Studies on
Microphone Array Processing and Time-Frequency Masking
for Robust Automatic Speech Recognition

Marco Kühne

School of Electrical, Electronic and Computer Engineering

August 7, 2009

This thesis is presented for the degree of
Doctor of Philosophy
of The University of Western Australia
## Contents

Abstract v
Acknowledgments vii
List of publications contained in this thesis and candidate’s contribution ix

### I General introduction

1 Introduction 3
   1.1 Motivation .................................................. 3
   1.2 Scope of research ........................................ 4
   1.3 Research objectives ..................................... 5
   1.4 Thesis organization ..................................... 5
   References .................................................... 7

### 2 Fundamentals of automatic speech recognition

2.1 The ASR problem ........................................... 9
2.2 Automatic speech recognition with hidden Markov models ........................................... 10
   2.2.1 Pre-processing and feature extraction ............... 11
   2.2.2 Acoustic modeling ..................................... 13
   2.2.3 Model evaluation and decoding ....................... 16
   2.2.4 Performance measures ................................. 18
2.3 Noise robust speech recognition ......................... 19
   2.3.1 Speech enhancement .................................. 19
   2.3.2 Robust feature extraction ............................ 20
   2.3.3 Model compensation techniques ...................... 23
2.4 Summary ..................................................... 28
   References .................................................... 29
3 Review of multichannel source separation techniques

3.1 Introduction

3.2 Methods

3.2.1 Beamforming

3.2.2 Independent component analysis

3.2.3 Sparse source separation

3.2.4 Auditory model inspired processing

3.3 Summary

References

II General discussion

4 Discussion and conclusions

4.1 Major findings and contributions

4.2 Limitations

4.3 Suggestions for future research

References

III Publications

1 Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition

2 Robust Source Localization in Reverberant Environments Based on Weighted Fuzzy Clustering

3 A Novel Fuzzy Clustering Algorithm using Observation Weighting and Context Information for Reverberant Blind Speech Separation

4 A Novel Evidence Model for Missing Data Speech Recognition with Applications in Reverberant Multisource Environments
In order to deploy automatic speech recognition technology in real world scenarios it is necessary to handle cocktail-party-like environments with multiple speech and noise sources. Despite several decades of research the noise robustness of state-of-the-art speech recognizers still falls short in comparison with human capabilities. This dissertation focuses on the development of computational models for automatic speech separation and recognition in multi-talker environments. Several aspects of cocktail-party processing are studied, ranging from source localization over source separation to speech recognition. The approach considered aims for a closer integration of microphone array processing and missing feature techniques for noise robust speech recognition. Front- and back-end processing are linked together through the consistent application of time-frequency masking for both source separation and speech recognition. The use of cluster algorithms for automatically estimating these masks on the basis of spatial localization cues is investigated for anechoic and reverberant data. The incorporation of spatial observation weights and time-frequency context information is proposed as a means to increase the localization and segmentation performance of standard fuzzy cluster algorithms, particularly in echoic conditions. While the former helps to improve the localization accuracy by ignoring noisy observations the latter smoothes the fuzzy cluster membership levels by exploiting the high correlation of neighboring mask points. The resulting robust fuzzy cluster algorithm is integrated into a source separation system which combines the advantages of time-frequency masking with the separation capabilities of adaptive beamforming. The thesis also investigates a novel evidence model in the form of the bounded-Gauss-Uniform mixture probability density function for missing data speech recognition. In a number of simulated cocktail-party scenarios it is observed that this new evidence model offers superior recognition accuracy over existing evidence functions and a related binaural segregation model.
List of publications contained in this thesis and candidate’s contribution


The candidate was responsible for the original idea, the software implementation and conducted the reported experiments. The candidate also wrote the manuscript, which was commented on by the co-authors.


The candidate was responsible for the original idea, the design and implementation of all described experiments. The candidate wrote the manuscript, which was commented on by the co-authors.


The candidate was responsible for the original idea, the design and implementation of all conducted experiments. The article was also written by the candidate and was then further commented on by the co-authors.

**Article 4:** M. Kühne, R. Togneri and S. Nordholm. A Novel Evidence Model for Missing Data Speech Recognition with Applications in Reverberant Multisource Environments. *Submitted for publication to IEEE Transactions on Audio, Speech, and Language Processing.*

The candidate was responsible for the original idea, the design and implementation of all conducted experiments. The article was written by the candidate and was then further commented on by the co-authors.
Part I

General introduction
Chapter 1

Introduction

1.1 Motivation

Automatic processing of human speech by machines has been a dream of mankind for centuries. Because speech is the most natural means by which humans communicate, research in automatic speech recognition (ASR) technology has attracted worldwide interest over the last couple of decades. Dramatic advances in computer hardware paired with steady progress in the field of computational natural language processing have made it possible to move the development of speech technology beyond simple laboratory prototype systems. Today, speech-driven interactive communication devices are commercially available, albeit for strongly domain-limited tasks and services, such as name dialing, dictation, telephone number queries or bus timetable retrieval.

However, it is fair to say that the ASR problem is far from being solved. Despite the worldwide efforts of numerous research groups, state-of-the-art ASR technology still performs far from human capabilities. Most ASR systems are trained on clean, anechoic speech and tend to produce errors when confronted with sound mixtures of speech and noise. According to the study of Lippmann [1997], ASR error rates are often more than an order of magnitude higher than those of human listeners in both quiet and noisy environments. For example, while speech recognizers achieved error rates of 40% for spontaneous speech in quiet environments and 23% for read sentences recorded in noise, human error rates remained well below 5% under similar conditions. Lippmann also points out that the inability of modern machine recognizers to distinguish between individual voices and non-speech sounds represents a fundamental weakness that prevents these systems to perform in real-life listening situations.

On the other hand, people with normal hearing possess the ability to focus on and segregate a particular voice or sound from a mixture of voices and background noise. This perceptual phenomenon is known as the "cocktail-party effect" [Cherry, 1953] and has been widely studied in the field of hearing research [Bee and Micheyl, 2008; Bronkhorst, 2000; Culling and Summerfield, 1995; Drullman and Bronkhorst, 2004; Yost, 1997, 2000]. Replicating this remarkable ability of the human auditory system with computational models has proved to be extremely difficult. To date, no technical system has been developed that is capable of human-like performance in environments that include multiple
speakers, room reverberation, and high levels of non-stationary noise sources. This may be attributed to the high complexity and interdisciplinary nature of the problem requiring knowledge from the fields of psychoacoustics, signal processing and pattern recognition. Not surprisingly, the problem has been studied intensively by a number of research groups worldwide. From a neuroscience perspective the cocktail-party effect has been of great interest [Feng and Ratnam, 2000; Moore, 2003; Yost, 2000] because it provides us with a better understanding of the human auditory system and the underlying neurobiological mechanisms related to the cognitive processes of speech perception and speech production. While computational solutions vary in their degree of physiological plausibility several promising lines of research have emerged that offer practical algorithms capable of cocktail-party processing [see Divenyi, 2005, for a detailed discussion]. This includes blind source separation techniques from the field of statistical signal processing [Bell and Sejnowski, 1995; Comon, 1994], spatial filtering approaches [van Veen and Buckley, 1988] or biologically motivated solutions from the field of computational auditory scene analysis (CASA) [Bregman, 1990; Cooke, 1993; Ellis, 1996].

