD algorithm. Adami et al. [2002a] only state:

> "... training is done offline using a noisy database."

The only other hint as to the data used for training the MLP can be found in the toolset documentation, i.e. the README file for the VAD parameters which the authors make available for users of the VAD tool. A comment there reads:

> "... SpeechDatCar-Italian data was omitted from the VAD training set due to its license terms ..."

I can therefore only assume that the SpeechDatCar corpus was (in part) used for training the MLP weights.

The SpeechDatCar corpus contains speech from multiple languages (US English, German, Spanish, Italian, etc.) recorded in cars using four microphones: one close-talking microphone (used as the reference) and three distant microphones at fixed positions in the car, all sampling at 16 kHz. In addition, a GSM speech signal sent from the car was recorded at 8 kHz. This configuration applies to all the databases.

At least 300 participants (per language set), both female and male, read instructions and commands from a prompt. Each speaker uttered the type of commands typically found for controlling (mobile) devices in a car, such as voice activity keywords, isolated and connected digits, dates, people and place names or phrases containing embedded keywords. Annotation of the speech and silence regions was carried out using human transcribers. Details can be found at `http://www.elda.org`.

SpeechDatCar is a highly variable corpus with many hours of speech data from different people, languages, environments and noise conditions, but principally out-of-domain with respect to meeting data.

It is therefore of interest to determine how the QIO-FE VAD algorithm performs if it is trained on in-domain data, especially because MLP-based VAD has been shown to perform well on speech from meeting data recorded using individual headset microphones [Dines et al., 2006].

Please note that I carried out re-training of the MLP to check whether the QIO-FE VAD algorithm could be improved if the MLP weights were trained on in-domain data.

Modification of the QIO-FE toolkit parameters and flow is very limited, and working in depth with MLP is beyond the scope of my research.

For modification of the QIO-FE VAD algorithm I decided to use the audio data and reference transcriptions from the AMI and AMIDA meeting corpus [Renals, 2010] to train the MLP weights and to test them again on the RT data. The AMI and AMIDA meetings have been recorded with either two or three microphone arrays. The first array is placed in the middle of the four participants of the meeting (I use only scenario meetings), the second in front of the whiteboard and the third at the remote location (of the fourth meeting participant of the AMIDA meetings).

Only the first array is used for the purpose of this research. The eight-channel audio data from the meetings needs pre-processing, i.e. noise reduction (QIO-FE tools) and acoustic beamforming (BeamformIt) are used to enhance the audio and reduce it to a mono channel. Next, MFCC feature vectors are generated and the MLP is trained using the ICSI Quicknet toolset [The International Computer Science Institute, 2010].

Training an MLP requires setting a parameter to specify how much of the training data is to be used for training and how much for cross-validation. I verified the best ratio in an initial experiment and achieved the best VER using approximately 10% for cross validation, a number which was confirmed by one of the developers of the QIO-FE (Stéphane Dupont, personal communication, 13 February 2013)

The final VER performance of the QIO-FE VAD MLP, trained on four AMI meetings (three for training, one for cross validation) and 32 AMI meetings (28 for training, four for cross validation) and tested on the NIST RT data is presented in Figure 5.7. ‘QIO FE VAD’ shows the results using the original weights, ‘3-1’ shows the VER results for re-training using four AMI meetings and ‘28-4’ for re-training using 32 AMI meetings.

The results in Figure 5.7 show that using the QIO-FE-trained MLP weights achieves the best overall VER on the NIST RT data. It also shows that increasing the training data from four meetings (approximately two hours of training data) to 32 meetings (approximately 16 hours of training data) actually does result in a degradation of the VAD performance. This would indicate that the increased amount of training data (AMI/DA meetings) leads to a greater mismatch on the test data (NIST RT meetings), despite the fact that nine of the 24 meetings from the test data are very similar to the training data. Surprisingly, the VER for EDI, IDI and TNO meetings in the NIST RT data also increases when VAD is carried out using the weights trained on the AMI data.

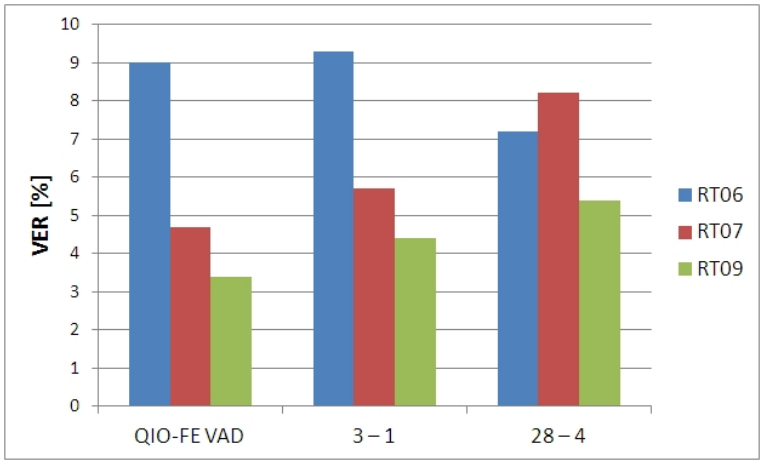


Figure 5.7: [%] VER for the QIO-FE VAD (retrained on AMI meeting data) on the NIST RT meeting data

I conclude that it is more important to train the MLP on varied test data which incorporates many different speakers and environments than a limited amount of in-domain data. This result also reinforces the decision to use the QIO-FE VAD with its own MLP weights for best VAD in meetings.

Please note that the great potential of MLP-based VAD has also recently been demonstrated by Ng et al. [2012] when developing a speech activity detection system for the DARPA RATS program. MLP-based VAD outperformed GMM-based VAD on the RATS (Radio Traffic Collection System) data comprising five different languages transmitted over eight different radio channels. The authors evaluated the MLP-based and GMM-based systems individually and also implemented a combined system with which they achieved 5% relative improvement over the MLP-based VAD system and 9% over the GMM-based system. The best EER (equal error rate) achieved on the RATS data which contains mostly CTS was 1.42%, i.e. 1.42% false alarms and 1.42% missed speech.

5.5 Summary and conclusions

This chapter reviewed and analysed multiple well-known and commonly used VAD algorithms for speaker diarisation and speech recognition. These were the ITU-T P.56 standard based on the energy threshold, the LRT scheme based on frequency bands proposed by Sohn, the MLP-based method developed for the QIO-FE, the GMM-HMM-based algorithm provided by the SHoUT toolkit and my own modification of the periodicity-based AZR method.

When comparing these VAD schemes with each other and with no VAD I found that, when designing a VAD algorithm, great care is required in order for it to perform well, especially in meetings.

Note that some of the VAD methods presented in this chapter were developed for telephone speech and automotive applications and not for distant speech (which typically contains a reverberation tail at the end of a vowel).

The results presented show that VAD based on GMM-HMMs and MLPs performs considerably better than methods based on speech activity or speech periodicity. Measured on the RT07 and RT09 data, I found that overall MLP performs best both in terms of the lowest mean and lowest variance.

Retraining the MLPs using in-domain meeting data did not achieve an improvement in VER.

Chapter 6

Determining the number of speakers in a meeting

Intuitively I would expect that the performance of speaker diarisation, that is, finding 'who spoke when' in a multiparty conversation, relies heavily on correctly determining the number of speakers, a parameter which is not known a priori. Surprisingly though, most state-of-the-art diarisation systems show little interest in finding this parameter, despite its potential use as a stopping criterion for the speaker clustering step.

This chapter starts by looking at prior research on determining the number of active speakers for diarisation and the reasons why there has so far been limited success in finding this number. This is followed by the presentation of a novel algorithm that is able to determine the number of active speakers in a meeting (or any other multiparty conversation) recording using a microphone array of known geometry. The new algorithm was verified on the NIST RT AMI meeting data and on the 2012_MMA corpus, specifically on the single and dual speaker tasks. The chapter concludes by presenting the results that were achieved followed by an analysis and discussion.

6.1 Prior work

In experiments, Meignier et al. [2006] found that automatically estimating the number of speakers during the clustering process generates a 4% increased absolute diarisation error than clustering with the optimal number of speakers. The authors used a diarisation engine that is based on HMMs and GMMs, and segmentation and clustering was

carried out using the Bayesian information criterion (BIC). It is therefore compatible with the ICSI, SHoUT and LIUM diarisation systems and achieves comparable results. The experiments show that the minimal DER is not achieved with the true number of active speakers but some other (higher) number, which is sub-optimal for downstream processes. Diarisation systems are optimised for lowest error rate which is achieved by clustering to a higher number of speakers than are really present, a result confirmed in the experiments by Sinclair and King [2013].

Sinclair and King showed that segmentation and clustering based on GMM-HMMs (as used by most current state-of-the-art diarisation systems) suffers from bad models as a result of training on impure data. Their research shows that conversational speech labelled with the correct number or too few speakers results in a sudden increase in DER.

Huijbregts et al. [2012] performed Oracle experiments on the SHoUT diarisation system. In their experiments the authors substituted single components of the system with Oracle components, i.e. the ground truth. Experiment 4 presented in Huijbregts et al. [2012] comes closest to providing the diarisation system with the correct number of speakers as the merge stopping criterion, but

> "fixing the number of speakers to the reference number is not ideal, because if the system makes a merging mistake, trying to cluster up to the reference number of speakers will not give you the best DER (Marijn Huijbregts, personal communication, 23 November 2011)".

The designers of the SHoUT and ICSI diarisation systems found that using the true number of speakers as the stopping criterion for the segmentation and clustering processes does not lead to the best DER. As a consequence of this, neither of these toolkits supports defining the final number of clusters.

Determining the number of speakers in a multi-party conversation for improved diarisation is not a new problem. Early experiments carried out by Ben et al. [2004] found that the task is very difficult (if not impossible) to solve using acoustic features alone. The acoustic signal does not inherently contain any location information, but the direction of arrival (DOA) of that signal and the time difference of arrival (TDOA) linked to this DOA do.

As discussed in Section 2.4, incorporating TDOA features into the diarisation task – that is, combining the audio (MFCC) and localisation (TDOA) features – has been explored by several researchers and can lead to significant DER improvements. Most

algorithms integrate the TDOA values as a parallel feature stream and do not attempt to determine the number of active speakers per se. Sinclair and King [2013] demonstrated that allowing more speakers, i.e. clusters, for the diarisation tasks leads to improved results as this keeps the main speaker models more pure and allows for segments with less confidence to be assigned to impure models. In addition, the last critical merge does not need to be carried out and can therefore not go wrong, thus avoiding model impurity and a significantly increased error rate.

For diarisation systems such as developed by Pardo et al. [2012], Friedland et al. [2012], Huijbregts et al. [2012] etc., determining the number of speakers N_{spkr} would only be of interest if the clustering stopping criterion could be set to N_{spkr} plus some arbitrary offset, where the offset would need to be determined by experiments.

At present, only two research groups appear to have tackled the problem and have developed algorithms that determine the number of active speakers in meetings. Nwe et al. [2012] developed an offline diarisation system that analyses the histogram of the TDOA values (generated by a DSB beamformer and smoothed using Viterbi alignment) of an entire recording, discards bad histograms that contain few different peaks (i.e. speakers) and clusters the quantised good histograms, where the number of clusters corresponds to the number of speakers. For N microphones and $(N-1)!$ microphone pairs, this method generates $(N-1)!$ possible histograms and therefore $(N-1)!$ hypotheses for the number of speakers present. The authors then used cluster fusion to determine the correct number of speakers after which a sophisticated cluster merging and purification scheme was employed to carry out the diarisation, producing the best known results on the RT09 data and a good estimate of the number of active speakers.

Nwe et al. [2012] state that "if speakers do not move during a meeting and audio quality is good then the number of significant peaks in the TDOA histogram represents the number of speakers and the TDOA features are very informative". However, they also claim that if speakers move to different places such as they believe to be the case in the RT07 and RT09 recordings, then the performance of the suggested scheme decreases "because the segments of an individual speaker are distributed into two or more clusters".

I analysed three of the four meeting videos which were claimed to contain moving speakers and can confirm that this is not the case, i.e. speakers do not change their position. My own research points more to the fact that TDOA values are prone to

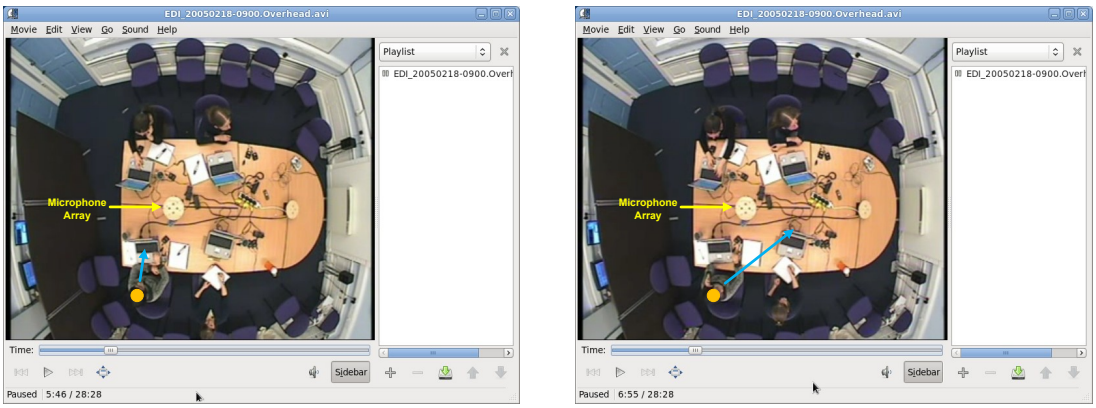
error in noisy environments and that the natural movement of a speaker's head during talking and reverberation could well result in histogram peaks at positions where no real speaker is present.

Figures 6.1, 6.2 and 6.3 show exactly this scenario. Figure 6.1 shows two screenshots of the NIST RT06 meeting 'EDI_20050218-0900' at times 5:46 (a) and 6:55 (b). The active speaker and the direction of her speech are indicated by blue arrows, the position of the microphone array is indicated by a yellow arrow. Figures 6.2 and 6.3 show the direction of the arriving sound derived from the TDOAs of the array. On the x-axis are 36 sectors on the circle, the y-axis shows the time in seconds and the z-axis the activity per sector. Figure 6.3 shows a zoomed version of Figure 6.2 for the time 300–350 s (a) and 700–900 s (b). The four speakers in the meeting sit at the positions 3, 8, 20 and 24. Figure 6.1 shows three distinct sharp peaks for speakers 3, 8 and 20, and a much wider peak for speaker 24. When analysing the video of the meeting recording, I found that speaker 24 moved her head between two distinct positions, either addressing the other meeting participants (Figure 6.1 (a)) or talking in the direction of the whiteboard (Figure 6.1 (b)). The width of the active sectors indicate that turning her head by 45° while talking results in TDOA changes which give the impression that the speaker moved by 30° around the circular array. Details of how to determine the active sector of a speaker from the TDOAs will be presented in Section 6.2.

An alternative to using TDOAs as a sound source location is to use DOA from power-steered beamforming. Ishiguro et al. [2012] took DOAs from ICA (independent component analysis) algorithms and filtered the result using the bag-of-words (BoW) model to determine the number of (active) speakers. ΔBIC-based segmentation and clustering then leads to improved diarisation results as the cluster stopping criterion is known.

The system implemented by Ishiguro et al. [2012] works online, achieving approximately 34% DER on a subset of the AMI meetings. No results are presented as to the number of speakers detected compared to the true number.

The next section presents a novel method for speaker diarisation which explicitly calculates the number of speakers by estimating their location using a microphone array.



(a)Talking to the group (b) Talking to the whiteboard

Figure 6.1: Analysis of TDOA during a meeting with the speaker (orange dot) turning her head while talking, blue indicates the direction of speech (EDI_20050218-0900)

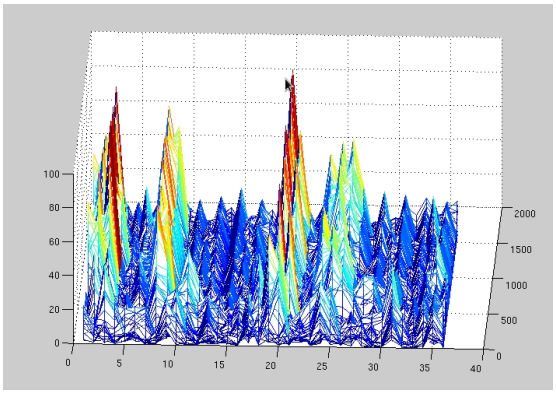
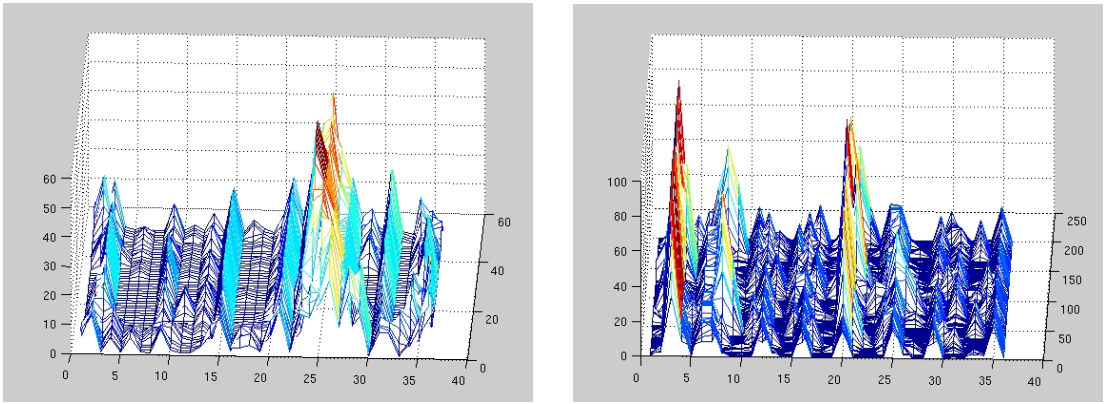


Figure 6.2: Histogram of angle of arrival of sound on microphone array for a complete meeting (EDI_20050218-0900)



(a) Speaker 2 (b) Speaker 1, 3 and 4

Figure 6.3: Detailed histogram of angle of arrival of sound on microphone array for a complete meeting (EDI_20050218-0900)

6.2 TDOA analysis

The TDOA is the time difference of the direct wavefront from a given sound source arriving at two different microphones. An established method for TDOA estimation is generalised cross correlation with phase transform (GCC-PHAT, cf. Section 2.1.3) and the estimates produced may be further improved by Viterbi smoothing [Anguera et al., 2007]. If the relative location of the microphones is known, given the TDOA values for a pair of microphones, simple geometry may then be used to calculate the angle of arrival of the signal in relation to the microphones. In fact, due to rotational symmetry, for two microphones a single delay estimate results in **two** possible angles of arrival – the correct one and another reflected on the axis of the two microphones.

If a recording is done with an eight-channel microphone array, then seven microphone pairs can be analysed in relation to one reference channel. The reference channel can be any of the eight microphone channels and is usually chosen once and fixed at a specific position. One option for choosing the reference channel is to select the input with the highest signal level, therefore assuming that this microphone is closest to the current active speaker. Another possibility is to analyse the cross-correlation of every microphone pair (resulting in $(M-1)!$ possible pairs) and to select the microphone with the highest overall cross-correlation score. Both schemes are sub-optimal, as the first requires changing the reference if the active speaker changes or it results in running with a non-optimal reference channel if the active speaker is not in front of the selected reference microphone. The second scheme requires large amounts of computation and is also not suitable for online processing as the complete audio needs to be available to determine the reference microphone.

As stated above (cf. Section 2.5.3), BeamformIt determines the reference channel by calculating the cross correlation of every microphone pair for the entire recording and then selects the one with the highest value. The mdm tools simply choose the microphone with the highest averaged energy. Both tools select the reference channel once and keep it fixed for the entire recording. The reference channel is the channel with TDOA = 0.

In either case, if one of the M microphones is chosen as the reference microphone, then the TDOA values of *M-1* microphone pairs can be calculated. The GCC-PHAT algorithm is most widely used for this. The TDOA values can only take discrete values and are subject to spatial and frequency aliasing. Given the speed of sound c and the

difference D_{ij} of a microphone pair, two possible angles of arrival can be calculated for a given TDOA value. These angles then need to be aligned to a reference direction, after which they can be evaluated and the most likely direction of arrival (DOA) determined. The following list summarises the proposed procedure:

1. determine reference channel
2. look up pair of angles from TDOA value
3. align each pair of angles with the reference direction
4. create a sector activity map with entries for each angle
5. evaluate this sector activity map to determine the most likely sound source location
6. write sector activity table

First, the reference channel is determined, looking at either a single set of TDOA values or the entire TDOA data. The reference channel is the channel with a TDOA value of 0. Then, given a circular microphone array with a diameter of $D = 20$ cm and eight uniformly distributed microphones around this circle, the distance between each microphone pair D_{ij} can be calculated. After this, given the $TDOA(n)$ of an incoming signal $x(n)$ (sampled at the frequency Fs), the angle of arrival φ for a specific TDOA value can be calculated as

$$\varphi = \arccos\left(\frac{TDOA(n)}{N}\right), \tag{6.1}$$

where N is the distance (in audio samples) of the microphone pair i and j between which the $TDOA(n)$ was measured, i.e.

$$N = \frac{D_{ij}Fs}{c}, \tag{6.2}$$

where c is the speed of sound in air (set to $343 m/s$ for this work).

An example result for TDOA = (-3, -2, -1, 0, 1, 2, 3) for the microphone pair (1-2, 1-3, ..., 1-8) is (2.57, 2.17, 1.85, 1.57, 1.29, 0.98, 0.57) rad or (147, 124, 106, 90, 74, 56, 33) degrees. It can now be seen that α runs from 0 to π with $t = -3$ being close to π and $t = +3$ being close to 0. This means that a signal from the front will produce a

maximum positive TDOA value and a signal from the back a maximum negative one, while a signal from the side does not incur a time difference ($TDOA = 0$).

It is very important to note that α runs from 0 to π and not to 2π. This means that it is not possible to tell whether a signal at a microphone pair arrives from the left or right, as is illustrated in Figures 6.4 and 6.5. Further, note that the precision of the angle derived from the TDOA is greatest for signals coming from the side, while it is least for signals coming straight from the front or back.

Extrapolating from a microphone pair to a microphone array with eight microphones, seven pairs of angles can be derived from seven different TDOA values, as shown in Figure 6.6.

In Figure 6.6 the direction of sound is indicated by a blue arrow. Microphone 2 has been selected as the reference microphone and the seven possible microphone pairs are (2-3, 3-4, 2-5, 2-6, 2-7, 2-8, 2-1). Given a small positive TDOA value for the microphone pair 2-7, two possible angles φ_1 and φ_2 can be determined. Different TDOA values for the other pairs will result in a typical scenario as shown in Figure 6.6.

6.3 Sector activity (SA)

The task at hand now is to find the true direction of arrival (DOA). I have explored five different filtering methods to find the correct DOA. These are *all*, *optimised*, *wtia* (winner-takes-it-all), *Gaussian* and *mixed*.

For the *all* algorithm, all 14 angles are entered into the SA table and no filtering is applied. For *optimised*, the sum of all the squared distances of each DOA (wrapped on the full circle) is calculated and, of the two possible DOAs, the one with the larger deviation to the minimal sum is discarded. For *wtia*, all but one DOA – that with the smallest deviation from *optimised* – are discarded. For *Gaussian*, the circular mean and variance of every DOA is calculated and the table entries are re-calculated based on the smallest variance and the corresponding mean. For *mixed*[1], only the DOA closest to the circular mean is entered into the table.

The directivity pattern of an MVDR superdirective beamformer [Bitzer and Simmer,

[1] I first implemented my own function and later found an identical function called *circ_r* as part of CircStat, a Matlab™ toolbox for circular statistics [Berens, 2009]. The results for the last two are, as expected, identical.

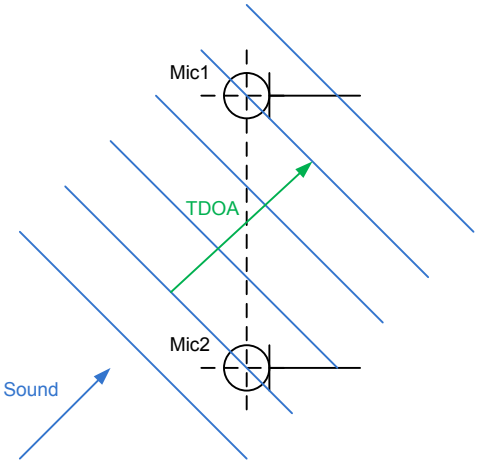


Figure 6.4: TDOA for stereo microphone with sound coming from the left

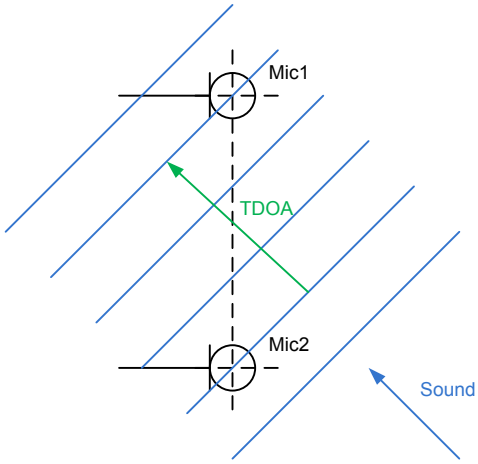


Figure 6.5: TDOA for stereo microphone with sound coming from the right

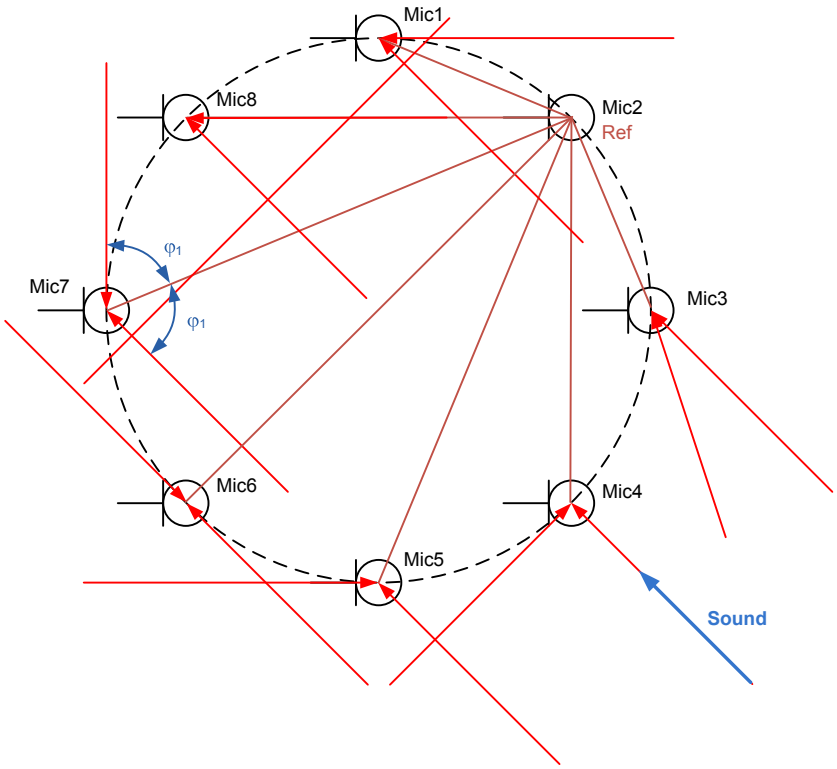


Figure 6.6: Angle of arrival calculation from the TDOAs for an eight-channel microphone array: the direction of sound is indicated by a blue arrow, microphone 2 has been selected as the reference microphone and the seven possible microphone pairs are (2-3, 2-4, 2-5, 2-6, 2-7, 2-8, 2-1), red arrows show the 14 possible angles of arrival.

2001] for an 8-element microphone array with a diameter of 20 cm and a sample rate of 16 kHz (the conditions used in the NIST recordings) shows a main lobe width of 10°. In order to identify the angle of the speakers in relation to the array, I therefore create a sector activity map of $N = 36$ possible sectors, one every 10°. The TDOA values for each microphone pair are estimated every 256 ms and the angle of arrival values calculated. A counter value in the sector corresponding to that angle is then incremented. I accumulate counts in five second windows with one second overlap and record the highest scoring sector for each window, thus calculating the sector with the most activity (the active sector) over five seconds for every second of recording.

The TDOA value outputs from the beamformer is only processed when speech is detected, i.e. when the VAD output is active. A speaker entry is generated for 36 possible SA sectors (one every 10°), i.e. one speaker is entered into the diarisation output for the highest scoring sector on the circle at that time. If the highest scoring sector changes during a single active VAD segment then this segment is split and two entries are made to the diarisation output, one for the first sector (or speaker) and another for the second. A typical SA map is shown in Figure 6.7.

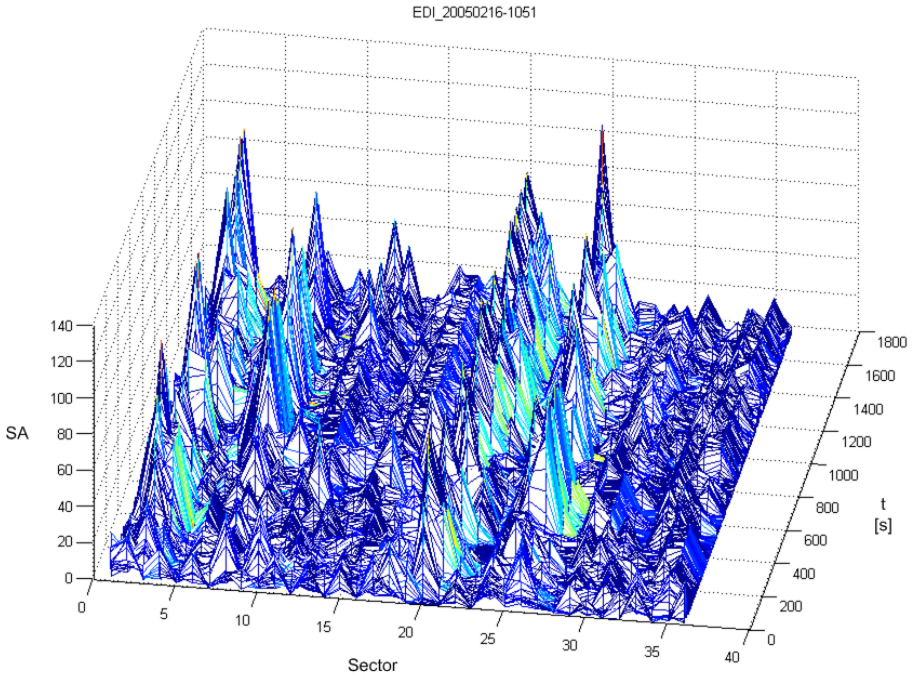


Figure 6.7: Sector activity (SA) map from NIST RT06 meeting EDI_20050216-1051

The proposed methods is called VAD+SA and the diarisation error rate for the different DOA filtering schemes is presented in Section 6.3.1 below.

6.3.1 Preliminary results

A subset of the NIST RT meetings (those recorded at the University of Edinburgh, IDIAP and TNO – see Table 5.1 for a full list) was recorded using an eight-element circular microphone array of 20 cm diameter. These meetings are the only ones in the NIST RT data set for which the relative locations of the microphones are known and will henceforth be referred to as the "NIST RT AMI meetings".

All diarisation experiments were carried out similarly to the VAD experiments presented in Section 5.3. Again, two different pre-processing methods for improved verification of the algorithms were used, i.e. two different noise reduction tools (QIO-FE and the mdm tools) and two acoustic beamformers (BeamformIt and the mdm tools), as shown in Figure 6.8.

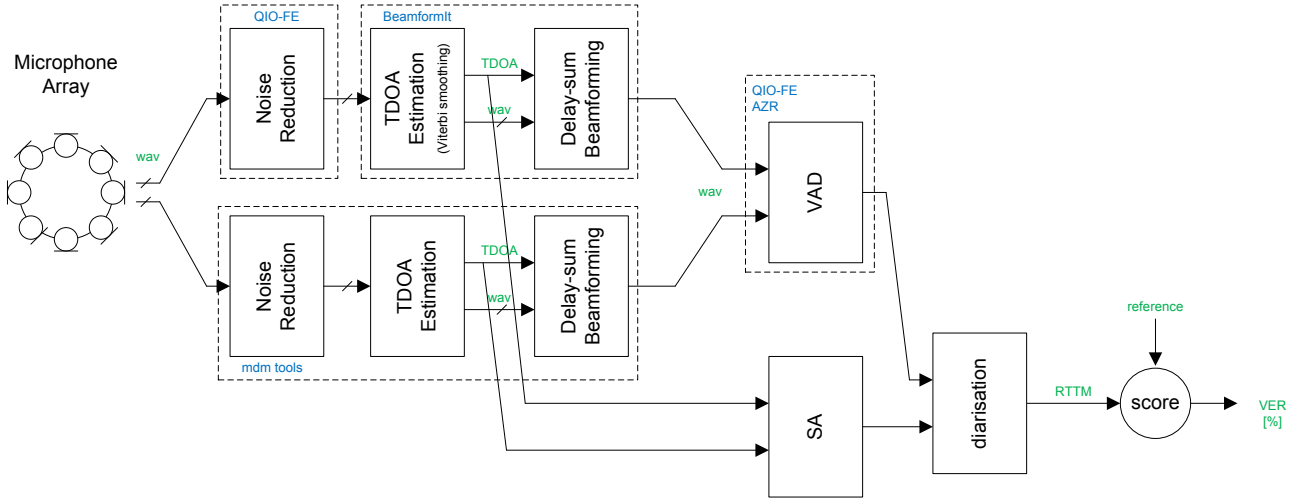


Figure 6.8: Flow diagram for the evaluation of the algorithm to determine the number of speakers in a meeting

In the first experiment, the six methods for filtering the TDOA values from the beamforming on the circular array were evaluated in relation to the diarisation error rate. Results are presented using two different pre-processing systems as well as two different VAD schemes. These are AZR and the best-performing VAD algorithm on the NIST data, the QIO-FE VAD (see Chapter 5 for details).

Two different beamformers and two VAD schemes are analysed in order to confirm the results from previous experiments on diarisation (cf. Section 3.2.2 on page 76) and VAD (cf. 5.3 on page 107).

Figure 6.9 shows the DER for the six DOA smoothing methods *all*, *optimised*, *wtia*, *Gaussian* and *mixed*. Interestingly, the *all* scheme works best and any attempt to 'improve' the TDOAs worsens the DER.

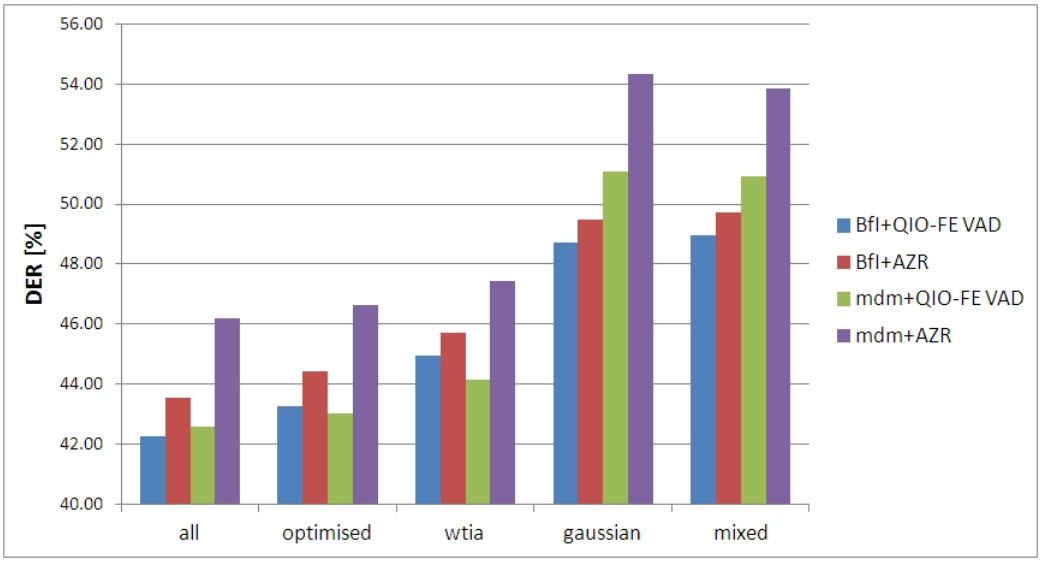


Figure 6.9: Sector activity map analysis for the NIST RT AMI meetings

Please note that the best performing VAD and acoustic beamforming tools remain the best methods to generate the sector activity map.

Next, the results from experiments using speaker clustering are presented in Figure 6.10. The speech segments found during sector activity detection with a maximum of 36 speakers are clustered using the ΔBIC criterion, i.e. adjacent speech segments are analysed as to whether they belong to the same speaker and merged if they do. For this, if $\Delta BIC > 0$ (as per Equation 2.41) then the speaker ID of the new speech segment is replaced with the speaker ID of the previous speech segment.

The clustering is carried out over a look-back window of the previous eight speech segments. With perfect TDOA estimation, the proposed method should be able to determine whether the two speech segments which are compared are from the same speaker or whether a speaker change has occurred. A look-back window of eight segments (compared to the theoretically necessary two) is selected to allow for the very short speech segment length in a typical conversation. In these experiments I tried to determine the best ΔBIC penalty λ, as per Equation 2.41.

Figure 6.10 shows the DER after speech clustering on all EDI, TNO and IDI meetings for the RT06, RT07 and RT09 evaluations. The first column (VAD+SA) shows the best

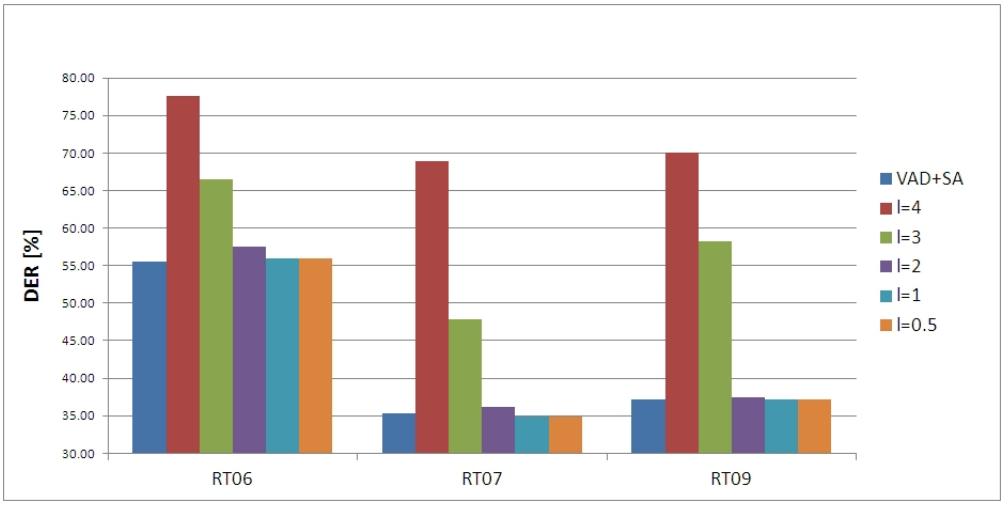


Figure 6.10: Clustering analysis for the NIST RT AMI meetings

results from the TDOA analysis, i.e. the *all* scheme shown in Figure 6.9 for QIO-FE VAD and BeamformIt acoustic beamforming.