Regardless of the particular approach, it is clear that enabling ASR technology to perform in cocktail-party scenarios would have far-reaching consequences in many important areas, such as hearing aid design, automatic transcription of broadcast news and audio information retrieval. It would not only revolutionize the way people interact with computers, cars and other electronic devices but could also improve the quality of life for persons with cochlear hearing impairments. It is this potential for widespread practical applications which provides the motivation for this thesis.

1.2 Scope of research

This dissertation focuses on the development of computational models for automatic speech separation and recognition in multi-speaker environments. The scope of the research is primarily confined to cocktail-party scenarios with up to seven concurrent talkers in anechoic environments and two-source configurations in reverberant enclosures. The focus is restricted to static cocktail-party scenarios with fixed but unknown speaker positions. These locations need to be estimated by the separation algorithm when necessary. Although it is recognized that moving objects are an important aspect of real-life acoustic scenes, the case of moving speakers lies beyond the scope of this thesis. Regarding the kind of source separation technique, consideration will be limited to multi-channel approaches that are based on microphone array processing. In order to acknowledge the physical design constraints that occur in practice, the thesis only considers linear microphone arrays with a maximum of six sensor elements and an aperture length comparable to the human head.
Regarding ASR performance assessment, the recognition tasks are restricted to low vocabulary problems, such as connected digit recognition. For this reason, the work will focus entirely on improving the robustness at the acoustic level enabling an easier and more transparent analysis of the recognition outcomes when compared with large vocabulary tasks. This is also justified by the study of Lippmann [1997], who concluded that human listeners seem to perform more robust low-level acoustic phonetic modeling than machines when confronted with short segments of speech containing only little contextual information.

Whilst for practical applications the question of real-time processing is of paramount importance, the thesis imposes no restrictions on the computational complexity of the algorithms and all developed techniques operate via batch-processing.

### 1.3 Research objectives

The primary goal of this research is to investigate novel signal processing and pattern recognition techniques for improving the performance of ASR systems under reverberant multi-source conditions. More formally, the thesis aims at achieving the following objectives:

- To evaluate existing approaches and available literature for the fields of multi-channel acoustic source separation and robust automatic speech recognition.
- To design and implement a source separation system capable of locating and separating multiple speech sources under anechoic and moderate reverberant conditions.
- To develop new strategies for a more effective exploitation of the source separation results in the speech decoder with the help of statistical model compensation techniques.
- To evaluate the overall system consisting of source separation stage and speech recognition engine under controlled laboratory conditions with additive and convolutive noise distortions.
- To discuss and compare the outcomes of the recognition experiments with related systems proposed in the literature.

### 1.4 Thesis organization

This thesis is divided into three parts.
Part I serves as a general introduction, in which we state the research problem as well as the specific aims and overall objectives of the dissertation. Chapter 2 presents important background information regarding the field of automatic speech recognition and surveys available techniques for improving ASR robustness under noise. Chapter 3 introduces the reader to the area of multi-channel source separation and discusses its main lines of research along with their benefits and shortcomings. Particular attention is given to the applicability of these methods as speech recognition front-ends.

Part II of this dissertation consists of a general discussion, which summarizes our major findings and contributions and places them within the context of previous work. The reader is advised to consult Part III (publications) of the thesis prior to reading this section. The discussion further identifies the limitations of the work and highlights future research directions.

Part III of this dissertation comprises of four selected articles written during the course of this research project. While some interdependencies exist each article is self contained and focuses on a separate research problem. The following paragraphs provide brief summaries of these articles and demonstrate how they are related to each other.

Article 1 starts with the adaptation of a two-channel blind source separation technique as front-end for missing data speech recognition. Attenuation and delay parameters extracted from an omnidirectional microphone pair are used for demixing multiple sources via time-frequency masking. Rather than using these masks for waveform resynthesis they are passed to a missing data decoder that performs marginalization of unreliable declared features. It is found that speech recognition accuracy can be retained at high levels for up to three concurrent talkers in an anechoic environment.

Article 2 is concerned with the problem of sound source localization in reverberant enclosures, which is an important prerequisite for many spatial signal processing algorithms. We extend the two-sensor system, used in Article 1, to a linear microphone array and replace the histogram-based localization step with a weighted fuzzy c-means cluster algorithm (wFCM). In order to increase the algorithm’s robustness against sound reflections observation weights are incorporated to emphasize reliable cues over unreliable ones. In simulated localization experiments, the proposed wFCM algorithm successfully located two speech sources for a range of angular separations and various room reverberation times.

Article 3 investigates the use of spatial filtering techniques in combination with time-frequency masking. Borrowing upon recent ideas from the blind source separation community, adaptive filters of a linearly constrained minimum variance (LCMV) beamformer are designed for an improved interference suppression. The LCMV weights are estimated blindly through a novel fuzzy clustering algorithm (wCFCM). This algorithm further extends the weighted fuzzy clustering approach, developed in Article 2, by incorporating
local correlations between neighboring time-frequency points into the mask estimation process. Speech localization and separation experiments demonstrate the superiority of the proposed method over standard fuzzy clustering, both in terms of localization accuracy as well as speech separation performance.

Article 4 presents a novel evidence model for missing data speech recognition in the form of the bounded-Gauss-Uniform mixture probability density function. Based on the source separation system developed in Article 3, we discuss a simple scheme for estimating the mixture density parameters of this new type of evidence model. The overall system is tested with an extended spectral feature set for its applicability in multi-source reverberant environments. Through a number of experiments it is observed that the new evidence model offers superior recognition accuracy over previously proposed evidence functions and a related binaural segregation model.

References


1. Introduction


Chapter 2

Fundamentals of automatic speech recognition

Automatic speech recognition (ASR) is a computerized process which aims to transcribe an unknown incoming speech signal through a sequence of linguistic units. In this chapter, we present the necessary background information for understanding this process from an engineering perspective. The first part of the chapter serves to introduce the fundamental concepts used in state-of-the-art ASR systems. After formulating the ASR task as a statistical optimization problem, we take a closer look at the architecture of a modern HMM-based speech recognition system. We begin with the feature extraction stage and describe the procedure for converting a speech waveform into a sequence of feature vectors. Next, we discuss how the speech feature distributions can be modeled with HMMs and what algorithms are required for model training and testing. The second part of this chapter is concerned with the problem of robust speech recognition in noise. We briefly discuss several techniques from the fields of speech enhancement, robust feature extraction and model compensation.