The remaining columns show the results after speech clustering for λ values of 4, 3, 2, 1 and 0.5 (as defined in Equation 2.38). The SA output from the TDOA analysis contains 36 possible speakers. I expected the speech clustering algorithm to reduce this to the real number (which is known to be four for all EDI, TNO and IDI RT meetings). Unfortunately, the ΔBIC algorithm was unable to merge the apparent 36 speakers down to four, as shown in Table 6.1.

Table 6.1: Number of speakers detected from the VAD+SA output before and after speech clustering for the NIST RT AMI meetings

	Meeting	Number of speakers detected	
		before clustering	after clustering
RT06	EDI_20050216-1051	17	17
	EDI_20050218-0900	22	20
	TNO_20041103-1130	29	28
RT07	EDI_20061113-1500	28	26
	EDI_20061114-1500	23	23
RT09	EDI_20071128-1000	24	24
	EDI_20071128-1500	24	24
	IDI_20090128-1600	15	15
	IDI_20090129-1000	24	24

At best the merging algorithm managed to reduce the number of speakers present by

a maximum of two, but most of the time no clustering took place. I carried out a series of experiments using the ΔBIC as the decision criterion as to whether to merge two different speech segments, using the acoustic features (MFCCs) as well as GMMs trained on these same speech segments. Unfortunately, the ΔBIC needs a few seconds of speech data from each of the two speech segments to produce reliable results. This is not what is observed in real speech from meetings, where the average speech segment length is around 1.5 s and the mean speech segment length can be as low as 1 s, as shown in Figure 6.11.

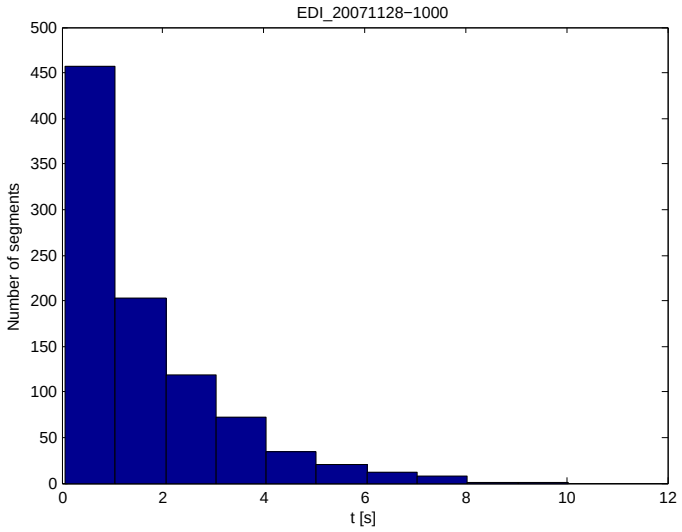


Figure 6.11: Analysis of speech length in NIST RT09 meeting EDI_20071128-1000

My experiments on speaker clustering using the ΔBIC criterion show that the proposed scheme only works well for minimum speech segment lengths of 3 s or more. When compared with the average speech segment length of around 1.5 s in meetings, the ΔBIC criterion fails to process most of the speech segments correctly, as demonstrated in Table 6.1.

The limitations of the ΔBIC criterion not only apply to clustering but also speech segmentation. My experiments on speech segmentation using the ΔBIC criterion (both on speech features, i.e. MFCC vectors, and GMMs of MFCC vectors) showed that Equation 2.41 will almost always peak for any speech segment input. This results in undesired splitting of the speech segments, making the speech segments even shorter and the clustering of speech more difficult.

Note that for the results presented in Figure 6.10 speech segmentation is bypassed as it only splits everything into minimum length segments, leading to further degradation of the output, as discussed above.

In order to overcome the problem of comparing short segments of speech I generate a speaker matrix, as presented in the next Section.

6.4 Speaker matrix (SM)

Assuming that there is a speaker in every SA sector and given a pure speaker model for each of them then, if the ΔBIC value of every incoming speech segment with the reference speech segment of each sector is calculated according to Equation 2.41, then only the true speaker will lead to a positive ΔBIC value.

If I now create a speaker matrix (of dimension $N \cdot N = 36 \cdot 36$) in which one axis is the active sector (from the SA map) of the incoming speech segment while the other is the reference speech segments, then in theory only elements on the diagonal of the speaker matrix will lead to positive ΔBIC values.

Leading on from here, if a counter value at position (x_i, y_j) of the speaker matrix (with $0 \leq x_i < N$ and $0 \leq y_j < N$) is incremented if $\Delta BIC > 0$, then every entry in the speaker matrix represents a speaker position.

This leads to the following algorithm:

1. load VAD output and sector activity (SA) map
2. select reference speech segment for each sector
3. create empty speaker matrix
4. for each incoming speech segment
   (a) calculate ΔBIC for each reference speech segment
   (b) if $\Delta BIC > 0$ then add 1 in speaker matrix at that position

First the VAD output and SA map are loaded. Then a reference speech segment needs to be selected for each sector. To keep this simple and to avoid sophisticated algorithms to determine the best reference segment, the longest speech segment is chosen for every

sector from the VAD output and SA map. This reference speech segment will be stored as s_i.

After this, using the VAD output, one speech segment after another is loaded in sequence and its sector looked up in the SA map. This sector corresponds to j and the corresponding speech segment is s_j. Now, for $0 \leq n < N$ the ΔBIC can be calculated according to Equation 2.41 and the count value of the speaker matrix (SM) at position (i, j) incremented if $\Delta BIC > 0$.

6.5 Speaker diarisation

In the ideal case the speaker matrix would only have entries on its diagonal because the originally assigned sector would be the same as the sector with the highest ΔBIC score. The indices of the entries would then correspond to sectors with speakers. In reality this is not the case, as shown in Figure 6.12.

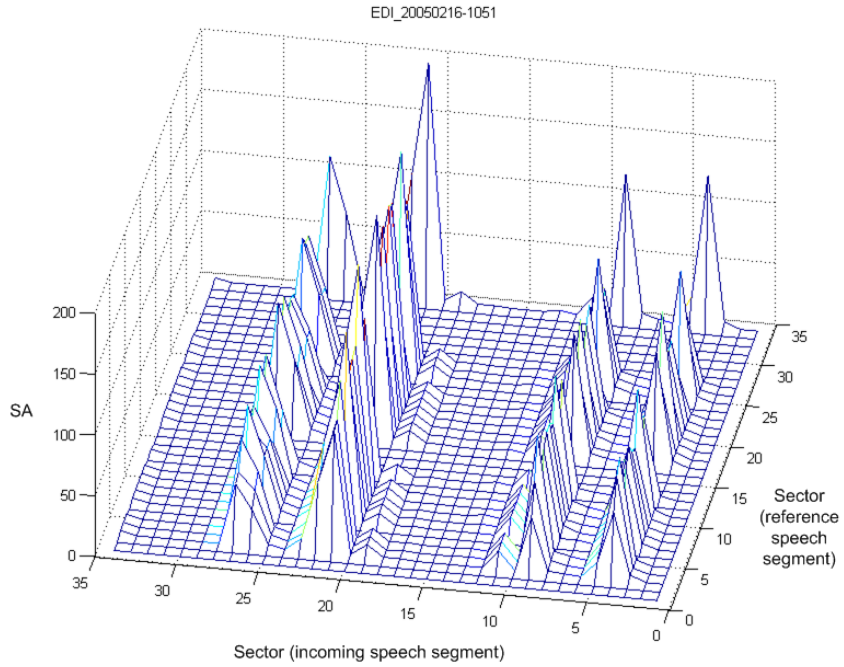


Figure 6.12: Speaker matrix (SM) from NIST RT06 meeting EDI_20050216-1051

The peaks in the SM tend to cluster in rows. In order to identify the sectors with speakers I look for peaks in the entries on the diagonal of the sector matrix. The

indices of the peaks correspond to the sector where I estimate a speaker is located – these are the speaker sectors, shown in Figure 6.13.

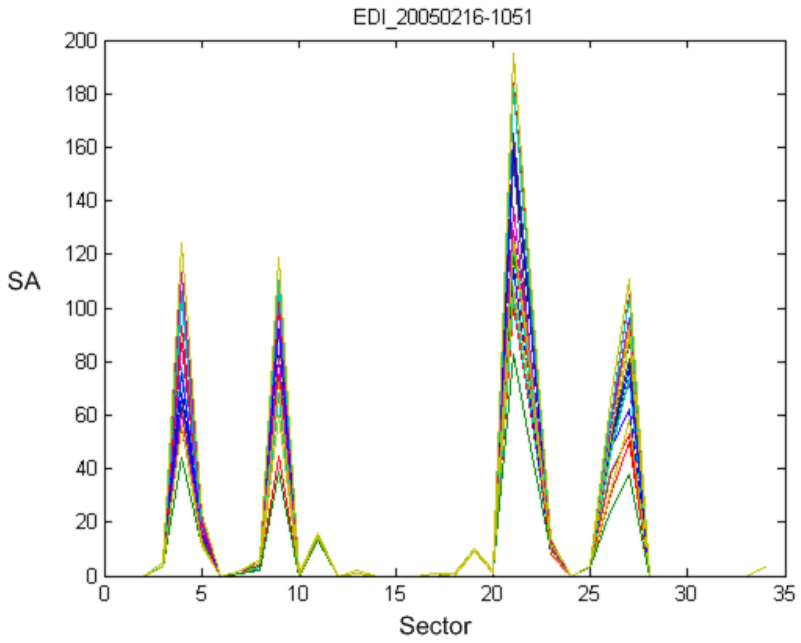


Figure 6.13: Speaker matrix peaks from NIST RT06 meeting EDI_20050216-1051

Given the VAD output, the sector activity map and the speaker matrix, the number of speakers and their position can now be determined after which clustering can be performed to complete the diarisation process. This is carried out according to the following algorithm:

1. load VAD, SA and SM
2. find peaks on the diagonal of the speaker matrix
   (a) determine number of speakers
3. for each incoming speech segment
   (a) assign incoming speech segment to closest speaker (i.e. SM peak)

After loading the VAD output, the sector activity map and speaker matrix I find the maximum peak on the diagonal of the speaker matrix $SM_{peak} = max(SM_{ii})$ for $0 \leq i < N$. Then, setting a threshold of $SM_{th} = 0.1 \cdot SM_{peak}$, I look for any other value on the SM diagonal that exceeds SM_{th}. These are the speakers and their positions. I determined the threshold parameter 0.1 used to calculate SM_{th} by visually analysing the noise levels of a few speaker matrices.

In the last diarisation step, the clustering process, I load the output of the VAD+SA system described above and replace the speaker entries (from the SA) with the position of the closest speaker, looking at the distance on the circle.

To summarise, in the first pass over the data the sector activity is calculated. In the second pass the speaker matrix is generated and the speakers and their positions determined. Finally, in the third pass, each speech segment is assigned to the closest speaker found in the speaker matrix, thus generating the diarisation output. The complete process is shown in Figure 6.14.

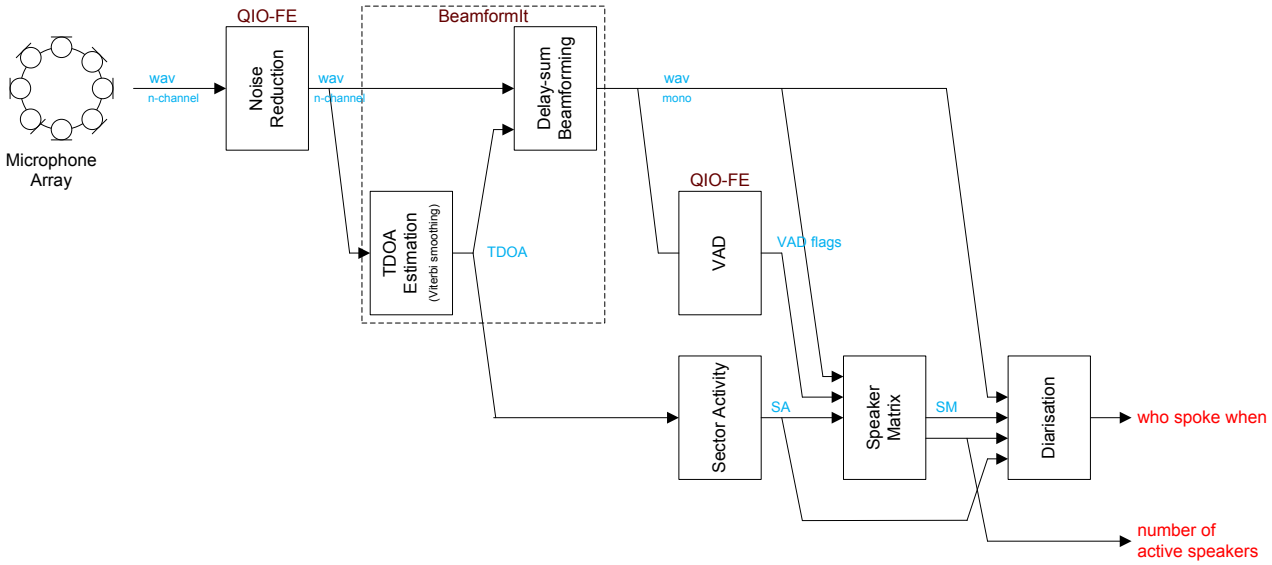


Figure 6.14: Flow diagram to determining the number of speakers in a meeting

A different description of the proposed algorithm can also be obtained from the Patent application [Zwyssig, 2012].

The beauty of this straightforward process is that it can be converted into an online system by means of a few simple steps, as described in the next section.

6.6 Online diarisation

Online, never-ending or incremental (real-time) speech processing systems start in some pre-trained status and process the incoming speech on-the-go without stopping and with limited delay to the output while learning from the data that is processed and

therefore improving the output. Such systems need to be real-time, i.e. it must take less time to compute the output from an incoming speech segment than the length of said speech segment, otherwise delays would accumulate and the system eventually stall. The time it takes to process a segment of data is called latency. Online operation therefore requires that the latency t_l of a speech segment of length t_s is smaller or equal to t_s, i.e. $t_l \leq t_s$. Online operation also demands that the data is processed continuously and in real-time without prior availability.

Naturally such a system does need some time to process the data, i.e. signal processing delays should be limited and acceptable to the receiver of the output, e.g. downstream processes.

The most complex, effective and sophisticated online diarisation (and speaker and speech recognition and speech understanding) system known at present was designed by researchers at the NTT (Nippon Telegraph & Telephone Corp, see [Ishiguro et al., 2012] for details), as presented in Sections 2.4 and 6.1.

It is important that the algorithm presented in this chapter is suitable for online processing. The few steps necessary to do this are presented in what follows.

Determining the number of (active) speakers in a meeting, as presented above, requires some pre-processing, i.e. noise suppression, acoustic beamforming and VAD. All of these algorithms work in an online manner while introducing some signal delay of a few hundred milliseconds. It can therefore be assumed that the acoustic signal, the TDOAs from the beamforming and the VAD output are available simultaneously for online processing.

Calculating the sector activity is therefore straightforward, implying a further signal processing delay of a few seconds (c.f. Section 6.3).

The critical component in online processing is the speaker matrix which requires reference speech for each sector. This is obviously not available from the beginning. On the contrary, diarisation following the VAD+SA principle outlined above requires nothing more that the sector activities and the VAD output.

I therefore suggest starting the online diarisation system using the VAD+SA scheme while gathering reference speech data for the individual sectors of the speaker matrix. As soon as a speech segment of suitable purity arrives for a sector then the speaker matrix can start operating as outlined in the VAD+SA+SM scheme. During the start-

up phase, if only a few speakers appear present from the speaker matrix, clustering can be executed as a mixture of the VAD+SA and VAD+SA+SM algorithms, i.e. if the distance of an incoming speech segment to the closest speaker exceeds a pre-defined threshold, then it is not assigned to a speaker in the matrix but remains at the location as per the VAD+SA algorithm.

The reference speech segment for a sector can also be updated if a 'better' speech segment arrives, and the counts in the speaker matrix should incur a forget-me factor, i.e. should decrease over time as defined by a constant τ_{SM}.

The novel algorithm to determine the number of active speakers presented here is therefore not just straightforward but also well suited to off- and online computing.

6.7 Diarisation results

This section presents the results from the speaker diarisation experiments that were carried out following the flow defined in Figure 6.14. Pre-processing of the eight-channel audio data was carried out by performing single channel noise reduction using the QIO-FE followed by acoustic beamforming using BeamformIt, after which VAD was performed on the single enhanced audio channel using the QIO-FE.

Four diarisation experiments were conducted on the EDI, IDI and TNO meetings from the RT06, RT07 and RT09 data sets (the "NIST RT AMI meetings"), a total of nine meetings. First the DER of the direct output of the sector activity map (VAD+SA) was calculated, i.e. diarisation was performed with a fixed number of 36 speakers. This was followed by the evaluation of the new algorithm (VAD+SA+SM) that uses the number of speakers determined from the speaker matrix. Finally, in order to provide baseline results, two open source diarisation systems, the SHoUT speech recognition toolkit and the LIUM speaker diarisation system, were used to perform the diarisation.

The two baseline systems provide a comparison of top-down vs. bottom-up diarisation: SHoUT uses bottom-up processing and LIUM top-down. The SHoUT diarisation system was chosen here (in preference to ICSI) as it is open source software.

The results of the four experiments presented in Table 6.2 and Figures 6.15 and 6.16 show that providing the diarisation system with an accurate estimate of the active number of speakers present results in a considerable improvement in the diarisation output.

Table 6.2: [%] DER, VER, FA, MS and estimated number of speakers (spkrs) for each meeting for the NIST RT AMI meetings. FA denotes false alarm, MS denotes missed speech.

			VAD+SA (basic)					VAD+SA+SM		SHoUT		LIUM	
	Meeting	spkrs	DER	VER	FA	MS	spkrs	DER	spkrs	DER	spkrs	DER	spkrs
RT06	EDI_20050216-1051	4	48.1	9.0	2.9	6.1	36	31.0	4	45.3	10	65.7	7
	EDI_20050218-0900	4	53.2	8.9	2.5	6.4	36	30.4	5	49.8	9	67.2	7
	TNO_20041103-1130	4	65.3	8.0	1.8	6.2	36	57.1	7	46.8	12	61.7	2
	avg (RT06)		55.5	8.6	2.4	6.2		39.4		47.4		64.9	
RT07	EDI_20061113-1500	4	44.6	4.7	4.2	0.5	36	31.8	4	56.7	14	72.5	1
	EDI_20061114-1500	4	27.5	6.3	5.7	0.6	36	20.3	4	23.4	10	64.1	6
	avg (RT07)		35.3	5.6	5.0	0.6		25.6		38.7		67.9	
RT09	EDI_20071128-1000	4	34.6	3.9	3.2	0.7	36	16.7	4	23.4	8	56.2	3
	EDI_20071128-1500	4	46.0	5.3	4.7	0.6	36	35.1	5	31.0	13	82.3	2
	IDI_20090128-1600	4	27.8	1.6	0.7	0.9	36	12.0	4	23.8	9	19.7	19
	IDI_20090129-1000	4	41.7	4.9	4.0	0.9	36	25.5	4	34.4	14	41.7	12
	avg (RT09)		37.2	3.9	3.1	0.8		21.9		28.0		48.7	
	avg (all)		42.3	5.6	3.2	2.4		27.8		35.8		57.1	

The basic VAD+SA method achieves an improvement of 26% relative / 15% absolute compared to the LIUM tool. The VAD+SA+SM outperforms both SHoUT and LIUM, giving an improvement of 51% relative / 29% absolute compared to LIUM and 22% relative / 8% absolute compared to SHoUT. In addition, the number of speakers estimated by the VAD+SA+SM system almost considerably closer to the actual number of speakers in the meeting than either of the other systems.

It is unfortunately difficult to compare the DERs presented in Table 2.3 with state-of-the-art performance as the authors of these diarisation systems do not provide detailed results. I have managed to obtain comprehensive results from the SHoUT diarisation toolkit (Marijn Huijbregts, personal communication, 23 November 2011) and compared my findings (presented here) with those achieved by SHoUT on the NIST RT06, RT07 and RT09 meeting data. The figures presented here match the results achieved by Marijn Huijbregt with less than 2% absolute DER error. Marijn Huijbregt and I have not managed to achieve identical DERs despite detailed tuning of many parameters.

Next, a detailed analysis of the results presented in Table 6.2 is carried out. This is done using boxplots for easier evaluation, where the median is shown as the interface

between the brown and red boxes, the lower quartile as a brown box, the upper quartile as a blue box and the minimum and maximum DER as whiskers.

Figure 6.15 shows the average DER listed for the RT06, RT07 and RT09 data sets, and Figure 6.16 shows the median, lower and upper quartile and minimum and maximum DER combined for all RT data sets for the different algorithms.

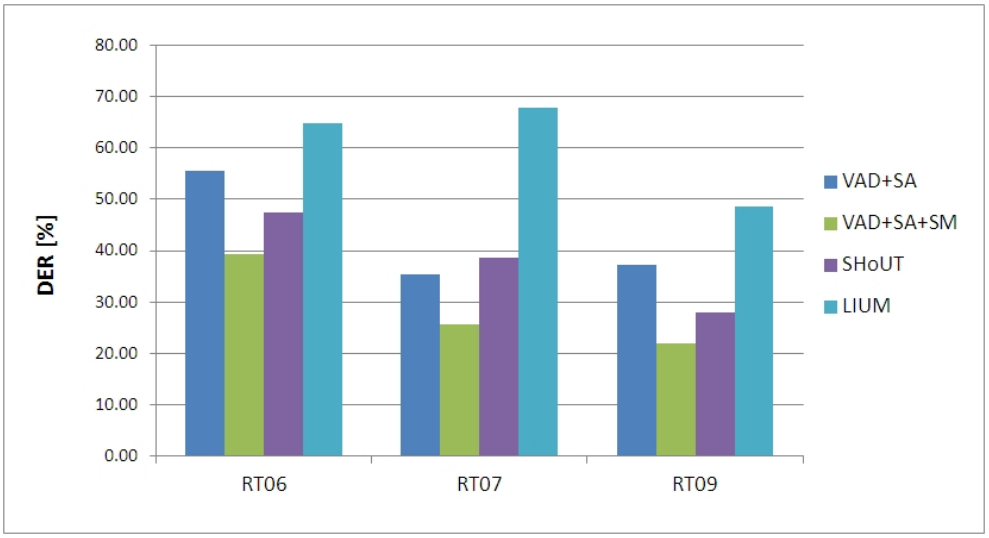


Figure 6.15: [%] DER for all algorithms for the NIST RT AMI meetings

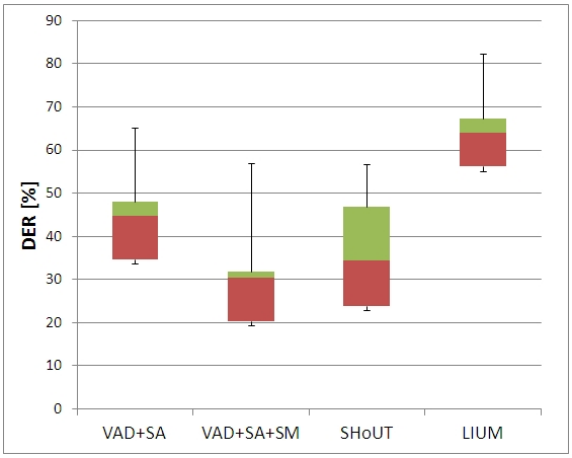


Figure 6.16: [%] DER variance for the complete NIST RT AMI meetings

As shown in Figures 6.15 and 6.16, the VAD+SA and VAD+SA+SM algorithms consistently outperform the LIUM diarisation tool, while the VAD+SA+SM algorithm also outperforms the SHoUT tool – both by a reduced average error and reduced mean and interquartile range. Note though that diarisation using the VAD+SA algorithm corresponds to diarisation using location features only, diarisation with the SHoUT and

LIUM toolkits uses acoustic features only, while the VAD+SA+SM algorithm uses acoustic and location features.

Huijbregts et al. [2012] report 26.6% DER for the SHoUT diarisation system on the complete test set of the NIST RT09 using acoustic features only. The 26.6% DER reported is comparable with the 28.0% DER presented in Table 6.2. No published results exist for the performance of the SHoUT tool on the NIST RT data using acoustic and localisation features. The ICSI tools achieve 17.2% DER on the RT09 data set using acoustic, other speech and localisation features (see Table 6.2). Unfortunately, no data is available as to the number of speakers (or final number of clusters) detected for any diarisation system on the NIST RT09 evaluation data.

This makes it difficult to compare the proposed VAD+SA+SM algorithm with state-of-the-art diarisation systems, not least because the proposed algorithm does not contain segmentation and clustering.

The main benefit of the VAD+SA+SM algorithm is that it provides any diarisation system with a reliable and accurate estimate of the number of speakers present. This information can be of significant advantage to any diarisation system as well as any downstream processes such as speaker identification or speech recognition.

In addition to the NIST RT meetings, it is also of interest to verify the novel algorithm for determining the number of speakers on the 2012_MMA corpus (cf. Chapter 4).

I verified the VAD+SA+SM algorithm on the four subsets of the 2012_MMA corpus, the single speaker task WSJ and the two overlapping speakers task MSWSJ in the two different environments, that is, in a meeting room and in a hemi-anechoic chamber. The experiments were carried out following the flow outlined in Figure 6.14. Acoustic beamforming was carried out using BeamformIt and the mdm tools. The results are presented in Table 6.3.

Twelve single speakers were recorded for the WSJ and WSJ_anechoic data sets, i.e. the true number of speakers is one. Six pairs of speakers were recorded for the MSWSJ and MSWSJ_anechoic data sets, i.e. the true number of speakers is two. Table 6.3 shows the number of recordings for the different number of speakers found for five different microphone arrays, using the VAD+SA+SM algorithm.

In the first row for the WSJ dataset, for the analogue microphone array with a diameter of 20 cm, sampled at 16 kHz, using the BeamformIt tool, the correct number of one

Table 6.3: Number of single speaker recordings (WSJ + WSJ_anechoic) and number of dual speaker recordings (MSWSJ + MSWSJ_anechoic) for different numbers of speakers detected using the acoustic beamformers BeamformIt and mdm tools on the 2012_MMA corpus. Numbers in red indicate that the correct number of speakers was detected; orange indicates near correct detection.

Data set	Microphone	Diameter [cm]	Fs [kHz]	Number of speakers detected																		
				BeamformIt							mdm tools											
				1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10	11	12	13
WSJ	Analogue	20	16	10	1	1					9	2	1										
	Analogue	4	96	12									2	5	4			1					
	Digital	20	16	9	3						4	8											
	Digital	4	96			1	6	3	2			2	7	3									
	Digital	4	48		1	3	2	3	2	1			11	1									
WSJ_anechoic	Analogue	20	16	10		2					10	1	1										
	Analogue	4	96		5	1	5	1				1	3	1	3	2	2						
	Digital	20	16	12							11	1											
	Digital	4	96		5	2	2	1	2						1	3		3	3	1			1
	Digital	4	48		2	3	7				2	3	2	2	1		1				1		
MSWSJ	Analogue	20	16		6							3	3										
	Analogue	4	96	2	4									2	2		2						
	Digital	20	16		6							6											
	Digital	4	96	2		1	1	1	1					4	2								
	Digital	4	48		1	2	1	1	1				2	2	2								
MSWSJ_anechoic	Analogue	20	16	1	3	2						4	2										
	Analogue	4	96		1	2	2	1						1		2	1	1	1				
	Digital	20	16		6							6											
	Digital	4	96		2	4							4	2									
	Digital	4	48		1	4	1						1	2	2	1							

speaker was found in 10 of 12 recordings, while in one recording two speakers and in another recording three speakers were detected. For the mdm tools, the correct number of one speaker was found in 9 of 12 recordings, while in two recordings two speakers and in another recording three speakers were detected.

In the third row for the MSWSJ dataset, for the digital microphone array with a diameter of 20 cm, sampled at 16 kHz, using either beamformer, the correct number of two speakers was detected in all six recordings.

As shown in Table 6.3, the proposed algorithm to determine the number of speakers in a meeting performs very well on the analogue and digital microphone arrays of diameter 20 cm. BeamformIt unfortunately only supports audio input sampled at 8 or 16 kHz. Acoustic input at any other sample rate is automatically re-sampled to 16 kHz. For the microphone arrays with diameter 4 cm sampled at 48 and 96 kHz, down-sampling to 16 kHz leads to a huge loss in localisation accuracy. Indeed, if we calculate the maximum possible TDOA as per Equation 6.2, then, given the specifications used for the recordings of the 2012_MMA corpus, we get the following Table 6.4.

Table 6.4: Maximum possible TDOA values for a given microphone array dimension and audio sample rate

D [cm]	Fs [kHz]	TDOA_{max}
0.20	16	$\pm$9.3
0.04	16	$\pm$1.9
0.04	48	$\pm$5.6
0.04	96	$\pm$11.2

When analysing the TDOA values that were generated using GCC-PHAT and the two-stage Viterbi smoothing, I found them to be very stable for the analogue microphone array with a diameter of 4 cm (IMR only) despite the small range. Unfortunately, this was not the case for the audio signal in the hemi-anechoic chamber and for the digital MEMS microphone array with a diameter of 4 cm. These TDOA values exceed the maximum possible values most of the time which explains the poor performance of the algorithm using BeamformIt on out-of-domain data.

The mdm tools support acoustic beamforming at any sample rate. The TDOA estimates should therefore work well for microphone array speech recorded at 48 and 96 kHz. The results for the mdm tools presented in Table 6.3 unfortunately show poor

performance of the algorithm on all 4 cm microphone arrays.

I conclude that two requirements are crucial for the correct output of the proposed algorithm to determine the number of speakers. First, the acoustic data needs to be sampled at a sufficiently high frequency to give good TDOA estimates and, second, TDOA smoothing as implemented in BeamformIt [Anguera et al., 2007] is essential. Given these two factors, the sensitivity of the algorithm to the recording environment should disappear.

6.8 Summary and conclusions

This chapter looked at the importance of knowing how many speakers are present in a multi-party conversation, particularly relating to speaker diarisation, i.e. who spoke when. A review of state-of-the-art diarisation systems shows that these systems do not attempt to find this number and that creating a diarisation output using acoustic models with the true number of speakers does not result in the best DER performance.

Obtaining a good estimate of the true number of active speakers is nevertheless a useful parameter for diarisation and downstream processes, as was demonstrated by the work presented in this chapter and also proved by the current best-performing speaker diarisation system.

I have proposed a TDOA-based algorithm to determine the number of active speakers in a meeting and applied this to the diarisation task. The proposed algorithm outperforms ΔBIC-based diarisation tools due to its improved estimation of the number of speakers in the meeting. The algorithm is computationally less expensive than ΔBIC-based methods and can be easily adapted so as to require only a single pass over the data, making it suitable for online processing.

The proposed algorithm is not restricted to meeting recordings but ports well to any multi-party conversation recorded with multiple microphones where the microphone positions are known and the individual audio channels are synchronised. Assuming a circular microphone array, two requirements are essential for the algorithm to work correctly: first, the microphones must be placed at a sufficient distance for lower sample rates (i.e. 20 cm for Fs = 16 kHz) or sampled at higher rates for smaller geometries (96 kHz for 4 cm) and second, smoothing of the TDOA values must be applied.

The algorithm performs well on NIST RT data and the 2012_MMA corpus and gives a much improved estimate of the active number of speakers compared to two publically available state-of-the-art diarisation systems , the SHoUT bottom-up and the LIUM top-down diarisation tools.

Please note that in the NIST RT data and the 2012_MMA corpus speakers typically remain in a fixed location and do not move around. Such a restriction is not realistic and moving speakers would significantly increase the error rate for TDOA-based systems.

Chapter 7

Speech separation

7.1 Introduction

Overlapping speech is common in normal human conversations such as meetings or any situation involving two or more people [Shriberg et al., 2001]. Humans are exceptionally good at localising speakers and separating overlapping speech, even with monaural hearing. This is due to the sound characteristic in the ear produced by the shape of the pinna and the sound shade from the head and torso. Humans also constantly move their head which assists sound source localisation and speech separation. These human capabilities greatly surpass any machine algorithm, particularly in noise [Good and Gilkey, 1996].

Attempts to mimic human speech localisation and separation have not so far produced satisfactory results [Cooke et al., 2010, Barker et al., 2012] and a more practical and successful solution to the problem appears to be to use arrays of multiple microphones. If an audio signal is sampled at an appropriate frequency (observing Nyquist's law) and at the correct physical intra-microphone distance (adhering to the wavelength of sound derived from the frequency range of human speech and the speed of sound in air) then, using three or more microphones, speech can be localised and separated with good accuracy [McDonough et al., 2008a].

This chapter presents speech separation and recognition experiments carried out on the multiple microphone array corpus of single and overlapping speech (2012_MMA, cf. Chapter 4) where the effect of post-filtering, echo suppression and binary masking are looked at.

The work presented in this chapter was carried out in collaboration with Friedrich Faubel from the Saarland University and divided as follows. Recording and file preparation were carried out by myself and speaker localisation and speech separation were performed using the Saarland University Beamforming Library which they very kindly made available. Executing the speaker localisation and speech separation is computationally very demanding and was therefore divided between the computers available to Friedrich Faubel and myself. I then conducted the speech recognition, model adaptation and scoring.

7.2 Prior work

A review of overlap in natural speech and overlapping speech corpora was presented earlier in Section 2.5.2. The work described there gives an analysis of the amount and nature of overlapping speech in meeting conversations, but does not look at detecting overlapping speech.

This section reviews detection and separation of overlapping speech.

Prior to speech separation, overlapping speech in multi-party conversations will need to be detected and, in the case of speaker diarisation, attributed to the correct speaker. Correctly detecting overlapping or simultaneous speech is an open research topic. Several groups have carried out research in detecting overlapping speech in meetings and they usually report their results in terms of an improved DER.

Zelenak et al. [2012] present a review of the state-of-the-art of overlapping speech detection. The authors divided the research roughly into two main areas, overlapping speech detection using close talking microphones (i.e. each speaker present wears a headset or lapel microphone) and overlap detection on distant speech data.

Overlap detection using distant microphones can be carried out using acoustic and intra-microphone TDOA data, as proposed by Boakye et al. [2008], Vipperla et al. [2012] and Zelenak et al. [2012].

In the original ICSI system [Friedland et al., 2012], VAD was carried out using two HMM-based GMMs, one for speech and one for non-speech. Boakye et al. [2008] modified the ICSI diarisation system's VAD component by adding an additional GMM trained to detect overlapping speech. Testing the modified VAD and diarisation engine

on the AMI development set, the authors measured a DER reduction of 1.3% absolute and 3.6% relative from 38.1% to 36.8% using the speech from the single distant (far-field) microphone, i.e. for the sdm condition.

A detailed analysis shows that the modified system actually had an increased false alarm detection rate from 0% to 1.8%, an increased speaker error from 19.8% to 20.3% and a decreased missed speech error from 18.3% to 14.6%, resulting in an overall DER improvement. Boakye et al. [2008] failed to report the actual overlap detection rate of their method.

Looking at the proposed system and the findings presented in Sinclair and King [2013] it could be inferred that the system presented by Boakye et al. [2008] is actually unable to detect overlapping speech. Splitting the speech model into single and overlapping speech components is more likely to generate a 'purer' speech model, resulting in a significant decrease in missed speech at the cost of increased false alarms and speaker error.

Vipperla et al. [2012] used an oracle overlap detection component to investigate the detection capability of their diarisation engine. A similar effect to that seen in the ICSI system can be observed, i.e. significantly reduced missed speech at the cost of increased false alarms and speaker error. Vipperla et al. [2012] used pure (non-overlapping) speech for each given speaker to learn base models using spectral magnitude features, therefore training a speaker model a priori. Incoming speech was then classified using convolutive, non-negative sparse coding (CNSC) to capture spectro-temporal patterns. Overlapping speech was detected if two speaker models activated, i.e. if their CNSC activation energy exceeded a pre-defined threshold. The proposed scheme requires that prior speaker models exist for each speaker in the test data.

Zelenak et al. [2012] complemented their GMM-HMM baseline system with an overlap detection component which combines spectral (audio) and spatial (TDOA) features. The spectral features used are MFCCs, spectral flatness and the prediction error from the LPC residual signal generation. The spatial features used are the TDOA coherence value, i.e. the value of the principal peak of the GCC, the coherence dispersion ratio and the delta value of two adjacent TDOA estimations.

The authors trained three GMM-HMM models, a silence model, a speech model and an overlapping speech model and performed diarisation using Viterbi decoding. Dimensionality reduction and normalisation using sequential principal component analysis

(PCA) was necessary to merge the two parallel streams in the GMM-HMM model.

The authors claim that:

> "... [they] have observed that the time delay estimates produced by the GCC-PHAT jump from one speaker to another at a very high rate as one source dominates due to the non-stationarity of the voice."

This statement is very surprising as it does not agree with my observation that TDOAs are stable over a time period of several hundred milliseconds, justifying the typical choice for calculating TDOAs every 256 ms for the BeamformIt tools and every 500 ms for the mdm tools. Zelenak et al. [2012] also observed that the median overlap duration in the AMI meeting corpus is rather short at 0.46 s. The proposed algorithm managed to correctly detect 20% of the overlapping speech on the AMI data set and approximately 5% on the RT09 data set. Overall the DER decreases from 42.2% for the baseline system by 1.2% absolute/2.7% relative to 41.0% for the proposed improved diarisation system with overlap detection.

Looking at the detailed false alarm, missed speech and speaker error figures, I noticed the same behaviour as presented by Boakye et al. [2008] and Vipperla et al. [2012], i.e. a reduced false alarm rate compared to an increased missed speech and speaker error rate, again indicating that the DER improvement might well be attributed to 'purer' speaker models than to overlap detection.

Unfortunately, the improvements suggested by these researchers are limited and few results could be found that actually report on how much overlap was really detected. Detecting overlapping speech therefore remains an open research problem.

The challenges to modern ASR systems posed by overlapping speech will be looked at next.

The most systematic work in the field, using recordings of overlapped speech, is based on the multi-channel Wall Street Journal audio visual (MC-WSJ-AV) corpus [Lincoln et al., 2005], released for the second PASCAL Speech Separation Challenge (SSC2).

Initial experiments on these recordings demonstrated that the speech recognition WER for overlapping speech can easily be double or triple that of a comparable single speaker scenario [Himawan et al., 2008, McDonough et al., 2008a]. More recent experiments on the single speaker part of the MC-WSJ-AV corpus have shown that it is important for distant speech recognition to use sophisticated front-end processing on multiple input channels [Kumatani et al., 2012]

Speech recognition using a single distant microphone suffers significantly from additive noise and reverberation. Overcoming this speech degradation is difficult using back-end compensation techniques such as vector Taylor series (VTS), cMLLR or a combination of both [Gales and Wang, 2011], or Bayesian feature enhancement [Krueger et al., 2012]. Kolossa et al. [2011] achieved promising results by combining sophisticated front-end processing and back-end compensation techniques.

Although there has been a lot of recent research activity in single speaker distant speech recognition (e.g. the CHiME challenge [Barker et al., 2012]), this has typically involved the artificial creation of data by convolving close-talking speech recordings with a multi-channel room impulse response and then adding noise. Ideally, however, the corpora would be recorded in different natural environments in order to capture the way in which speakers change their speaking style in noise and reverberation [Pelegrín-García et al., 2011], and this has motivated the collection of the 2012_MMA corpus.