2.1 The ASR problem

Speech recognition by machines is often formulated as a pattern recognition problem [Rabiner and Juang, 1993], in which a simultaneous segmentation and classification of the incoming acoustic waveform into linguistic units is performed. To illustrate this process, Fig. 2.1 presents an overview of a typical ASR system. The first stage, called feature extraction, produces meaningful patterns from the speech waveforms. A training phase is required in which the system learns the acoustic characteristics of speech sounds from training data. This process results in a set of acoustic models which, in combination with a dictionary and a language model, are used to segment and classify the input speech waveforms into linguistic units in the recognition stage.

More formally, given an observation sequence $O = (o_1, o_2, \ldots, o_T)$, the task of the ASR system is to determine the most likely word sequence $\hat{W} = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_n)$ among the set of all possible word sequences $W$. For this problem, the optimal statistical classification strategy [Duda et al., 2001] is given by

$$\hat{W} = \arg \max_{W \in \mathcal{W}} P(W|O),$$

(2.1)
where \( W = (w_1, w_2, \ldots, w_n) \) denotes a possible word sequence hypothesized by the speech recognizer. This expression is not computable directly in practical implementations. Instead, Bayes’ rule is applied to rewrite (2.1) as follows

\[
\hat{W} = \arg\max_{W \in W} \frac{P(O|W)P(W)}{P(O)},
\]

(2.2)

where \( P(O|W) \) is called the acoustic score, \( P(W) \) is the language score and \( P(O) \) denotes the prior probability of the occurrence of feature vector sequence \( O \). The term \( P(O) \) can be neglected for the maximization in (2.2) because it is independent from \( W \), resulting in

\[
\hat{W} = \arg\max_{W \in W} P(O|W)P(W).
\]

(2.3)

Another common assumption is that the language model \( P(W) \) is independent from the mismatch between training and testing environment. Therefore, the search for the most likely word sequence is in fact governed solely by the acoustic matching, e.g.,

\[
\hat{W} = \arg\max_{W \in W} P(O|W).
\]

(2.4)

This assumption is not critical for the connected digit tasks investigated in this study, but allows for a faster and easier analysis of the recognition outcomes. In the following section, we describe the statistical framework that is used for modeling the acoustic score \( P(O|W) \) in virtually all modern ASR systems.

## 2.2 Automatic speech recognition with hidden Markov models

This section presents an overview over current state-of-the-art methods in ASR systems that are based on hidden Markov Models (HMMs). We start with the feature extraction
process and describe the procedure of encoding a speech waveform into a sequence of feature vectors. The section then continues with a short review of the definition and basic properties of HMMs and explains common methods for model training, evaluation and decoding. For a more in-depth discussion of HMMs in speech recognition applications, the interested reader is referred to the relevant publications [Huang et al., 1990; Rabiner, 1989; Rabiner and Juang, 1993; Young et al., 2006].

### 2.2.1 Pre-processing and feature extraction

All modern speech recognition systems perform some sort of feature extraction before attempting classification of the incoming speech sound waves. Rather than using the incoming speech waveforms directly, further signal processing steps are usually required to eliminate unwanted side information. Depending on the recognition task, this could include the removal of phonetic identity, background noise and transmission channel effects as well as speaker characteristics or emotions. Ultimately, the goal of the feature extraction is to transform the signal waveform into a data representation appropriate for statistical modeling. Ideally, the extracted features should encode all the relevant linguistic information of the speech signal while rejecting non-discriminatory elements that are unimportant for recognition purposes.

Feature extraction is an important component of ASR systems because the recognition performance ultimately depends to a great extent on the quality of the extracted features. Information lost at this stage can usually not be recovered in later processing stages. Consequently, since the early 1980's a vast amount of research has been conducted in pursuit of feature representations that best satisfy the above mentioned criteria. Within the HMM paradigm, short-time spectral analysis techniques based on filterbank analysis or linear predictive coding (LPC) have been most successful.

In the following, we concentrate on the description of mel-frequency cepstral coefficients (MFCCs), which have become the de-facto standard feature representation in modern ASR. The following description is based on the Hidden Markov Toolkit (HTK) [Young et al., 2006] filterbank implementation, which has been used throughout this thesis for speech parameterization. The entire procedure for encoding an incoming speech waveform into MFCCs is summarized in Figure 2.2.

### Framing

Although the statistical properties of speech signals change with time, they are considered to be short-time stationary with acoustic events occurring on a time scale of 5-200 ms. In order to comply with this stationarity assumption, the speech waveforms are split into shorter segments, called frames. Typically, frame lengths vary between 20-40 ms. The dis-
2. Fundamentals of automatic speech recognition

![Diagram of the acoustic feature extraction process](image)

**Figure 2.2:** Illustration of the acoustic feature extraction process for converting a waveform into a sequence of frame-synchronous MFCC feature vectors suitable for processing with HMMs.

The distance between consecutive segments is called frame shift, and is usually set to 10 ms. This choice is motivated by the movements of the human speech articulators, which change their positions at a rate of approximately 100 Hz.

**Pre-emphasis**

In order to compensate for the negative spectral slope of approximately 20 dB per decade, a pre-emphasis filter is often applied to each frame prior to spectral analysis. In HTK, this feature can be enabled by specifying a pre-emphasis coefficient, which should be in the range \([0, 1)\). For all experiments reported in this thesis, the pre-emphasis coefficient was set to 0.97.

**Windowing**

The pre-emphasized samples within each frame are then multiplied by a Hamming window function in order to attenuate discontinuities at the frame edges.

**Filterbank analysis**

Psychophysical studies have demonstrated that the human ear resolves frequencies non-linearly across the audio spectrum [Zwicker and Fastl, 1999]. In HTK, a simple Fourier transform based filterbank is implemented, which approximates the desired non-linear frequency resolution of the human ear. First, each window of speech data is analyzed for its spectral content using a Fourier transform. Then, the phase information is discarded by taking the magnitude or squared magnitude of each Fourier transform coefficient. The magnitude spectrum is then passed through a triangular filterbank in order to convert the linear frequency axis to the non-linear mel-frequency scale. The bank of overlapping filters is linearly spaced on the mel-frequency axis. The number of filters may vary between 13 and 24 depending on the bandwidth of the input signals [Davis...
and Mermelstein, 1980]. In order to simulate the non-linear loudness perception of the human auditory system [Zwicker and Fastl, 1999], the natural logarithm is applied to the output of each triangular filter.