The following sections present algorithms and results on overlapping speech separation. First, (multiple) speaker localisation and speech separation algorithms are discussed after which the experimental setup and data used are presented. This is followed by the results achieved in our experiments and an analysis of them. The chapter closes with a summary.

7.3 Speech separation

The next section describes the proposed speech separation algorithm which separates overlapping speech using a combination of spatial filtering and crosstalk cancellation as originally suggested by Himawan et al. [2008] and McDonough et al. [2008a]. This is achieved using a two-stage approach in which (1) an initial beamformer separates the speech based on spatial diversity after which (2) a cross-talk canceller further improves the separation by post-processing the beamformer outputs. The section closes with a description of the speaker localisation system.

7.3.1 Superdirective Beamforming

Beamforming provides an elegant way to extract the signal from a desired source through spatial filtering. In the MSWSJ data sets the task is to separate two speak-

ers located at directions

$$\mathbf{a}_k = [\cos\theta_k \cos\phi_k \;\; \sin\theta_k \cos\phi_k \;\; \sin\phi_k]^T, \quad k \in \{1,2\} \tag{7.1}$$

with θ_k and ϕ_k denoting the azimuth and elevation in relation to the array.

In order to use SDB for speech separation, a beamformer is pointed at each of the speakers. The beamformer outputs $Y_1(\omega,t)$ and $Y_2(\omega,t)$ are obtained according to Equations 2.9, 2.10 and 2.11 (cf. Section 2.1.2), and the corresponding separated speech signals $y_1(t)$ and $y_2(t)$ are recovered using inverse Fourier transformation followed by overlap-and-add.

As mentioned above, a two-stage approach is used for improved speech separation, i.e. a post-processing stage follows the superdirective beamforming. Two post-processing methods are proposed: (a) cross-talk cancellation using binary masking and (b) residual echo suppression. The next two sections give details of these algorithms.

7.3.2 Cross-Talk Cancellation

Overlapping speakers tend to use different frequency bands at a point in time [Belouchrani and Amin, 1998]. Yilmaz and Rickard [2004] demonstrated that this effect can be used successfully for speech processing. The authors employed a post-processing step in which the beamformer outputs $Y_k(\omega,t)$ are multiplied by a binary mask M_k whose components $M_k(\omega,t)$ identify which frequencies a speaker uses at time t, i.e.

$$\hat{S}_k(\omega,t) = M_k(\omega,t) \cdot Y_k(\omega,t), \quad k \in \{1,2\}. \tag{7.2}$$

Near perfect demixing would be possible if the true masks were known [Yilmaz and Rickard, 2004]. In practice, $M_k(\omega,t)$ needs to be estimated. This can be achieved by comparing the power at the beamformer outputs $Y_1(\omega,t)$ and $Y_2(\omega,t)$ and then allocating the time-frequency unit (ω,t) to the stronger output [Himawan et al., 2008], where

$$\hat{M}_k(\omega,t) = \begin{cases} 1, & |Y_k(\omega,t)|^2 \geq |Y_l(\omega,t)|^2 \quad \forall \quad l \\ 0, & \text{otherwise} \end{cases}. \tag{7.3}$$

The performance of binary masking can be improved in practice in two steps: (1) smoothing $|Y_k(\omega,t)|^2$ over time by convolving with a triangular filter kernel and (2) through Welsh averaging of the smoothed masks $\hat{M}_k(\omega,t)$. This results in

$$\bar{M}_k(\omega,t) = \alpha\, \bar{M}_k(\omega,t-1) + (1-\alpha)\, \hat{M}_k(\omega,t) \tag{7.4}$$

with $\alpha = 0.9$. Yilmaz and Rickard [2004] found that the optimum window length for time-frequency masking is about 1024–2048 samples (at a sampling rate of Fs = 16 kHz). We therefore used an FFT of length $L = 2^{\log_2(F/32)}$ with a window shift of $L/32$.

7.3.3 Residual echo suppression

Residual echo suppression is an alternative method to suppress the second of two overlapping speakers when carrying out speech recognition for one speaker. Enzner et al. [2002] define residual echo as the remainder of insufficient echo cancelling in hands-free telephone equipment. They propose that residual echo suppression is performed in two stages: first the residual echo is estimated using the coherence function after which it is suppressed using Wiener filtering.

Siegwart et al. [2012] found that residual echo suppression, used as a special speech separation post-filter, leads to improved speech separation and WER on the overlapping speaker task of the MC-WSJ-AV corpus.

In the first step, the cross-power spectral density (CSD) $\bar{\Phi}_{y_1y_2}$ and the power spectral densities (PSD) $\bar{\Phi}_{y_1y_1}$ and $\bar{\Phi}_{y_2y_2}$ are calculated from the instantaneous CSD and PSD Φ values using Welsh averaging, as

$$\bar{\Phi}(\omega) = \alpha\, \bar{\Phi}(\omega) + (1-\alpha)\, \Phi(\omega) \tag{7.5}$$

with α set so as to get a decay time of 25 ms (for best performance as per Siegwart et al. [2012]), taking account of the sample rate of the input audio signal and the FFT frame size and shift.

The coherence $\gamma_{y_1y_2}$ between Y_1 and Y_2 can now be calculated as

$$\gamma_{y_1y_2}(\omega) = \frac{\bar{\Phi}_{y_1y_2}(\omega)}{\sqrt{\bar{\Phi}_{y_1y_1}(\omega)\bar{\Phi}_{y_2y_2}(\omega)}}. \tag{7.6}$$

The residual part of Y_i that is contained in Y_j is then approximated as

$$\hat{R}_i(\omega) = \gamma_{y_1y_2}(\omega) \cdot Y_j(\omega), \quad j \neq i. \tag{7.7}$$

Replacing $|Y_i|^2$ with $\bar{\Phi}_{y_jy_j}$, the residual power $\hat{\Phi}_{r_ir_i}$ can be obtained by taking the magnitude square as

$$\hat{\Phi}_{r_ir_i}(\omega) = \underbrace{\frac{|\bar{\Phi}_{y_1y_2}(\omega)|^2}{\bar{\Phi}_{y_1y_1}(\omega)\bar{\Phi}_{y_2y_2}(\omega)}}_{=|y_{y_1y_2}|^2} \bar{\Phi}_{y_jy_j}(\omega). \tag{7.8}$$

Given the power spectrum of the residual echo, the Wiener filter equation can be solved, i.e. the clean speech power $\Phi_{s_is_i}$ is estimated as $(\Phi_{y_iy_i} - \beta\,\hat{\Phi}_{r_ir_i})$ where β denotes the residual overestimation factor. The Wiener filter transfer function is therefore:

$$H_i(\omega) = \frac{max(\bar{\Phi}_{y_iy_i}(\omega) - \beta\,\Phi_{r_ir_i}(\omega), 0)}{\bar{\Phi}_{y_jy_j}(\omega)}, \quad i \in \{1,2\} \tag{7.9}$$

with $\beta = 0.8$ (for best performance as per Siegwart et al. [2012]).

$H_i(\omega)$ is then multiplied by the corresponding beamformer output Y_i as per Equation 2.18.

7.3.4 Speaker localisation with a superdirective SRP-PHAT

Speaker localisation is the estimation of an acoustic point source given multiple microphones and their positions. This section looks at speaker localisation using the SRP-PHAT as defined above in Section 2.1.4.

Previously (cf. Section 2.1.4) I stated that:

> "For source localisation using the SRP, a simple delay-sum beamformer searches a predefined spatial region looking for peaks in the power spectrum."

The SRP can therefore be calculated as per Equations 2.14 and 2.15 (repeated here for readability) as:

$$P(\boldsymbol{q}) = \int_{-\infty}^{+\infty} |Y(\omega)|^2 \mathrm{d}\omega \tag{7.10}$$

The location estimate is then found from

$$\hat{\boldsymbol{q}}_s = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}) \tag{7.11}$$

One location estimate is calculated for the WSJ data sets while two estimates are required for the MSWSJ data sets. Single and multiple speaker localisation are looked at next.

7.3.4.1 Multiple speaker localisation with superdirective SRP-PHAT

We propose using a superdirective variant of the steered response power with phase transform (SRP-PHAT) method described by DiBiase [2000] and DiBiase et al. [2001] to localise multiple speakers. The original SRP-PHAT is briefly revisited here in order to explain the novel approach (c.f. Section 2.1.4). Our method essentially steers a DSB into each possible direction $\mathbf{a}_1(\phi,\theta)$ and then calculates the total power at the beamformer output as

$$P_1\left\{\boldsymbol{a}_1(\phi,\theta),t\right\} = \int_{-\infty}^{\infty} \|\boldsymbol{w}^H(\omega)\cdot\tilde{\boldsymbol{X}}(\omega,t)\|^2 d\omega \tag{7.12}$$

where

$$\tilde{X}_i(\omega,t) = \frac{X_i(\omega,t)}{|X_i(\omega,t)|}. \tag{7.13}$$

$\tilde{X}_i(\omega,t)$ is a whitened version of $X_i(\omega,t)$ and $\boldsymbol{w}^H(\omega)$ for the DSB can be calculated as per Equations 2.6, 2.7 and 7.1.

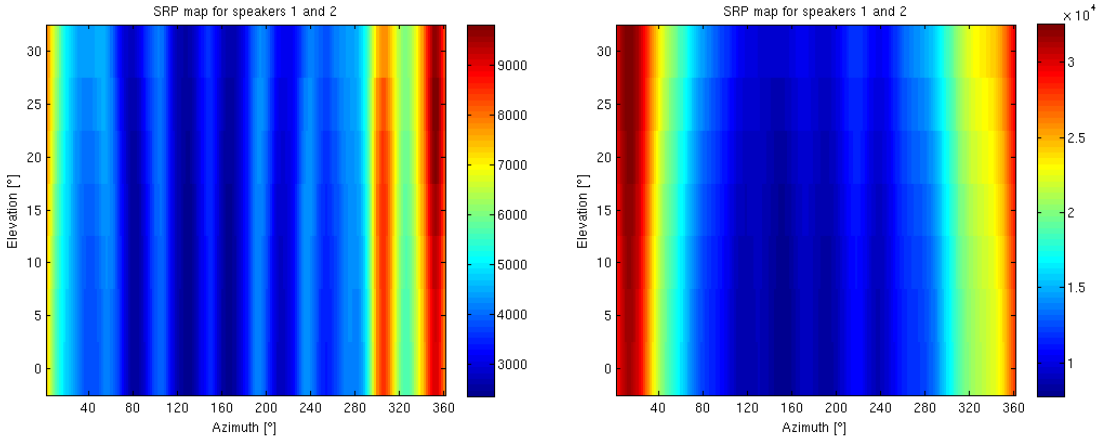
Once the whole range

$$\theta \in [0,2\pi], \quad \phi \in [-\pi/2,\pi/2] \tag{7.14}$$

has been scanned, the speaker is assumed to reside at that location $\mathbf{a}_1(\phi,\theta)$ where $P\{\mathbf{a}_1(\phi,\theta),t\}$ is maximised.

The idea of the newly proposed superdirective SRP is now to simply replace the weight vector $\mathbf{w}(\omega)$ of the DSB by that of a SDB as per Equation 2.10. This should in principle improve the localisation in reverberant environments.

Figure 7.1 (a) shows an SRP for two active speakers speaking with similar energy levels while (b) shows the SRP where the second speaker is hidden by either a sidelobe of the first one or speaks at a much reduced energy level.



(a) Two equal speakers

(b) One dominant speaker

Figure 7.1: SRP map for dual speaker localisation showing two equal vs. one dominant speaker

Unfortunately, scenario (b) is more often found in practice than (a) and our proposed 'multiple speaker localisation method using superdirective SRP' needs to be modified for improved robustness, as described in the next Section 7.3.4.2.

7.3.4.2 Robust multiple speaker localisation

As shown in Figure 7.1 (b), the maximum SRP as per Equation 7.12 usually only produces one active speaker as the second active speaker is suppressed by the first one. We therefore propose a novel robust multiple speaker localisation method.

Once the location of the first (stronger) speaker has been found, we perform a second SRP iteration in which one beamformer w_1 is fixed on the position of the first speaker. A second beamformer w_2 scans all possible directions for the second speaker. During

calculation of the response power $\int |Y_2(\omega,t)|^2 d\omega$ in a particular direction, the effect of the first speaker is cancelled by processing the output

$$Y_2(\omega,t) = \mathbf{w}_2^H(\omega)\tilde{\mathbf{X}}(\omega,t) \tag{7.15}$$

with the binary masking method presented in Section 7.3.2. This effectively restricts the localisation to those time-frequency units which are not used by the first speaker. The SRP of the second speaker is therefore calculated as:

$$P_2\{\mathbf{a}_2(\phi,\theta),t\} = \int_{-\infty}^{\infty} \hat{M}_2(\omega,t) \cdot \|\mathbf{w}_2^H(\omega) \cdot \tilde{\mathbf{X}}(\omega,t)\|^2 d\omega \tag{7.16}$$

with $\mathbf{a}_2(\phi,\theta)$ scanning all possible directions.

In a similar way to Equation 7.3, the mask $\hat{M}_2(\omega,t)$ is set to 1 if the output power $|Y_2(\omega,t)|^2$ of the second (scanning) beamformer exceeds the output power $|Y_1(\omega,t)|^2$ of the first beamformer (which is fixed on the first speaker). The mask is otherwise set to 0.

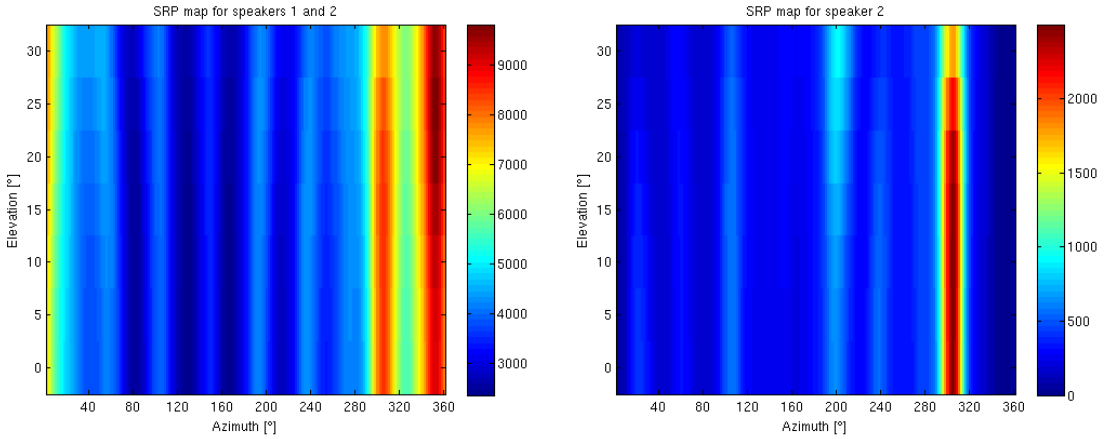
The proposed method is remotely related to cancelling the GCC peaks corresponding to the first speaker, as presented by Oualil et al. [2012].

Figure 7.2 (a), which is identical to Figure 7.1 (a), shows the results from the first SRP map while Figure 7.2 (b) shows the second SRP map after the first speaker has been masked.

Note that for WSJ data sets, the proposed method is being used to search for the position of one speaker, while for the MSWSJ data sets two speakers need to be located. The proposed algorithm will find the number of peaks (and therefore speakers) which it is set to look for, independently of the number of active speakers. Finding the true number of active speakers, particularly on a frame basis, is an open research problem as stated earlier (cf. Section 7.2).

7.4 Experiments

The proposed algorithms were verified on the WSJ and MSWSJ data sets of the 2012_MMA corpus (see Chapter 4 for a detailed description). Two research questions were addressed in the proposed speech separation experiment: (1) taking the recordings from both



(a) First SRP map

(b) Second SRP map - after masking the first speaker

Figure 7.2: Robust speaker localisation using the newly proposed two-pass SRP and masking method

the IMR and the hemi-anechoic chamber, what is the effect of adding noise and reverberation to the interfering speech of the two speakers in a meeting room ($T_{60} = 180$ ms, Steve Renals and Mike Lincoln, personal communication, 2012) and an almost noiseless and reverberation-free environment and (2), using the 20 cm and 4 cm microphone arrays, what is the effect of the reduced diameter of the array on the speaker localisation and speech separation performance? The reduction in the dimension of the microphone array is compensated by increasing the microphone sample rate[1].

In order to verify these algorithms, the speech output quality was measured as WER on the ASR performance. This is a much more practical measure for human understanding of the results than verifying a speech separation system by giving SNR and speech perception figures or the precision of the localisation.

The speech recognition results presented here were produced following the same setup described in Zwyssig et al. [2010] to ensure validity of the experimental data and in order to be able to compare the results. Baseline experiments were also carried out with the MC-WSJ-AV corpus. This corpus was recorded with the eight-channel analogue array with a diameter of 20 cm, the same array as used for the 2012_MMA corpus.

The recordings from the WSJ and MSWSJ data sets were processed as illustrated in Figure 7.3.

[1]Note that the audio signal sample rate was increased by a factor of six from the DMMA.2 to the DMMA.3 while the dimension was only reduced by a factor of five. This is not expected to cause any difficulty and will be analysed in the upcoming experiments.

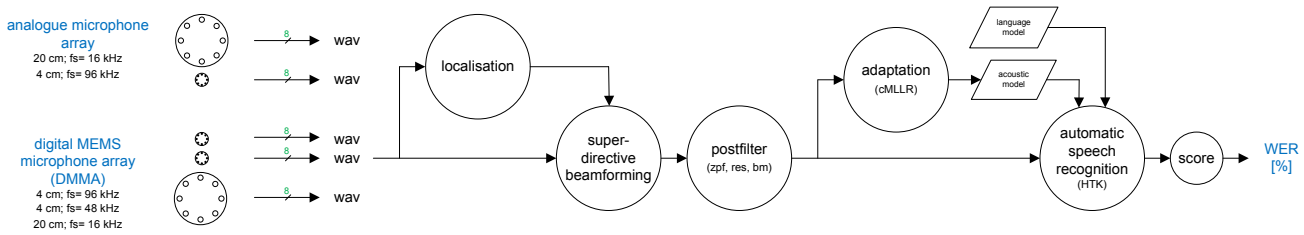


Figure 7.3: Flow diagram for speech separation and ASR experiment

First, sound source localisation was carried out with the newly proposed method using the audio signals from the eight channels. Beamforming and post-filtering was then performed and one or two speakers were extracted from the audio inputs. Speech recognition was carried out on the post-filtered signal and acoustic model adaptation performed using the adaptation recordings. Recognition and scoring were conducted with a context-dependent GMM-HMM system using the HTK toolkit.

Note that no filtering is applied to the audio signal before performing localisation, therefore guaranteeing no modification of the input audio signal phase for best localisation performance.

There is an acoustic mismatch between the WSJCAM0 training data and the microphone array recordings which form the test data. To address this we used the adaptation sentences recorded to carry out a two pass cMLLR adaptation of the model means and variances, similar to my previous experiments (cf. Section 3.2.1 and [Zwyssig et al., 2010]). We adapted the models to the individual channels and to the speakers, pooling the 17 adaptation sentences recorded by each speaker. The recognition experiments were then performed on the 5,000 word (closed vocabulary) sub corpus of WSJCAM0 from the matched array.

Modifications were necessary for the overlapping speaker experiments because the identity and position of the individual speakers were not known. cMLLR adaptation was therefore carried out for a speaker pair and not the individual speakers. This is necessary because the identity of the individual speakers of the MSWSJ data sets is not known a priori unless speaker identification has been carried out.

7.5 Results and discussion

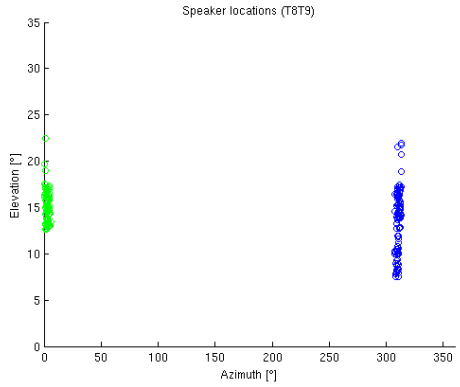
This section presents the results and discussion of the proposed speaker localisation and speech separation algorithms. First, the localisation performance is shown followed by the presentation of the WER of the experiments and a detailed analysis.

Correct localisation is not the prime objective of our proposed speaker localisation algorithm. The main aim is to direct one acoustic beam for single speaker and two acoustic beams for overlapping speakers into the direction of arrival of the sound for best speech recognition performance. It is nevertheless interesting to see how the localisation performs. This is illustrated in Figure 7.4.

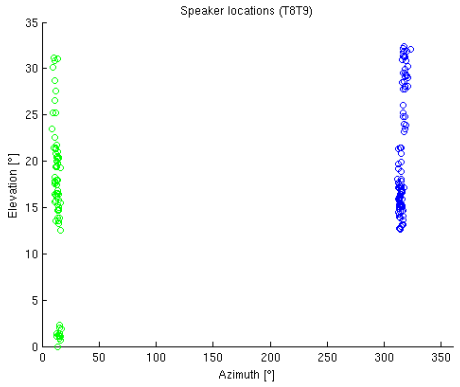
Figure 7.4 shows that the proposed robust speaker localisation is working well, independent of the array geometry or audio sample rate. The localisation accuracy in the azimuth is correct within $\pm 2^{\circ}$ while the elevation accuracy varies much more. The elevation accuracy is not so precise because the microphone arrays used are flat, therefore not allowing good elevation resolution, as shown in Figures 7.1 and 7.2. Figures (a)-(e) show the speaker locations detected for the speakers T8 and T9 from the MSWSJ data set and (f) shows the location for the speaker T9 from the MC-WSJ-AV stat data set. Speaker T9 is reading prompts from 6 different positions, which is clearly visible in the figure.

Good azimuth accuracy is much more important for better acoustic beamforming than elevation accuracy, particularly for beamforming at higher audio frequencies. Looking at Figure 2.1 (cf. Section 2.1) we can see that the acoustic beam is much narrower for higher audio frequencies. A deviation of the localisation in the azimuth will lead to an attenuation and distortion of the audio signal and subsequent degradation on speech separation. Localisation in the elevation is less important as the audio beam is much wider and therefore less sensitive to errors (see Figures 7.1 and 7.2) allowing the speakers to sit or stand without localisation degradation.

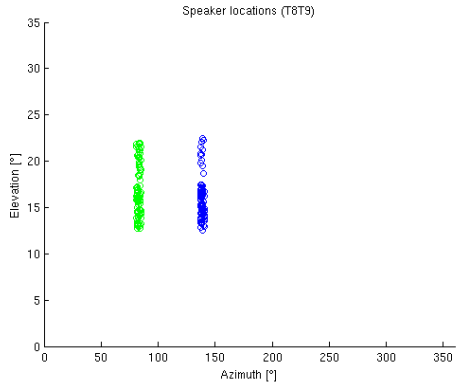
Speaker localisation using SRP-PHAT is a means of acoustic scene analysis which searches for the peaks in the acoustic power map. Multiple sources will appear as multiple peaks where we find a main and possibly several competing sources. The main source can be a person while an overlapping speaker is a competing source. Depending on the acoustic properties (such as reverberation time) of the room where the localisation is carried out, other sound sources will appear. One possible source is



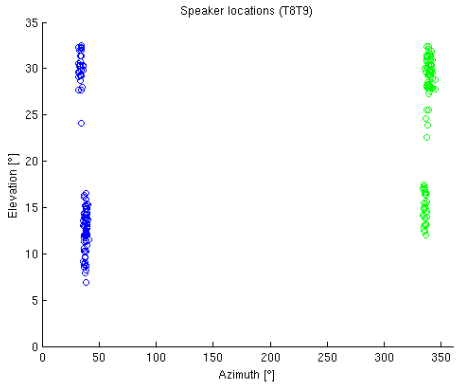
(a) Analogue microphone array (20 cm 16 kHz)



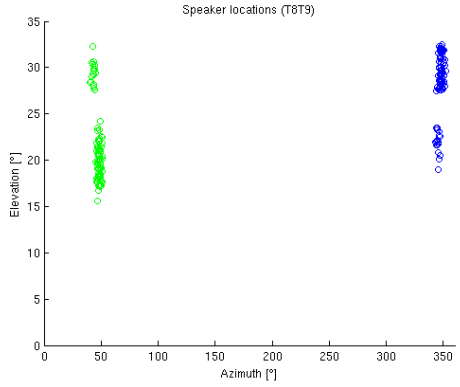
(b) Analogue microphone array (4 cm 96 kHz)



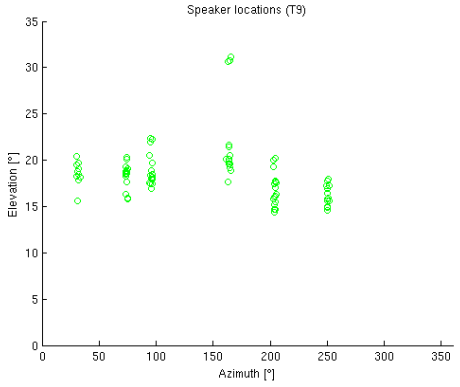
(c) Digital microphone array (20 cm 16 kHz)



(d) Digital microphone array (4 cm 96 kHz)



(e) Digital microphone array (4 cm 48 kHz)



(f) Analogue microphone array (20 cm 16 kHz)

Figure 7.4: Speaker localisation distributions using the proposed robust speaker localisation algorithm for the five different microphone arrays with two speakers (a-e) and for a single moving speaker with one microphone array (f). Green circles show the position of the first speaker, blue circles show the position of the second speaker.

"ghosts" in the SRP map which may be generated by constructive interferences in the acoustic map domain. Determining whether a peak in the SRP map is a real speaker or an interference is an open research problem [Brutti et al., 2010].

We have analysed every recording of single and overlapping speech and in no case did the proposed localisation algorithm fail to detect the correct position of the one or two speakers. The acoustics in the IMR and hemi-anechoic chamber where the recordings for the 2012_MMA corpus were carried is very good, resulting in fault-free localisation.

Speech recognition performance is looked at next.

First, we compared our algorithm with the baseline results achieved on the MC-WSJ-AV corpus, specifically on the single and overlapping tasks (cf. Section 2.5.2).

State-of-the-art speech recognition performance using the single stationary speaker data of the MC-WSJ-AV corpus is 12.2% WER [Kumatani et al., 2012]. Our results on the single speaker data (2012_MMA, WSJ, see Table 7.2), ranging from 13-25%, are in line with those for all five microphone arrays using simple cMLLR adaptation[2].

Word error rates achieved for the overlapping speech recognition task (olap) of the MC-WSJ-AV corpus are presented in Table 7.1.

Table 7.1: Overlapping speaker WER [%] from the ASR experiments on the MC-WSJ-AV corpus

Adaptation	None WER [%]	channel WER [%]	speaker & channel WER [%]
SDB	90.3	67.2	67.2
SDB+ZPF	87.6	63.2	63.2
SDB+RES	81.7	55.3	58.9
SDB+BM	73.8	46.3	48.6

SDB denotes superdirective beamforming on the two speaker locations, SDB+ZPF superdirective beamforming followed by Zelinski postfiltering [Zelinski, 1988], SDB+RES superdirective beamforming followed by residual echo suppression and SDB+BM superdirective beamforming followed by binary masking. Note that [%] WER figures reported in Table 7.1 averaged for nine speaker pairs.

[2]Note that the digital MEMS microphone array DMMA.2 (d = 20 cm, Fs = 16 kHz) is a prototype only and shows increased noise and therefore also increased WER. The issues have been resolved with the new array DMMA.3 (d = 4 cm).

For the overlapping speaker scenario Himawan et al. [2008] achieved a 58% WER for both speakers and 40% for the better speaker. McDonough et al. [2008a] achieved a 39.6% WER using a sophisticated 4-pass ASR system. Our best result of 46.3 % WER for SDB+BM, achieved with simple two-pass MLLR adaptation, outperforms the similar system presented by Himawan et al. [2008] but cannot match the recognition performance of the sophisticated 4-pass ASR system by McDonough et al. [2008a].

Next, the performance of the proposed speech separation algorithms was analysed on the WSJ and MSWSJ data sets of the 2012_MMA corpus. The WERs achieved are presented in Table 7.2.

Our results on the WSJ data sets using superdirective beamforming (SDB) are similar to previous results achieved on the first DMMA.1 [Zwyssig et al., 2010], where I demonstrated that simple cMLLR adaptation to the channel (i.e. microphone array type) can be used to achieve almost identical speech recognition performance. The WERs using SDB only can be improved by a few percent using Zelinski postfiltering (SDB+ZPF). Using speaker and channel adaptation the WERs obtained from the different microphone arrays are very closely matched.

For the multi-speaker MSWSJ speech separation task we achieved a lowest WER of around 35%, again only using simple cMLLR adaptation to the channel. These results were obtained with both residual echo suppression (RES) and binary masking (BM).

The best results were achieved by using SDB and residual echo suppression (SDB+RES) or binary masking (SDB+BM). Residual echo suppression appears to be more effective for analogue microphones, while binary masking works better for the MEMS microphones. Speaker and channel adaptation is not effective for overlapping speech recognition due to the data not being from one, but two speakers. Channel-only adaptation is more effective as there is more adaptation data.

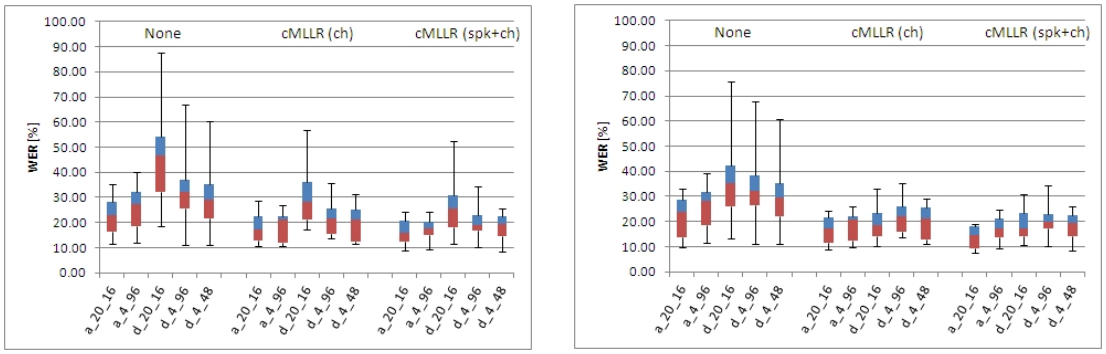
Next, a detailed analysis of the results presented in Table 7.2 is carried out. This is done using boxplots for easier evaluation of the results, where the median is shown as the interface between the brown and red boxes, the lower quartile as a brown box, the upper quartile as a blue box and the minimum and maximum WER as whiskers.

Figures 7.5, 7.6 and 7.7 present a detailed analysis of the WERs for the four algorithms SDB, SDB+ZPF, SDB+RES and SDB+BM on the WSJ and MSWSJ data sets of the 2012_MMA corpus.

Table 7.2: [%] WER results from the ASR experiments on the single (WSJ) and overlapping speaker (MSWSJ) data sets (from the 2012_MMA corpus) in an IMR and hemi-anechoic chamber

	Corpus	**WSJ (IMR)**					**WSJ (hemi-anechoic)**				
	Microphone array	Analogue		Digital			Analogue		Digital		
	diameter [cm]	20	4	20	4	4	20	4	20	4	4
	Fs [kHz]	16	96	16	96	48	16	96	16	96	48
	Adaptation	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]
SDB	None	23.2	26.3	45.3	32.3	29.4	18.0	20.6	37.1	21.1	20.8
	cMLLR (channel)	17.9	18.2	29.7	21.4	20.0	16.4	17.6	26.3	17.9	17.9
	cMLLR (speaker & channel)	16.1	17.3	25.6	19.7	18.2	14.4	15.8	24.9	15.0	15.6
SDB+ZPF	None	21.8	26.3	35.3	33.0	29.6	18.0	20.5	36.1	21.0	20.7
	cMLLR (channel)	16.8	18.1	19.3	21.7	20.0	17.0	16.8	25.9	17.9	18.0
	cMLLR (speaker & channel)	13.9	17.0	18.7	20.1	18.2	14.7	14.9	23.8	14.9	15.6

	Corpus	**MSWSJ (IMR)**					**MSWSJ (hemi-anechoic)**				
SDB	None	93.4	105.0	97.2	108.8	108.6	93.7	104.8	97.8	107.9	104.7
	cMLLR (channel)	66.7	81.5	64.1	80.9	82.1	67.6	79.4	60.0	81.7	80.0
	cMLLR (speaker & channel)	67.7	83.6	63.0	85.8	85.9	67.4	81.4	59.4	83.1	82.3
SDB+ZPF	None	88.2	102.7	90.2	105.4	107.2	90.4	102.9	94.2	106.3	102.8
	cMLLR (channel)	56.2	77.1	43.2	78.7	79.5	64.3	76.7	59.1	78.9	77.8
	cMLLR (speaker & channel)	55.8	80.5	43.5	81.5	83.4	64.5	78.7	58.4	79.6	80.2
SDB+RES	None	65.3	66.2	72.5	66.9	64.9	58.8	65.2	71.8	72.0	63.9
	cMLLR (channel)	35.4	36.3	39.4	31.9	34.1	30.9	37.6	44.5	49.0	37.8
	cMLLR (speaker & channel)	36.1	37.0	40.8	35.0	36.1	32.4	43.1	45.2	50.8	39.1
SDB+BM	None	59.9	63.2	58.4	60.3	60.3	61.9	75.8	66.6	71.8	62.9
	cMLLR (channel)	31.9	35.8	32.7	33.5	33.5	40.3	47.0	42.4	46.2	42.6
	cMLLR (speaker & channel)	34.3	38.7	34.9	35.4	35.2	39.4	48.0	42.8	48.5	44.0



(a) SDB

(b) SDB and Zelinski postfiltering

Figure 7.5: [%] WER of the speech recognition experiments on the IMR WSJ data set (2012_MMA corpus)

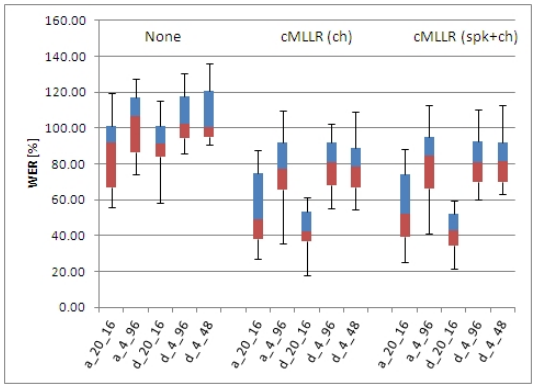
Figure 7.5 compares the WER performance using (a) SDB and (b) SDB+ZPF processing on the WSJ data set. The type of adaptation is indicated as 'None' for no adaptation, 'cMLLR (ch)' for microphone channel adaptation and 'cMLLR (ch+spk)' for channel and speaker adaptation. For 'cMLLR (ch)' adaptation all adaptation sentences for a microphone array type are pooled for the model adaptation while for 'cMLLR (ch+spk)' the adaptation sentences for the individual microphone array and speaker are pooled for the adaptation.

The microphone array type is indicted as <type>_<diameter>_<Fs>. a_20_16 is therefore the analogue microphone array with a diameter of 20 cm, sampled at 16 kHz while d_4_96 is the digital microphone array with a diameter of 4 cm, sampled at 96 kHz, i.e. the DMMA.3.

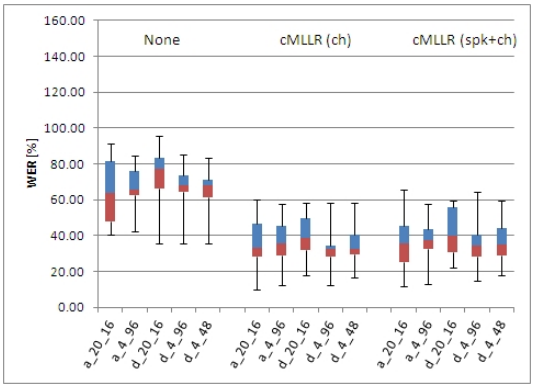
For the WSJ data sets, cMLLR adaptation leads not only to an improved mean and median WER, but also to a significantly lower deviation. This is also the case for the MC-WSJ-AV data set presented in Table 7.1 and the first experiments on the DMMA.1 presented in Zwyssig et al. [2010].

Figure 7.6 (a) and (b) and Figure 7.7 (a) compare the different postfiltering methods on the MSWSJ data set. As stated above, SDB+RES and SDB+BM achieve best WERs on the speech separation task. While the average values of the two schemes are comparable, the variance of the WERs is almost half using binary masking compared to residual echo suppression.

Finally, Figure 7.7 (a) and (b) compare the WERs of the binary masking method (SDB+BM) from a meeting room (IMR) and a hemi-anechoic chamber. Intuitively,

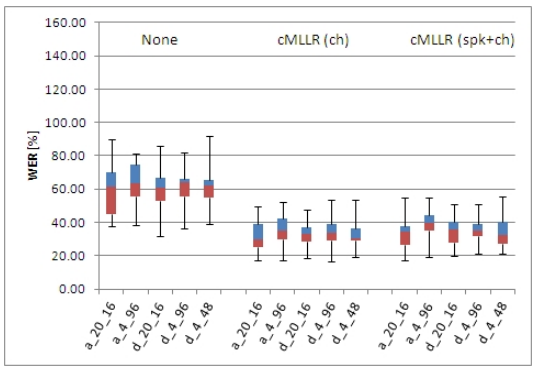


(a) SDB and Zelinski postfiltering

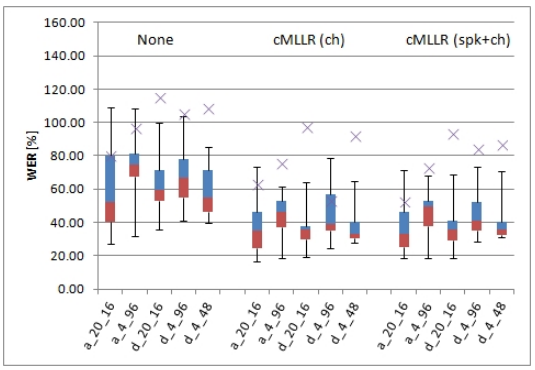


(b) SDB and residual echo suppression

Figure 7.6: [%] WER of the speech separation experiments on the IMR MSWSJ data set (2012_MMA corpus)



(a) SDB and binary masking in IMR



(b) SDB and binary masking in hemi-anechoic chamber

Figure 7.7: [%] WER of the speech separation experiments when performing binary masking on the MSWSJ data (2012_MMA corpus)

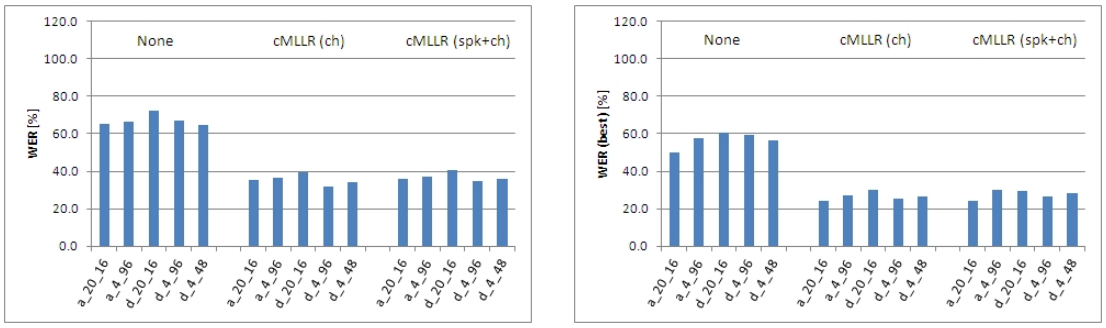
the WER performance in an echo-free environment should be better than in a 'normal' meeting room. Looking at the results presented here, the contrary could be concluded, i.e. that the performance in the IMR is better than in the hemi-anechoic chamber.