**Orthogonalization**

The log-mel-frequency spectrum is then converted to the cepstral domain by applying the discrete cosine transform (DCT). This serves mainly to decorrelate the elements of the feature vector, which is highly desirable when feature distributions are modeled with a statistical framework. Decorrelated features allow the use of diagonal covariance matrices and require less training data due to the reduced number of parameters. The cepstral transformation also allows to separate the vocal tract from the excitation information by retaining only the lower 10-13 cepstral coefficients.

**Temporal information**

It has been shown that the temporal evolution of spectral/cepstral characteristics is of great importance for improving HMM-ASR performance in practice [Furui, 1986]. Typically, first and second order derivatives of the static features are calculated by means of regression filters. These dynamic features, also referred to as delta and acceleration coefficients [Young et al., 2006], are then appended to the static coefficients forming a composite feature vector.

### 2.2.2 Acoustic modeling

In the context of probabilistic speech recognition, acoustic modeling denotes the process of building statistical representations for the speech sounds of natural languages. Acoustic models play an essential part in the computation of the likelihoods $P(O|W)$ in Equation (2.4). The use of HMMs as acoustic models has been the most dominant paradigm in speech technology over the last decades.

An HMM, named after the Russian mathematician Andrei Andrejewitsch Markov, is a stochastic finite state automaton defined by two random processes. The first process corresponds to a Markov chain, in which the transitions among the states are governed by a set of probabilities, called transition probabilities. The states of this Markov chain are not directly observable, e.g., they are hidden. Instead, a second random process generates output symbols (observations) at each time instant according to a state-dependent probability distribution.
2. Fundamentals of automatic speech recognition

In order to specify an HMM $\Lambda = \{A, B, \Pi\}$ completely, the following set of parameters is required:

- The number of states in the hidden Markov chain, $N$.

- A state transition probability distribution $A = \{a_{ij}\}$ describing the probability associated with moving from state $i$ to state $j$ with $1 \leq i, j \leq N$.

- A set of output probability distributions $B = \{b_i(o)\}$, $1 \leq i, j \leq N$, where $b_i(o)$ is the probability of observation $o$ while in state $i$.

- A set of initial state distributions $\Pi = \{\pi_i\}$, where $\pi_i$ is the probability of being in state $i$ at time $t = 1$.

Commonly, HMMs are categorized into different model types, depending on the choice of the output probability distributions. For example, all models that define $b_i(o)$ on a discrete probability space are referred to as discrete HMMs. Note that the number of observation symbols in the alphabet is finite for this model type. In contrast, continuous density HMMs are defined on continuous probability spaces leading to an infinite set of output symbols. In speech recognition, the most widely used state output probability density function is the Gaussian mixture model (GMM). This is motivated by the fact that Gaussian mixture PDFs have some nice mathematical properties which can be exploited to alleviate some of the added computational costs associated with using continuous distributions. To lessen the computational burden further, other HMM types, such as semi-continuous HMMs [Huang and Jack, 1988], have been proposed together with tied HMM models [Bellegarda and Nahamoo, 1990] that share some or all parameters of the Gaussian distributions across states. In this thesis, we have used continuous density HMMs, where each state has its own output distribution in form of a Gaussian mixture model. An example for a simple left-to-right HMM in HTK is shown in Figure 2.3.

The popularity of HMMs stems from their ability to model complex time-varying signals, such as speech, through a mathematically appealing framework. Three simplifying assumptions are usually made that guarantee mathematical and computational tractability. Firstly, the Markov assumption implies that the next state is dependent only upon a limited number of past states. For a first order HMM, which is the most common type, this can be formulated as

$$a_{ij} = P(q_{t+1} = j|q_t = i). \quad (2.5)$$

Secondly, the stationary assumption implies that the state transition probabilities are independent of time, e.g.,

$$P(q_{t+1} = j|q_t = i) = P(q_{t+2} = j|q_t = i), \quad (2.6)$$
2.2. Automatic speech recognition with hidden Markov models

Figure 2.3: The Markov generation model for a 5-state left-to-right HMM. The model moves through the state sequence $Q = 1, 2, 3, 4, 5$ in order to generate the observations $o_1$ to $o_5$. Note that the initial and final states are non-emitting to facilitate the construction of composite models [Young et al., 2006].

for any $t_1$ and $t_2$. And lastly, the output independence assumption implies that for an HMM $\Lambda$ the current observation is statistically independent of previous observations, such that

$$p(O|q_1, q_2, \ldots, q_T, \Lambda) = \prod_{t=1}^{T} p(o_t|q_t, \Lambda).$$

(2.7)

By using these assumptions, HMMs can model the spectro-temporal characteristic of speech signals in a rather elegant way. The signals are first split up into smaller segments which are supposed to be piecewise stationary. These segments are then analyzed for their spectral composition and form a sequence of feature vectors $O = (o_1, \ldots, o_T)$ (see Section 2.2.1). The temporal evolution of this observation sequence is modeled by the hidden Markov chain, mapping each of the observed feature vectors $o_t$ to one of the HMM states. During this mapping process, the spectral characteristics are taken into consideration by assigning the observations to their most likely states according to the state output distribution functions.

In this way, a natural language may be represented by modeling each word with a unique HMM. Sequences of words can be obtained by simply concatenating the HMMs of the constituent words of the sequence. However, more commonly, sub-word models based on phones are utilized as the basic HMM unit. In particular, for large vocabulary recognition, vast HMM networks are constructed through the concatenation of individual phone models. For the relatively simple connected digit recognition tasks investigated in this thesis, we resort to the whole-word model approach, where each individual digit is represented by its own HMM.

Before speech recognition experiments can be conducted the HMMs need to be created during a supervised learning phase, called model training. Training an HMM usually involves estimating the model parameters with the help of a phonetically transcribed
speech database. The algorithm of choice for maximum likelihood model training is the so-called Baum-Welch re-estimation procedure [Baum et al., 1970]. Starting from an initial guess, the Baum-Welch algorithm iteratively improves the HMM parameters by maximizing the likelihood of the training data until a convergence criterion is met. While other parameter estimation methods exist, such as discriminative model training [Huang and Soong, 1989; Juang and Katagiri, 1992], all digit HMMs used in this thesis have been trained using the maximum likelihood approach.

2.2.3 Model evaluation and decoding

The problem of model evaluation is concerned with answering the question of how well a particular model explains a given observation sequence. A comparison of different HMMs would then allow us to choose the most appropriate model from a set.