Unfortunately, the WER results in the hemi-anechoic chamber are affected by an outlier, shown by crosses in Figure 7.7 (b). If the outlier is removed from the statistics, as shown in Figure 7.7, then it becomes apparent that the type of room neither degrades nor improves the speech separation performance. This could also be explained by the fact that the IMR has been designed for best possible reverberation, as reflected by the very low reverberation time of $T_{60} = 180$ ms (compared to a recommended reverberation time of 0.6 to 1 second for a conference room[3]).

[3] http://www.acoustics.com/conference_room.asp

Note that the results reported here are averages of six speaker pairs. We observed that the WER for one speaker is usually significantly better then for the other one, e.g. the reported WER of 31.5% for the analogue microphone array of 20 cm diameter is a product of the average of 15% WER for the first better speaker and 57.2% WER for the second speaker. This has previously been observed during speech separation experiments on the MC-WSJ-AV corpus [Himawan et al., 2008].

Figure 7.8 shows the comparison of the average of two speakers (a) and the best speaker (b).



(a) Two speakers

(b) Better speaker

Figure 7.8: Comparison of WERs of two speakers vs. the better speaker alone on the IMR MSWSJ data (2012_MMA corpus)

7.6 Summary and conclusions

This chapter looked at overlapping speech separation in different acoustic environments using microphone arrays built from digital MEMS and analogue microphones.

First, a review of the detection and separation of overlapping speech was presented. I showed that detecting overlapping speech is a difficult problem and that approaches based on acoustic models have had limited success so far, both in determining regions of overlapping speech and the number of overlapping speakers.

Next, algorithms for speech separation, i.e. acoustic beamforming and postfiltering were reviewed. Assuming that the regions of overlapping speech and the number of sound sources are known, then acoustic beamforming alone is not sufficient to separate the speakers but postfiltering of the speech signals is also required. This section therefore reviewed acoustic beamforming and presented multiple postfiltering techniques,

i.e. residual echo suppression and binary masking.

After this, a novel algorithm to locate multiple speakers using binary masking was presented.

This novel algorithm and the different postfiltering techniques were verified on the MC-WSJ-AV and 2012_MMA corpora. All experiments were carried out following the setup used for the MC-WSJ-AV corpus in order to guarantee their validity and to compare the results with the state-of-the-art. Our speech separation experiments were carried out by performing ASR and simple constrained MLLR adaptation on the speech output, therefore allowing comparison of the results as WERs, a practical measure for human understanding.

We found that our newly proposed multiple speaker localisation algorithm works very well and that postfiltering is essential for good speech separation. Our experiments show that binary masking and residual echo suppression perform best and that the different methods produce similar results in an IMR and hemi-anechoic chamber.

To summarise, we demonstrated that the 2012_MMA corpus is a valuable extension to the existing MC-WSJ-AV corpus, allowing research in speech separation on natural speech using recordings from five different microphone arrays, including MEMS microphones. Using state-of-the-art speech separation, acoustic beamforming techniques, post-filtering and simple constrained MLLR adaptation, we have obtained baseline WERs in line with the state-of-the-art on the distant single speaker task, and demonstrated improved recognition accuracy on the overlapping speech separation and recognition task.

Chapter 8

Summary, conclusions and outlook

The main objective of this thesis was to study the effects of the increased self-noise (or decreased SNR) of digital MEMS microphones in speech processing applications.

MEMS microphones are in the process of replacing conventional analogue microphones in consumer electronic devices such as smartphones, tablets and mobile computers. These devices use speech processing extensively and it is therefore important to analyse how existing speech processing methods and algorithms can be adapted to audio data recorded with MEMS microphones.

The principal research question I aimed to answer was therefore:

> What are the effects of the increased self-noise of MEMS microphones on state-of-the-art voice activity detection, speaker diarisation, speech recognition and speech separation methods?

This thesis has shown that the decreased SNR does not affect voice activity detection or speaker diarisation and that well established speaker adaptation techniques such as MLLR are sufficient to adapt existing acoustic models (trained on clean speech) to the new acoustic environment and higher noise of speech recorded using MEMS microphones.

No public MEMS microphone speech corpora existed before work on this thesis started. Multiple experiments were therefore designed using four different microphone arrays, two of which were newly developed and manufactured. The newly manufactured arrays are the DMMA.2 and DMMA.3.

These microphone arrays were built using two different microphone types (analogue

capacitive and digital MEMS) and two different microphone array geometries (20 cm and 4 cm). The larger geometry would typically be found in an instrumented meeting room and the smaller one on a mobile device. The audio from the four arrays was sampled at three different sampling rates in order to compensate for the smaller array geometry (16 kHz for 20 cm and 48 kHz/96 kHz for 4 cm), therefore allowing good spatial resolution for superdirective acoustic beamforming.

The microphone arrays with a diameter of 20 cm were used to record a corpus of six meetings from which 72 minutes were annotated for speaker diarisation experiments. Later, after designing the DMMA.3 and its analogue counterpart, these were used to record the 2012_MMA corpus. The 2012_MMA corpus contains four subsets. These are (1) read WSJ sentences from a single stationary speaker (WSJ dataset); (2) read WSJ sentences from two overlapping speakers (MSWSJ dataset); four or six stationary speakers playing Settlers of Catan (Settlers dataset); and (4) four mobile speakers playing Warhammer 40,000 (Wargames dataset). The WSJ, MSWSJ and Settlers datasets were recorded in an instrumented meeting room and a hemi-anechoic chamber. The Wargames dataset was recorded in an instrumented meeting room and the players wore location tracking devices.

The AD_IMR meeting corpora contains confidential information and is only accessible to researchers at the University of Edinburgh. The 2012_MMA corpus was recorded with the aim of allowing unrestricted access to the wider research community and I am working with the LDC to release the recordings.

Using existing meeting recordings provided by the NIST RT challenges and the newly recorded MEMS microphone corpora (AD_IMR and 2012_MMA) I analysed the following speech processing methods:

- voice activity detection (VAD)
- speaker diarisation
- speech recognition
- speech separation

Voice activity detection: using the NIST RT data, an analysis and comparison of well-known, commonly used and novel VAD methods for speaker diarisation and speech recognition was carried out. It transpired that great care is required when designing

a VAD algorithm in order for it to perform better than without VAD, especially in meetings. The results showed that VAD based on Gaussian mixture models (GMM) and multilayer perceptrons (MLP) performs significantly better than methods based on speech activity level or speech periodicity. Overall MLPs perform best on the NIST RT07 and RT09 evaluation data, both in terms of the lowest mean and lowest variance.

Speaker diarisation: The AD_IMR meeting corpus was recorded and annotated in order to carry out speaker diarisation experiments. These experiments compared (1) the analogue microphones and digital MEMS microphones (DMMA.2) and (2) two different acoustic beamformers and TDOA smoothing techniques. It was shown that the digital MEMS microphone array achieves identical diarisation performance to the analogue array, that TDOA smoothing leads to improved diarisation and that superdirective beamforming is of no benefit to diarisation.

Identifying the number of speakers in a multi-party conversation is an important task when carrying out speaker diarisation. This thesis presented a novel algorithm for determining the number of active speakers in a meeting, given audio data recorded using a microphone array of specified dimensions, i.e. known number of microphones and their relative positions, and synchronised audio channels. The proposed algorithm works well on the NIST RT corpora and provides an accurate number of active speakers present in a meeting to downstream speaker diarisation, speaker identification and speech recognition systems. The algorithm is also well suited to online processing.

Speech recognition and **speech separation**: a novel multiple speaker localisation algorithm was presented and its performance measured along with the effect of three different post-filtering schemes on the speech separation task using the WSJ and MSWSJ datasets of the 2012_MMA corpus. Speech recognition and speaker adaption experiments showed that the proposed algorithms performed comparably to state-of-the-art speech recognition algorithms on the single speaker recognition tasks and outperformed state-of-the-art methods on the overlapping speech separation task.

The work presented on speech separation in this thesis is the outcome of collaboration between Friedrich Faubel from Saarland University and myself in which we have shown that our methods port well to speech recorded with digital MEMS and analogue microphones in both a normal meeting room and a hemi-anechoic chamber using simple cMLLR adaptation.

However, this thesis also highlighted that there are some unresolved issues.

VAD and speaker diarisation performance varies greatly over different meetings. The reasons for this are unknown and this thesis has shown that the amount of overlapping speech and minimum, average and maximum speech segment length are not correlated with the VAD and diarisation error rate.

In addition, the short average speech segment length of less than 1.5 s typically found in meeting conversation remains a major problem for speaker diarisation systems.

The DER is also normally lower for state-of-the-art diarisation systems if the number of speakers detected is overestimated and if short speech segments are ignored. Overestimating the number of speakers present in a conversation will cause problems for downstream processes as does ignoring short speech segments. Neither of these issues has been addressed so far.

This thesis has also shown that the problem of overlapping speech detection has not yet been solved.

To summarise, the main contributions of this thesis are:

- a study of VAD algorithms on meeting recordings
- a comparison of speaker diarisation performance using analogue and MEMS microphone arrays and different acoustic beamforming and TDOA smoothing methods
- the development of a novel algorithm to determine the number of active speakers in a meeting recorded using a microphone array of known geometry
- the development of a novel overlapping speaker localisation method (by F. Faubel) and a study of speech separation using analogue and MEMS microphone arrays and different acoustic beamformer post-filtering methods
- a corpus of read and conversational speech using analogue and MEMS microphone arrays and speaker localisation equipment

8.1 Outlook

Nowadays it is considered normal to use speech activated Internet search on mobile devices to e.g. find a particular tourist attraction while exploring a foreign city. Many also control their computers using oral commands and dictate e-mails and letters on a daily basis.

The speech technology required for voice activated search and dictation is gradually finding its way into business life and many people look forward to seeing their meetings recorded and analysed, relieving them of the chore of manually typing meeting minutes and lists of actions and decisions. The armed forces, police, fire brigade and health care professionals in particular require meetings to be recorded for analysis or future reference. These groups will also want to have instant access to records of past events during meetings.

Speech acquisition for speech processing technologies is changing from close talking or headset microphones to distant microphones, as observed by the AMI/AMIDA consortium.

MEMS microphones are omnipresent in todays' mobile devices such as smartphones, tablets and ultrabooks, laptops, headsets, gaming, cameras, televisions, hearing aids, etc. and have found their way into all speech processing domains so it is to be hoped that the analyses, methods and corpora presented here will be useful beyond this thesis.

Audio processing on mobile devices – using MEMS microphones or not – is also moving from mono to stereo or three channels. Smartphones today have at least two microphones, one in the region of the mouth and one or two on the opposite side, allowing advanced methods for e.g. noise suppression and echo cancelling. The research presented in this thesis made use of many, i.e. usually eight microphone channels. Adapting and testing the proposed methods and algorithms to two or three channels would expand their range of application.

Another trend is the incorporation of audio and video information for improved speech processing. Our speaking habits and gestures contain much useful information which could and should be exploited for e.g. better meeting analysis. This requires sophisticated video processing techniques which are known to suffer from harsh environments such as low or degraded video quality, varying lighting conditions or speaker feature detection in the presence of facial hair or (reflections from) glasses.

Nevertheless, video features are essential for conversation analysis and have therefore been included in all corpora presented in this thesis. Audio and video features can easily complement each other for best performance and should ideally be combined in the *instrumented meeting room in a briefcase*. The IMR in a briefcase would be a self-contained box performing high-quality audio and video recording as well as screen capture, and output the data in pre-processed compressed form for downstream processing and machine and human analysis.

I would therefore like to suggest several ways in which the tools and techniques presented in this thesis might be extended for further investigation, such as:

- Speaker localisation performance in the elevation would improve if the microphone array contained a 3D element. Adding an extra microphone to the DMMAs which is not in the horizontal plane could be a requirement for the design of the DMMA.4.
- The DMMAs could be extended to fulfill the requirements of the *IMR in a briefcase* by adding video recording capabilities. Two problems that would need to be addressed for this are: (1) choosing a video camera and a lens which allow panoramic recording and (2) transferring the large amounts of video data to the recording equipment.
- The corpora used for this thesis and for the rich transcription of meetings contain both audio and video data. Audio-visual (AV) processing could be used to increase the performance of VAD, speaker diarisation, speaker localisation, etc.
- The Settlers and Wargames components of the 2012_MMA corpus promise interesting new problems for AV VAD, speaker diarisation, speaker localisation, etc., one of which is overlap detection.
- Our work on speech separation showed good performance but did not look into overlapping speech detection. Overlap detection is an open research topic to which the Settlers and Wargames data sets could contribute.

Appendix A

Consent forms

Participants of the recordings for the 2012_MMA corpus were asked to fill in a form confirming that their recordings can be used for research. The form provided for the AMI/DA meeting recordings was used for the WSJ, MSWSJ and (first) Settlers datasets while a modified form was used for the Wargames recordings. Later participants of Settlers recordings were asked to fill in a newer form. These forms are attached here.

AMI Meetings Corpus Consent Form

Version 1.1, October 8, 2004

Multi-modal research requires large amounts of acoustic recordings of spoken language, along with high quality video, and other multi-modal data recordings. Our goal is to compile such a corpus. This corpus will include a large number of Non-Native-English-Speakers, and will therefore be unique from those compiled at other institutions.

We are asking that when you participate in meetings in our specially equipped recording rooms, you allow us to record the meeting data. You may record multiple meetings, but will only need to complete this form once. Your participation is voluntary and you may stop at any point. The data will initially be used by the AMI Project Partners. It is possible however that at a later stage we will make some or all of the data available to the wider research community, in both transcribed and digitised formats.

No one other than the project staff will have access to any forms you provide to us. However, your name and general demographics may be mentioned in the course of your meeting(s), and you may be recognisable to some people. For this reason it is impossible to completely guarantee anonymity for things you may say. Some general demographics are also typically included in the scientific documentation of corpora and in published findings (e.g. age, dialect information) however, under no circumstances will your name and contact information be divulged as part of the published demographic information.

Please remember that comments you make about people or companies can defame them or invade their privacy, even if you/they are not specifically named but are still recognizable, so it is your responsibility to monitor your speech/behavior. **If you are concerned about any of your data, please advise us immediately and we can arrange for you to review the meeting(s) online**. **On your request we can remove a part of a meeting.**

By signing this form, you agree to allow us to record you and accept responsibility for your conduct in the meeting(s). It is your responsibility to monitor your own speech and actions during the meeting(s), and advise us if any data should be removed.

To indicate that you wish to participate as outlined above, please complete the following:

I, (please print name).....

have read this form, agree to its content and agree to take part in the research on these terms.

Signature: ..... Date: .....

Age: (optional)..... Sex: .....

Are you a native English speaker?

☐ **Yes**, please indicate your country and region .....

☐ **No**, please indicate your native language.....

How many months have you spent living in an English speaking country? .....

Which English speaking country have you lived in?.....

Please list any other language influences (other languages spoken, dialects, etc) .....

.....

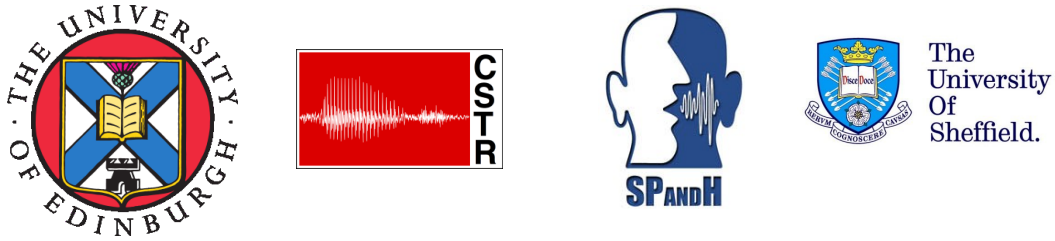
Please provide your email address (or other contact information) so that we can contact you if necessary.

.....

.....

.....

Figure A.1: AMI consent form used for 2012_MMA WSJ, MSWSJ and (first) Settlers data sets



Multi-modal research requires large amounts of acoustic recordings of spoken language, along with high quality video, and other multi-modal data recordings. Our goal is to compile such a corpus. This corpus will include usage of digital MEMS microphones and speaker tracking technology, and will therefore be unique from those compiled at other institutions.

We are asking that when you participate in our specially equipped recording rooms, you allow us to record the data. You may record multiple meetings, but will only need to complete this form once. Your participation is voluntary and you may stop at any point. The data will initially be used by the Universities of Edinburgh, Sheffield and their partners. It is possible however that at a later stage we will make some or all of the data available to the wider research community, in both transcribed and digitised formats.

No one other than the project staff will have access to any forms you provide to us. However, your name and general demographics may be mentioned in the course of the recording, and you may be recognisable to some people. For this reason it is impossible to completely guarantee anonymity for things you may say. Some general demographics are also typically included in the scientific documentation of corpora and in published findings (e.g. age, dialect information). However, under no circumstances will your name and contact information be divulged as part of the published demographic information.

Please remember that comments you make about people or companies can defame them or invade their privacy, even if you/they are not specifically named but are still recognizable, so it is your responsibility to monitor your speech/behaviour.

If you are concerned about any of your data, please advise us immediately and we can arrange for you to review the recording(s) online. On your request we can remove parts.

By signing this form, you agree to allow us to record you and accept responsibility for your conduct. It is your responsibility to monitor your own speech and actions during the recordings(s), and advise us if any data should be removed.

To indicate that you wish to participate as outlined above, please complete the following:

I, (please print name).....
have read this form, agree to its content and agree to take part in the research on these terms.

Signature: ..... Date: .....

Age: (optional)..... Gender: .....

Are you a native English speaker?

- ☐ Yes, please indicate your country and region .....
- ☐ No, please indicate your native language .....
  How many months have you spent living in an in an English speaking country? .....
  Which English speaking country have you lived in?.....

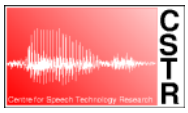
Please list any other language influences (other languages spoken, dialects, etc)
.....

Please provide your email address (or other contact information) so that we can contact you if necessary.

.....

Figure A.2: 2012_MMA *Wargames* gamers consent form

THE UNIVERSITY *of* EDINBURGH
informatics



Speaker Release Form

Contributor's Name: .....

Contributor's Age …….....

Contributor's Address: ……....

……

……

Date of Recording: ……..

The University of Edinburgh, a charitable body registered in Scotland with registration No: SC005336, Old College, South Bridge, Edinburgh EH8 9YL ("the University") will record audio and video of the Contributor taking part in game interactions with other participants for the purpose of speech and interaction data collection ("your Contribution").

1 You hereby agree to the recording of your Contribution and grant to the University all rights including, without limitation, copyright and performers' property rights in your Contribution and all consents necessary to enable the University to make the fullest use of your Contribution worldwide, in perpetuity, in any and all media, whether now known or hereafter developed or discovered, without liability, further payment or acknowledgement to you.