For a HMM with known parameters, $\lambda = \{A, B, \Pi\}$, the probability of generating an observation sequence $O$ for a particular state sequence $Q = q_1, q_2, \ldots, q_T$ is given by

$$p(O|Q, \Lambda) = \pi_{q_1} \prod_{t=1}^{T} a_{q_tq_{t+1}} b_{q_t}(o_t).$$  \hspace{1cm} (2.8)

Because the true state sequence is unknown, we have to sum (2.8) over all possible state sequences to calculate the probability of the observations as

$$p(O|\Lambda) = \sum_{Q} \left( \pi_{q_1} \prod_{t=1}^{T} a_{q_tq_{t+1}} b_{q_t}(o_t) \right).$$  \hspace{1cm} (2.9)

The direct evaluation of (2.9), which is exponential in $T$, is computationally too expensive for most practical systems. In practice, most HMM speech recognizers perform model evaluation using an approximation of (2.9), known as decoding.

Decoding usually refers to the problem of finding the hidden state sequence that was most likely to have generated the given observation sequence. Different from exact model evaluation, the summation in (2.9) is approximated by the single best state sequence. One of the most popular solutions to this problem is the Viterbi algorithm. Assuming Bellman’s principle of optimality [Bellmann, 1952], Viterbi decoding can be implemented with the help of dynamic programming as a shortest-path problem through a 2-dimensional graph, also called a trellis (see Fig. 2.4). According to this principle the shortest path or the most likely state sequence can be calculated recursively as a concatenation of optimal sub-paths.

Let $\phi_j(t)$ denote the likelihood of the most likely HMM state sequence ending in state $j$ at time $t$ after observing the partial feature sequence $o_1, o_2, \ldots, o_t$. The Viterbi algorithm is then given as follows:
2.2. Automatic speech recognition with hidden Markov models

![Diagram of Viterbi decoding algorithm](image)

Figure 2.4: Illustration of the Viterbi decoding algorithm [from Young et al., 2006, p. 10].

1. Initialization:

\[
\phi_j(1) = \pi_j b_j(o_1), \quad 1 \leq j \leq N, \tag{2.10}
\]

\[
\psi_j(1) = 0. \tag{2.11}
\]

2. Recursion:

\[
\phi_j(t) = \max_{1 \leq i \leq N} \{ \phi_i(t-1) a_{ij} \} b_j(o_t), \quad 1 \leq j \leq N, 2 \leq t \leq T, \tag{2.12}
\]

\[
\psi_j(t) = \arg\max_{1 \leq i \leq N} \{ \phi_i(t-1) a_{ij} \}, \quad 1 \leq j \leq N, 2 \leq t \leq T. \tag{2.13}
\]

3. Termination:

\[
\Phi^* = \max_{1 \leq i \leq N} \{ \phi_i(T) \}, \tag{2.14}
\]

\[
q^*_T = \arg\max_{1 \leq i \leq N} \{ \phi_i(T) \}. \tag{2.15}
\]

4. Optimal state sequence via backtracking:

\[
q^*_t = \psi_{t+1}(q^*_{t+1}), \quad t = T - 1, T - 2, \ldots, 1. \tag{2.16}
\]

The Viterbi approximation results in a computational efficient solution but is only valid when a single network path dominates all other competing state sequences. Furthermore,
when a network of HMMs is created from a set of individual word or phone models, the size of the resulting composite model can become quite large for large vocabularies. One way to significantly reduce computation for large networks is to only consider paths which have a good chance of being part of the optimal solution. Such a selective search technique is called beam search [Lowerre, 1976] or pruning [Lingyun and Limin, 2004]. The low vocabulary tasks investigated in this thesis led to relatively small composite HMM networks, such that pruning was deemed unnecessary.

2.2.4 Performance measures

The recognition performance of a speech recognizer is usually measured by comparing the hypothesized word string with a correct reference transcription. Fig. 2.5 shows an example for the three types of errors that may occur during such a matching process. While substitution errors are the result of a word being incorrectly identified, insertion and deletion errors are observed when the hypothesized label sequence either contains extraneous words or omits a correct word. In HTK, each type of error is associated with a pre-defined penalty score, such that the string matching process involves a search for the optimal label alignment with the lowest possible penalty. After optimally aligning recognized and reference label sequence with the help of a non-linear string matching algorithm (e.g., dynamic programming), various performance statistics can be reported. Using the HTK alignment tool HRESULTS [Young et al., 2006], the following two standard performance measures are computed throughout this thesis. The percentage correctness score is defined as

\[
COR = \frac{NUM - DEL - SUB}{NUM} \times 100%,
\]

(2.17)

where NUM is the total number of tokens in the test set and DEL and SUB denote the deletion and substitution errors, respectively. The second performance measure, the percent accuracy is defined as

\[
ACC = \frac{NUM - DEL - SUB - INS}{NUM} \times 100%
\]

(2.18)

and in contrast to (2.17) also considers insertion errors, denoted here as INS. The accuracy score is therefore considered the more representative performance measure.

<table>
<thead>
<tr>
<th>LAB:</th>
<th>FOUR</th>
<th>SEVEN</th>
<th>THREE</th>
<th>ZERO</th>
</tr>
</thead>
<tbody>
<tr>
<td>REC:</td>
<td>FOUR</td>
<td>OH</td>
<td>NINE</td>
<td>THREE</td>
</tr>
<tr>
<td>ERR:</td>
<td>I</td>
<td>S</td>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.5: Example for a substitution (S), insertion (I) and deletion (D) error between a hypothesized label sequence (REC) and the correct transcription (LAB).
2.3 Noise robust speech recognition

The statistical pattern matching approach, outlined in the last section, suffers from severe performance degradation when there is an acoustic mismatch between training and testing environment. This is known as the robustness problem. A vast amount of techniques have been proposed over the last couple of decades for improving the robustness of ASR systems. Robust ASR methods can be classified into the following three categories: speech enhancement, robust feature extraction and model compensation techniques. In the following, we provide background information relevant to the topic of this thesis by briefly discussing some of the methods of each category. A more comprehensive review can be found in Rabiner and Juang [1993] and Gong [1995].

2.3.1 Speech enhancement

Most speech enhancement techniques are designed to recover the clean speech from noisy signal observations. Usually, these methods make different uses of a priori information about the speech and the noise [Gong, 1995] and are integrated as a preprocessing step in conventional enhancement-recognition schemes. Depending on the number of available microphones, speech enhancement approaches can be divided into single-channel and multi-channel techniques. Here, we only describe spectral subtraction as an example for single-channel speech enhancement, because the focus of this research is on multi-channel source separation, which is discussed in more detail in Chapter 3.

Spectral subtraction

Spectral subtraction is a method for restoring the magnitude or power spectrum of a signal corrupted by additive noise [Vaseghi, 1996]. In its most basic form, an estimate of the average noise spectrum is subtracted from the noisy signal spectrum in order to recover the clean speech. The noise spectrum is estimated from the periods when only the noise is present, e.g., in speech pauses. For waveform restoration, the enhanced magnitude or power spectrum is combined with the phase of the noisy signal and transformed in the time domain using the inverse Fourier transform.

Berouti et al. [1979] proposed a more general spectral subtraction (GSS) scheme in which the clean speech spectrum estimate is given by

\[
|\hat{S}(t, f)|^\gamma = \begin{cases} 
|X(t, f)|^\gamma - \alpha|\hat{N}(t, f)|^\gamma, & \text{if } |X(t, f)|^\gamma - \alpha|\hat{N}(t, f)|^\gamma > \beta|\hat{N}(t, f)|^\gamma \\
\beta|\hat{N}(t, f)|^\gamma, & \text{otherwise.}
\end{cases}
\]  

(2.19)

Here, \(\hat{S}(t, f)\) is the estimate of the clean speech signal, \(X(t, f)\) is the noisy signal and \(\hat{N}(t, f)\) is the estimate of the average noise spectrum. The type of spectral subtraction
is determined through the exponent $\gamma$, which for $\gamma = 1$ leads to magnitude spectral subtraction and to power spectral subtraction for $\gamma = 2$. The term GSS stems from the two additional parameters $\alpha$ and $\beta$, where $\alpha$ is a noise over-estimation factor and $\beta$ is a floor parameter used to prevent negative estimates of the spectral subtraction output.

Spectral subtraction is simple and efficient, but suffers from some serious drawbacks [Gong, 1995; Vaseghi, 1996]:

1. The performance of spectral subtraction schemes depends heavily on the quality of the noise estimate. The reliable detection of noise-only segments represents a considerable challenge in practical implementations.

2. The application of spectral subtraction may introduce new artifacts, such as musical noise, through its non-linear rectification process in (2.19). Due to the noise masking capabilities of the human auditory system, such distortions may not be perceived by humans, but they can influence speech recognizers in a negative way.

3. The stationarity assumption of the noise limits the applicability of the method to rather artificial types of noise distortions. Consequently, spectral subtraction fails for most real-world noises which are characterized by non-stationary behavior.

For speech parameterization involving mel-frequency scaling, it has been suggested by Nolazco Flores and Young [1994] that the application of spectral subtraction preceding the mel-filter bank is superior than after mel-frequency scaling. Following this recommendation, we have employed a modified version of general spectral subtraction in the linear frequency domain as part of the feature extraction in Article 4 of this thesis. Our method estimates the noise through a multi-channel source separation technique, which brings the advantage of circumventing the stationarity assumption necessary in single-channel spectral subtraction.

### 2.3.2 Robust feature extraction

Robust feature extraction continues to be an active research topic which attracts much attention at international conferences and workshops. Robust feature extraction techniques have been the subject of intense research since the early 1990s and it is infeasible to provide a comprehensive review of all methods. The scope of this section is therefore limited to a brief overview of some selected techniques related to the topic of this dissertation. For more information about robust feature extraction methods, the interested reader is referred to the excellent review articles and book publications about this subject [Gong, 1995; Junqua and Haton, 1996, 1998; Junqua and van Noord, 2001; Stern et al., 1996, 1997; Viikki, 2001].
2.3. Noise robust speech recognition

**Cepstrum based representations**

A large number of alternative cepstral feature representations have been developed that build upon the large success of MFCC features. For example, Gu and Rose [2001] proposed perceptual harmonic cepstral coefficients (PHCC) for noise robust speech recognition. PHCCs are designed to emphasize pitch harmonics when estimating the spectral envelope of voiced speech segments. Another extension was developed by Milner and Vaseghi [1994] in the form of cepstral time matrices (CTMs), where a 2-dimensional DCT is applied to the log-spectrum in order to decorrelate both frequency and temporal feature dimension. More recently, a similar approach was proposed by Tyagi et al. [2003] under the name of mel-cepstrum modulation spectrum (MCMS). In MCMS, cepstral trajectories over the basis of sines and cosines produce the cepstral modulation frequency response of the signal. Significant increases in recognition performance over standard MFCCs and PLP-RASTA features, defined in the next paragraph, are reported for clean as well as various noise conditions. However, for the scope of this thesis, we have restricted our cepstral feature representation to the standard implementation provided by the HTK platform [Young et al., 2006].

**Perceptual Linear Predictive Analysis**

Hermansky [1990] developed a new technique for the analysis of speech, which he called perceptual linear predictive (PLP) analysis. PLP is a feature representation which implements several aspects from the field of psychoacoustics, such as perceptual frequency scales, the equal-loudness curve and the intensity-loudness power-law relation. The resulting auditory-like spectrum of speech is approximated by an autoregressive all-pole model and simulates the above mentioned well-known properties of human hearing through engineering approximations [Hermansky, 1990]. Comparisons between PLP and MFCC features have shown somewhat inconclusive results. While Holmes and Holmes [2002] reported comparable recognition results for both feature types, Milner [2002] observed advantages for MFCCs over PLPs. The PLP technique has also been used extensively in combination with RASTA processing, termed PLP-RASTA, for handling both additive and convolutional noise distortions (see Section 2.3.2).

**Frequency filtering**

Frequency filtering can be interpreted as a simple linear transformation that performs a convolution between logarithmically scaled filterbank energies (FBE) and the impulse function of a finite impulse response (FIR) filter [Nadeu et al., 2001]. Mathematically, this
operation can be expressed as

\[ F(b) = S(b) \ast h(b), \quad b = 1, \ldots, B \] (2.20)

where \( F(b) \) is the frequency filtered sequence of the filterbank values \( S(b) \), \( h(b) \) specifies a given impulse response sequence and \( B \) is the total number of frequency bands. Nadeu et al. [2001] investigated the following two FIR filters, called FF1 and FF2, for which the impulse responses are defined as

\[ h_{FF1}(k) = \{+1, -1\}, \] (2.21)
\[ h_{FF2}(k) = \{+1, 0, -1\}. \] (2.22)

The corresponding transfer functions for these two filters are given by

\[ H_{FF1}(z) = 1 - z^{-1}, \] (2.23)
\[ H_{FF2}(z) = z - z^{-1}, \] (2.24)

respectively. The filter FF2 was proposed by Nadeu et al. [2001] as the default choice for frequency filtering because it is computationally simple, data-independent and has been shown to work well for a broad range of conditions. Essentially, frequency filtered features correspond to the frequency derivatives of mel-scaled log-filterbank energies or, in other words, frequency filtering captures information related to the slope of the spectral envelope.