2 You acknowledge that the unscripted interactions being recorded during this data collection may include personal data (as defined in the Data Protection Act 1998) and you hereby grant the University permission to include the unscripted interactions in your Contribution, for the purposes set out in clause 1, above. Your name, age and address details from this form will not be stored as part of your Contribution.

3 In return for providing your Contribution, the University shall pay you a fee of £7 (seven pounds Sterling), which shall be in full and final settlement of any payments that are or may be due to you howsoever arising. No further sums of any nature for any reason shall be due to you for the provision of your service or the exploitation of your Contribution in any media at any time.

4 You acknowledge that your Contribution will be used in connection with the research activities of the Centre for Speech Technology Research at the University's School of Informatics.

5 You acknowledge and agree that the University shall be entitled to use, edit, copy, issue or make available to the public, add to, adapt, exploit or translate your Contribution at

Figure A.3: 2012_MMA *Settlers* gamers consent form (page 1/2)

the University's discretion. In particular, you acknowledge that the University shall be entitled to commercially licence your Contribution and/or works derived from your Contribution and that the University may distribute your Contribution on a worldwide basis, including by means of the internet.

6 In respect of your Contribution, you hereby irrevocably waive in favour of the University, its assignees and licensees, to the fullest extent permitted by law, the benefit of all moral rights and performers' rights arising under the Copyright, Designs and Patents Act 1988 or similar rights arising under the laws of any jurisdiction including, without limitation, the right to be identified as the performer in relation to your Contribution and the right to object to derogatory treatment of your Contribution.

7 You acknowledge that the University does not necessarily undertake to broadcast or otherwise exploit your Contribution.

8 You acknowledge that the University shall not be liable for any loss, damage or injury suffered by you in connection with your participation in the recording of your Contribution other than death or personal injury caused by the University's negligence.

9 The University's rights hereunder may be freely assigned or licensed by the University.

I hereby acknowledge and accept the above provisions of this release form.

Signed ……

Date ……

Recording Supervisor's name.....

Signed.....

Figure A.4: 2012_MMA *Settlers* gamers consent form (page 2/2)

Bibliography

A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. Qualcomm-ICSI-OGI front end archive. `http://www.icsi.berkeley.edu/Speech/papers/qio/`, 2002a. [Online; accessed February 2013].

A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, Jain. P., S. Kajarekar, N. Morgan, and S. Sivadas. Qualcomm-ICSI-OGI features for ASR. In *Seventh International Conference on Spoken Language Processing*, 2002b.

J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2003.

X. Anguera. BeamformIt, the fast and robust acoustic beamformer. `http://www.xavieranguera.com/beamformit/`, 2006. [Online; accessed February 2013].

X. Anguera, C. Wooters, and J. Hernando. Purity algorithms for speaker diarization of meetings data. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2006.

X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007.

X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27(3): 621–633, 2012.

C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1505–1512, 2006.

A. Belouchrani and M.G. Amin. Blind source separation based on time-frequency signal representations. *IEEE Transactions on Signal Processing*, 46(11):2888–2897, 1998.

M. Ben, M. Betser, F. Bimbot, and G. Gravier. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In *Seventh International Conference on Spoken Language Processing*, 2004.

O. Ben-Harush, I. Lapidot, and H. Guterman. Initialization of iterative-based speaker diarization systems for telephone conversations. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):414–425, 2012.

P. Berens. CircStat: a Matlab toolbox for circular statistics. *Journal of Statistical Software*, 31(10), 2009.

P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz, and A. Sixtus. Large vocabulary continuous speech recognition of Broadcast News - The Philips/RWTH approach. *Speech Communication*, 37(1):109–131, 2002.

J. Bitzer and K.U. Simmer. *Microphone Arrays*, chapter Superdirective microphone arrays, pages 19–38. Springer Verlag, 2001.

K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008.

S. Bozonnet, N.W.D. Evans, and C. Fredouille. The LIA-Eurecom RT'09 speaker diarization system: enhancements in speaker modelling and cluster purification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

M.S. Brandstein and H.F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1997.

M. Brauer, A. Dehé, T. Bever, S. Barzen, S. Schmitt, M. Füldner, and R. Aigner.

Silicon microphone based on surface and bulk micromachining. *Journal of Micromechanics and Microengineering*, 11:319, 2001.

M. Brookes. Voicebox: Speech processing toolbox for Matlab, 2011. URL `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. [Online; accessed February 2013].

A. Brutti, M. Omologo, and P. Svaizer. Multiple source localization based on acoustic map de-emphasis. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, 2010.

Cambridge University Engineering Department (CUED). HTK Speech Recognition Toolkit. `http://htk.eng.cam.ac.uk/`, 2012. [Online; accessed February 2013].

J. Chen, J. Benesty, Y. Huang, and S. Doclo. New insights into the noise reduction Wiener filter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (4):1218–1234, 2006.

S.S. Chen and P.S. Gopalakrishnan. Clustering via the Bayesian information criterion with applications in speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1998a.

S.S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998b.

S.W. Chin, K.P. Seng, and L.-M. Ang. Audio-visual speech processing for human computer interaction. *Advances in Robotics and Virtual Reality*, pages 135–165, 2012.

H. Christensen, J. Barker, N. Ma, and P. Green. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Proceedings of Interspeech*. Citeseer, 2010.

C. Cieri, D. Miller, and K. Walker. The Fisher corpus: a resource for the next generations of speech-to-text. In *LREC, The International Conference on Language Resources and Evaluation*, pages 69–71, 2004.

R.K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman, and M.C. Thompson. Measurement of correlation coefficients in reverberant sound fields. *Journal of the Acoustic Society of America*, 27(6):1072–1077, 1955.

M. Cooke, J.R. Hershey, and S.J. Rennie. Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1–15, 2010.

H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1365–1376, 1987.

V. Darre and R. Yussupov. SIRI technology exploration and implementation in healthcare. In *Proceedings of the Eleventh Annual Freshman Conference, University of Pittsburgh*, pages 2182–86, 2011.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

J.H. DiBiase. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Brown University, Providence, Rhode Island 02912, USA, 2000.

J.H. DiBiase, H.F. Silverman, and M.S. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 157–180. Springer, 2001.

J. Dines, J. Vepa, and T. Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of Interspeech*. Citeseer, 2006.

H. Do and H.F. Silverman. A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008.

J. Droppo and A. Acero. *Springer Handbook of Speech Processing*, chapter Environmental Robustness, pages 653–679. Springer Verlag, 2008.

G.W. Elko and J. Meyer. *Springer Handbook of Speech Processing*, chapter Microphone arrays, pages 1021–1041. Springer, 2008.

D.P.W. Ellis and J.C. Liu. Speaker turn segmentation based on between-channel differences. In *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*, pages 112–117. National Institute of Standards and Technology, 2004.

G. Enzner, R. Martin, and P. Vary. Unbiased residual echo power estimation for hands-free telephony. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2002.

ETSI. Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. `http://webapp.etsi.org/WorkProgram/Report_WorkItem.asp?WKI_ID=25817`, 2007. [Online; accessed March 2013].

N. Evans, S. Bozonnet, D. Wang, C. Fredouille, and R. Troncy. A comparative study of bottom-up and top-down approaches to speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):382–392, 2012.

J.G. Fiscus, J. Ajot, and J.S. Garofolo. The rich transcription 2007 meeting recognition evaluation. In *Multimodal Technologies for Perception of Humans*, pages 373–389. Springer, 2008.

C. Fox, Y. Liu, E. Zwyssig, and T. Hain. The Sheffield Wargames Corpus. In *Proceedings of Interspeech*. Citeseer, 2013.

D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd. The voice activity detector for the Pan-European digital cellular mobile telephone service. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1989.

G. Friedland, O. Vinyals, Y. Huang, and C. Muller. Prosodic and other long-term features for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):985–993, 2009.

G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M. Knox, and O. Vinyals. The ICSI RT-09 speaker diarization system. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):371–381, 2012.

S. Furui, L. Deng, M. Gales, H. Ney, and K. Tokuda, editors. *IEEE Signal Processing*, volume 29. IEEE, 2012a.

S. Furui, L. Deng, M. Gales, H. Ney, and K. Tokuda. Fundamental technologies in modern speech recognition. *Signal Processing Magazine, IEEE*, 29(6):16–17, 2012b.

S. Furui, J. Fiscus, G. Friedland, and T. Hain. Introduction to the special section on new frontiers in rich transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):353–355, 2012c.

G. Galatas, G. Potamianos, and F. Makedon. Audio-visual speech recognition incorporating facial depth information captured by the Kinect. In *European Signal Processing Conference (EUSIPCO)*. IEEE, 2012.

Mark JF Gales. *The generation and use of regression class trees for MLLR adaptation*. University of Cambridge, Department of Engineering, 1996.

M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2), 1998.

M.J.F. Gales and Y.Q. Wang. Model-based approaches to handling additive noise in reverberant environments. In *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011.

J.S. Garofolo. *TIMIT: Acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, 1993.

J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan. Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In *Proceedings of Interspeech*. Citeseer, 2010.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1992.

M.D. Good and R.H. Gilkey. Sound localization in noise: The effect of signal-to-noise ratio. *The Journal of the Acoustical Society of America*, 99:1108, 1996.

D. Graff. An overview of Broadcast News corpora. *Speech Communication*, 37(1): 15–26, 2002.

T. Hain, L. Burget, J. Dines, P.N. Garner, A.E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan. The AMIDA 2009 meeting transcription system. In *Proceedings of Interspeech*. Citeseer, 2010.

T. Hain, L. Burget, J. Dines, P. Garner, F. Grézl, A. El Hannani, M. Huijbregts, M. Karafiát, M. Lincoln, and V. Wan. Transcribing meetings with the AMIDA systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):486–498, 2012.

K.J. Han and S.S. Narayanan. A robust stopping criterion for agglomerative hierar-

chical clustering in a speaker diarization system. In *Proceedings of Interspeech*. Citeseer, 2007.

K.J. Han and S.S. Narayanan. Novel inter-cluster distance measure combining GLR and ICR for improved agglomerative hierarchical speaker clustering. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008.

M.H. Hennecke and G.A. Fink. Towards acoustic self-localization of ad hoc smartphone arrays. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, pages 127–132. IEEE, 2011.

H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

I. Himawan, I. McCowan, and M. Lincoln. Microphone array beamforming approach to blind speech separation. In *Machine Learning for Multimodal Interaction*, pages 295–305. Springer, 2008.

T. Hori, C. Hori, Y. Minami, and A. Nakamura. Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1352–1365, 2007.

T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, et al. Real-time meeting recognition and understanding using distant microphones and omni-directional camera. In *Spoken Language Technology Workshop (SLT)*, pages 424–429. IEEE, 2010.

M. Huijbregts. SHoUT speech recognition toolkit. `http://shout-toolkit.sourceforge.net/`, 2006. [Online; accessed February 2013].

M. Huijbregts and D. van Leeuwen. Large scale speaker diarization for long recordings and small collections. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):404–413, 2012.

M. Huijbregts, D.A. van Leeuwen, and C. Wooters. Speaker diarization error analysis using oracle components. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):393–403, 2012.

M.A.H. Huijbregts. *Segmentation, diarization and speech transcription: surprise data*

unraveled. PhD thesis, Centre for Telematics and Information, Technology University of Twente, 2008.

K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada. Probabilistic speaker diarization with bag-of-words representations of speaker angle information. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):447–460, 2012.

ITU-T. ITU G.191 : Software tools for speech and audio coding standardization. `http://www.itu.int/rec/T-REC-G.191-201003-I/en`, 2010. [Online; accessed February 2013].

ITU-T. ITU P.56, Objective Measurement of Active Speech Level. `http://www.itu.int/rec/T-REC-P.56/e`, 2011. [Online; accessed February 2013].

B.H. Juang. Speech recognition in adverse environments. *Computer speech & language*, 5(3):275–294, 1991.

D. Jurafsky and J.H. Martin. *Speech and language processing*, chapter Feature Extraction: MFCC Vectors, pages 329–336. Pearson International Edition, second edition, 2009a.

D. Jurafsky and J.H. Martin. *Speech and language processing*, chapter Decoding: The Viterbi Algorithm, pages 218–220. Pearson International Edition, second edition, 2009b.

B. Kingsbury, T.N. Sainath, and H. Soltau. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In *Proceedings of Interspeech*. Citeseer, 2012.

C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.

D. Kolossa, R.F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J.P. da Silva Neto, and R. Martin. CHiME challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques. In *CHiME Workshop on Machine Listening in Multisource Environments*, 2011.

M. Kotti, V. Moschou, and C. Kotropoulos. Speaker segmentation and clustering. *Signal processing*, 88(5):1091–1124, 2008.

A. Krueger, O. Walter, V. Leutnant, and R. Haeb-Umbach. Bayesian feature enhance-

ment for ASR of noisy reverberant real-world speech. In *Proceedings of Interspeech*. Citeseer, 2012.

K. Kumatani, J. McDonough, D. Klakow, P.N. Garner, and W. Li. Adaptive beamforming with a maximum negentropy criterion. In *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008.

K. Kumatani, J. McDonough, and B. Raj. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Processing Magazine*, 29(6):127–140, 2012.

L. Lee and R.C. Rose. Speaker normalization using efficient frequency warping procedures. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1996.

C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech & Language*, 9:171–185, 1995.

J. Lewis. Understanding microphone sensitivity. *Analog Dialogue*, 46-05, 2012.

M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti. The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2005.

F. Liu, G. Tur, D. Hakkani-Tür, and H. Yu. Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. *Journal of the American Medical Informatics Association*, 18(5): 625–630, 2011.

I. McCowan. Microphone arrays: A tutorial. `http://www.idiap.ch/~mccowan/arrays/index.html`, 2001. [Online; accessed February 2013].

I. McCowan. mdm-tools: Multiple distant microphone toolkit. CSTR internal use only, please contact `http://www.cstr.ed.ac.uk`, 2005.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, 2005.

I. McCowan, M. Lincoln, and I. Himawan. Microphone array shape calibration in diffuse noise fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):666–670, 2008.

J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow. To separate speech: A system for recognizing simultaneous speech. In *Machine Learning for Multimodal Interaction*, pages 283–294. Springer, 2008a.

J. McDonough, M. Wölfel, K. Kumatani, B. Rauch, F. Faubel, and D. Klakow. Distant speech recognition: No black boxes allowed. In *Voice Communication (SprachKommunikation), 2008 ITG Conference on*, pages 1–11. VDE, 2008b.

S. Meignier and T. Merlin. LIUM SpkDiarization: an open source toolkit for diarization. In *CMU Sphinx Workshop for Users and Developers (CMU-SPUD)*, volume 2010, 2010.

S. Meignier, D. Moraru, C. Fredouille, J.F. Bonastre, and L. Besacier. Step-by-step and integrated approaches in Broadcast News speaker diarization. *Computer Speech & Language*, 20(2):303–330, 2006.

M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002.

D. Mostefa. The CHIL audiovisual corpus for lecture and meeting. `http://www.limsi.fr/tlp/chil.html`, 2008. [Online; accessed February 2013].

T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, K. Vesely, P. Matejka, X. Zhu, and N. Mesgarani. Developing a speech activity detection system for the DARPA RATS program. In *Proceedings of Interspeech*. Citeseer, 2012.

T.L. Nwe, H. Sun, B. Ma, and H. Li. Speaker clustering and cluster purification methods for RT07 and RT09 evaluation meeting data. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):461–473, 2012.

M Omologo and P. Svaizer. Acoustic event localization using a crosspower-spectrum phase based technique. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1994.

Y. Oualil, M. Magimai-Doss, F. Faubel, and D. Klakow. Joint detection and locali-

zation of multiple speakers using a probabilistic steered response power. In *SAPA-SCALE Conference*, 2012.

J.M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multi-microphone meetings using only between-channel differences. In *Machine Learning for Multimodal Interaction*, pages 257–264. Springer, 2006.

J.M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computers*, 56(9):1212–1224, 2007.

J.M. Pardo, R. Barra Chicote, R. San-Segundo, R. Cordoba, and B. Martínez-González. Speaker diarization features: The UPM contribution to the RT09 evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2): 426–435, 2012.

D.B. Paul and J.M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

D. Pelegrín-García, B. Smits, J. Brunskog, and C.H. Jeong. Vocal effort with changing talker-to-listener distance in different acoustic environments. *The Journal of the Acoustical Society of America*, 129:1981–1990, 2011.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, et al. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011.

P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett. The DARPA 1000-word resource management database for continuous speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1988.

L. Rabiner and B Juang. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

S. Renals. M4 - Multimodal meeting manager. `http://spandh.dcs.shef.ac.uk/projects/m4/programme.html`, 2004. [Online; accessed February 2013].

S. Renals. Welcome to AMI: Augmented multi-party interaction. `http://www.amiproject.org/`, 2010. [Online; accessed February 2013].

T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1995.

R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278, 2000.

R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Seventh European Conference on Speech Communication and Technology*, 2001.

M.A. Siegler, U. Jain, B. Raj, and R.M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of DARPA Broadcast News Workshop*, page 11, 1997.

C. Siegwart, F. Faubel, and D. Klakow. Improving the separation of concurrent speech through residual echo suppression. In *ITG Symposium Speech Communication*, 2012.

K.U. Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 39–60. Springer, 2001.

M. Sinclair and S. King. Where are the challenges in speaker diarizaton? In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.

J. Sohn and W. Sung. A voice activity detector employing soft decision based noise spectrum adaptation. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1998.

J. Sohn, N.S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

H. Soltau, G. Saon, and B. Kingsbury. The IBM Attila speech recognition toolkit. In *Spoken Language Technology Workshop (SLT)*, pages 97–102. IEEE, 2010.

The International Computer Science Institute. QuickNet - ICSI Open Source Speech Tools, 2010. URL `http://www1.icsi.berkeley.edu/Speech/qn.html`. [Online; accessed February 2013].

D. Vijayasenan. *An information theoretic approach to speaker diarization of meeting recordings*. PhD thesis, École Polytechnique Fédéral de Lausanne, 2010.

R. Vipperla, J.T. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll. Speech overlap detection and attribution using convolutive non-negative sparse coding. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2012.

D. Ward, R.A. Kennedy, and R.C. Williamson. Constant directivity beamforming. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 3–17. Springer, 2001.

M. Wölfel and J. McDonough. *Distant Speech Recognition*. Wiley, 2009.

P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young. Large vocabulary continuous speech recognition using HTK. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1994.

P.C. Woodland, M.J.F. Gales, D. Pye, and S.J. Young. Broadcast News transcription using HTK. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1997.

C. Wooters. The ICSI meeting corpus. `http://www1.icsi.berkeley.edu/Speech/mr/`, 2009. [Online; accessed February 2013].

O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

S.C. Young. *Springer Handbook of Speech Processing*, chapter HMMs and related speech recognition technologies, pages 539–557. Springer Verlag, 2008.

M. Zelenak, C. Segura, J. Luque, and J. Hernando. Simultaneous speech detection with spatial features for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):436–446, 2012.

R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1988.

A. Zolnay, R. Schlüter, and H. Ney. Acoustic feature combination for robust speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2005.

E. Zwyssig. Digital microphone array — Design, implementation and speech recognition experiments. Master's thesis, The University of Edinburgh, 2009.

E. Zwyssig. Speaker diarisation in meetings — Second year progress report. Technical report, The University of Edinburgh, 2011.

E. Zwyssig. Signal processing method and apparatus, 2012. Patent Application GB1203810.5.

E. Zwyssig, M. Lincoln, and S. Renals. A digital microphone array for distant speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

E. Zwyssig, S. Renals, and M. Lincoln. Determining the number of speakers in a meeting using microphone array features. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2012a.

E. Zwyssig, S. Renals, and M. Lincoln. On the effect of SNR and superdirective beamforming in speaker diarisation in meetings. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2012b.

E. Zwyssig, F. Faubel, S. Renals, and M. Lincoln. Recognition of overlapping speech using digital MEMS microphone arrays. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.