When compared to filterbank and MFCC features, frequency filtering offers a range of advantages. FF based features not only have been shown to correlate well with perceptual data [Klatt, 1982], they also are effectively decorrelated, just like MFCCs. However, in contrast to cepstral coefficients, frequency filtering remains in the log-spectral feature domain and therefore retains the noise localization property of filterbank energies. This makes them especially suited for missing feature approaches [de Veth et al., 2001]. In Article 4 of this thesis, we investigate the use of FF2 based features for their use in missing data speech recognition when applied in isolation and in combination with FBE features.

**Removal of slow variations**

Removing convolutional noise caused by transmission channel distortions has been found to significantly improve the recognition accuracy [Hanson and Applebaum, 1993]. Filters that remove these slow variations in the feature vectors can be implemented in different parameter spaces. The log-spectral as well as the cepstral parameter space are especially appealing because convolutional noise in the time domain becomes an additive noise in these feature domains.
2.3. Noise robust speech recognition

The most popular method to remove slow variations in the cepstral domain is cepstral mean normalization (CMN). CMN works by simply subtracting the long term average from the cepstral feature vectors. It has been reported [Atal, 1974; Furui, 1981; Hanson and Applebaum, 1993] that CMN is very effective in practice and provides a simple way to compensate for long term spectral effects caused by different microphones and transmission channels.

Work by Hermansky [1991] on RASTA (RelAtive SpecTrAl) processing has established a similar compensation framework for the log-spectral feature domain. In RASTA, the time trajectories of the acoustic features in each frequency band are filtered to remove the constant additive offsets caused by the convolutional distortions in the time domain. Several variations of RASTA have been proposed applying it to PLP features [Hermansky, 1991], the mel-cepstral domain [Hirsch et al., 1991] or extending the method further to handle both additive and convolutional noise distortions [Hermansky et al., 1993]. Similar to CMN, RASTA has been shown to give good recognition results for telephone speech and microphone channel variations [Hermansky et al., 1993; Murveit et al., 1992].

2.3.3 Model compensation techniques

In contrast to speech enhancement or robust feature extraction, model compensation allows for the presence of the noise during the decoding process itself. The HMM paradigm provides a formal mathematical framework that is able to model the temporal and spectral variability in speech signals. This offers the possibility to accommodate for the noise by adapting the clean speech model parameters, such as the mean and variance of a Gaussian distribution, in a time-varying manner. Consequently, model compensation methods are well suited to deal with the often unavoidable discrepancy between training and testing conditions.

Parallel model combination

The parallel model combination (PMC) approach proposed by Gales and Young [1993] pursues the goal of adapting the mean and covariance statistics of the HMM state distributions, such that these better match the cepstral distributions of the incoming noisy speech. This is done by combining both clean speech and noise HMMs before recognition is attempted (see Fig. 2.6).

PMC is based on the assumption that the noise corruption is additive in the power domain. Because of this, the acoustic models of noise and speech are first converted from the cepstrum to the log-spectral domain and then back to the linear spectral domain. The combined models are then transformed back to the cepstral domain for recognition. In this way, acoustic models that better match the acoustic environment are generated.
PMC has been shown to produce considerable recognition improvements for simple and more complex noise environments [Gales, 1994; Gales and Young, 1996; Minami and Furui, 1995; Nolazco Flores and Young, 1994; Vaseghi and Milner, 1995]. However, most versions of PMC require a substantial amount of knowledge and data for training the noise model. Typically, observations with isolated noise samples are needed in order to adequately learn the noise model parameters from the data. Furthermore, the computational cost can become quite high for non-stationary distortions which require more complex HMM topologies for the noise model.

Multi-band speech recognition

Motivated by Fletcher’s theory of human speech perception [Fletcher, 1953], several approaches based on sub-band modeling have been developed for noise robust speech recognition [Bourlard and Dupont, 1996; Bourlard et al., 1996]. Multi-band systems are based on the assumption that colored or bandlimited noise only affects some of the frequency bands while leaving the remaining parts of the spectrum intact. If the noise corrupted sub-bands can be detected and ignored during recognition, the system’s robustness in noisy conditions can be improved [Bourlard et al., 1996; Misra, 2006; Morris et al., 2000].
2.3. Noise robust speech recognition

Fig. 2.7 shows the basic architecture of a multi-band system in which the spectral representation of the speech signal is split into several independently processed frequency bands. For example, feature type and processing may differ from band to band due to the use of different classifiers for every sub-band. However, the outputs of these sub-band classifiers need to be merged to reach a final recognition decision. A number of studies [Hagen, 2001; Hagen and Morris, 2004; Hagen et al., 1998, 2001] have investigated different weighting strategies for optimally combining the recognition outputs of the sub-band classifiers. Note, that the recombination of individual classification results can be performed at sentence, word or phone level.

Missing data techniques

A related approach to multi-band processing which is also inspired by the masking properties of the human auditory system has been developed in the form of missing data (MD) recognition for dealing with partial noise corruptions [Cooke et al., 1994; Lippmann and Carlson, 1997]. The basic idea in missing data ASR is to discard the noise corrupted parts of the spectro-temporal feature representation of the noisy speech signal and perform the recognition only on the remaining reliable feature components. This requires a segmentation procedure which partitions each feature vector $o$ into reliable or present and unreliable or missing parts, such that $o = (o_p, o_m)$. The HMM state distributions are trained in the usual way (see Section 2.2.2). Only during recognition the speech models are modified to cope with noise distortions. As usual, the HMM state distributions are assumed to be modeled by a mixture density function

$$ b(o|\Lambda_j) = \sum_{r=1}^{R} c_{jr} b(o|\Lambda_{jr}), \quad (2.25) $$
where $\Lambda_j$ are the model parameters for HMM state $j$, $R$ is the total number of mixture components and $c_{jr}$ denotes the mixture weight.

Using these assumptions, the following techniques can be applied to improve the robustness of the decoder in noisy environments:

**Marginalization** denotes the process of integrating the HMM state observation probability density function over the missing vector components. This can be formulated as [Cooke et al., 2001]

\[
b(o_p|\kappa, \Lambda_j) = \int b(o|\kappa, \Lambda_j) \, do_m = \int b(o_p, o_m|\kappa, \Lambda_j) \, do_m \tag{2.26}
\]

\[
= \sum_{r=1}^{R} c_{jr} b(o_p|\Lambda_{jr}) \int_{o_{low,m}}^{o_{high,m}} b(o_m|\Lambda_{jr}) \, do_m, \tag{2.27}
\]

where $\kappa$ represents any additional information about the clean feature value, such as lower and upper feature bounds $o_{low,m}, o_{high,m}$. If no other knowledge is available apart from the present data $o_p$, the integration is performed over the interval from $-\infty$ to $\infty$ reducing the integral to unity. This is called full marginalization. The performance of full marginalization often degrades considerably for low SNRs due to the lack of discriminatory information in spectro-temporal areas with large amounts of missing features. Bounded marginalization, on the other hand, exploits any extra available knowledge about the range of possible values the missing data could take by restricting the integration interval to $[o_{low,m}, o_{high,m}]$. The use of energy bounds rectifies the above mentioned problem in full marginalization and leads to a very significant improvement in speech recognition accuracy at most SNRs. For the case that $b(o|\Lambda_j)$ is represented by mixtures of diagonal covariance Gaussians, the evaluation of the marginal density and the Gaussian integrals is considerably simplified.

**Missing data imputation** is a general term for a variety of methods designed to handle the missing data problem [Little and Rubin, 1986]. Different to marginalization techniques, imputation aims at reconstructing the missing parts of the spectro-temporal feature representation. The estimation of the missing values can be performed based on the present data [Cooke et al., 1997], the HMM state distributions [Josifovski et al., 1999] or with the help of a separate clean speech model [Raj and Stern, 2005; Raj et al., 2000]. From a practical point of view, the last method is particularly attractive as it does not require modifications to the recognizer and allows to convert the reconstructed, complete feature vector to the cepstral domain. Several studies [Cooke et al., 1997, 2001; Josifovski et al., 1999; Raj et al., 2004] have reported inferior recognition results for imputation strategies when compared with
(bounded) marginalization techniques as long as recognition was performed on re-constructed log-spectra. However, once the spectra were converted to the cepstral feature domain followed by CMN, imputation strategies became superior to spectral marginalization [Raj, 2000; Raj and Stern, 2005].

Several extensions of these classic MD approaches have been developed in recent years. For example, Barker et al. [2000] extended the hard partitioning of feature elements in present or missing components to a fuzzy segmentation, where continuous values between 0 and 1 indicate the feature reliability. Other extensions perform marginalization in the cepstral domain [Häkkinen and Haverinen, 2001; Jun et al., 2004] or use an imputation approach based on PROSPECT features [Van Hamme, 2004] for acoustic modeling.

The approach pursued in this thesis is based on the work by Morris et al. [2001], who developed a probabilistic missing data framework for dealing with uncertain data. They introduced the notion of evidence modeling in order to express the belief to which each possible feature datum represents the true clean data value by a probability density function. This provided a theoretical basis for the mostly heuristically motivated marginalization schemes discussed above. For example, the bounded marginalization equation can be derived from this concept by using a two-component mixture evidence model consisting of a Dirac-delta function and a Uniform distribution [Morris et al., 2001]. More importantly, evidence modeling also offered new insights into how data preprocessing and decoding could be linked together more efficiently. However, so far, the concept has not received much attention and the simple Dirac-Uniform mixture PDF can still be considered the predominant evidence model in missing data recognition. Further details on this topic are given in Article 4 of this dissertation.

**Uncertainty decoding**

Uncertainty decoding (UD) refers to a broad class of algorithms that exploit the notion of feature uncertainty for noise robust ASR [Liao and Gales, 2008]. The particular UD technique discussed here, assumes that the feature uncertainty can be quantified along with the enhanced feature value during the enhancement process. The uncertainty is then propagated to the decoding stage, where the clean speech model variances are compensated accordingly.

To understand this, we consider the UC method proposed by Deng et al. [2005], where the basic idea is to account for the imperfections of speech enhancement by integrating the HMM state observation probability over all possible feature values. This can be formulated as

\[
\int b(o|q, m)p(o|\theta) \, do,
\]

(2.28)

where \(o\) is the unknown clean feature vector, \(\theta\) denotes the parameter vector characteriz-
Fundamentals of automatic speech recognition

2. Fundamentals of automatic speech recognition

The speech enhancement procedure and $b(o|q, m)$ is the $m$-th mixture component of the observation distribution in HMM state $q$. It was suggested by Deng et al. [2005] that a Gaussian distribution $G(o; \hat{o}, \Sigma_\hat{o})$ is a good choice for modeling $p(o|\theta)$. Here, the enhanced speech value is denoted as $\hat{o}$ and the uncertainty due to the speech enhancement algorithm is given by the variance $\Sigma_\hat{o}$. As discussed in Section 2.3.3, the observation density of each HMM state is commonly modeled as a mixture of Gaussians. Under these assumptions, Deng et al. [2005] have shown that the compensated observation likelihood is given by

$$
\int G(o; \mu_{q,m}, \Sigma_{q,m})G(o; \hat{o}, \Sigma_\hat{o}) \, do = G(\hat{o}; \mu, \Sigma_{q,m} + \Sigma_\hat{o}).
$$

(2.29)

As evident from (2.29), the net effect of UD on the observation likelihood calculation is a dynamic variance compensation which increases the variance of the gaussian mixture component. As a result, the more the enhanced feature deviates from the clean value the less it will contribute to the overall likelihood. UD is computationally very efficient when diagonal covariance matrices are used for $\Sigma_{q,m}$ and $\Sigma_\hat{o}$.

It has been shown [Arrowood, 2003; Benitez et al., 2004; Deng et al., 2005] that the utilization of the speech feature uncertainty can lead to a significant improvement in the ASR accuracy on small vocabulary tasks. Similar to missing data techniques (see Section 2.3.3), the performance improvement was particularly prominent when the variance of the enhanced features was known a priori [Deng et al., 2005]. However, unlike missing data techniques, UD can operate in the cepstral feature domain which better fits the diagonal covariance assumption. In Article 4, we present a novel approach that combines the variance compensation ability of UC with the bounded marginalization feature of MD-ASR and investigate its performance on a small vocabulary task involving both additive and convolutive noise distortions.

2.4 Summary

In this chapter, we have described some of the fundamental concepts used in modern ASR technology. First, the task of finding the most likely word sequence for an unknown input utterance was formulated as a statistical pattern matching problem. Next, we discussed the main building blocks of a state-of-the-art HMM recognition system. Starting with the feature extraction stage we continued with a description of the acoustic modeling process using HMMs. Strategies for model training and evaluation were reviewed along with a discussion of the quality measures used to assess the speech recognition performance. The second part of the chapter was devoted to the problem of noise robust speech recognition caused by the acoustic mismatch between training and testing environment. Several approaches for reducing the mismatch through speech enhancement
algorithms, noise robust feature extraction and model compensation techniques were summarized. The case of multi-channel source separation as a preprocessing technique will be discussed in more detail in the next chapter.

References


2. Fundamentals of automatic speech recognition

Germany, 1997.


J.-C. Junqua and J.-P. Haton. *Robustness in Automatic Speech Recognition: Fundamentals and


Chapter 3

Review of multichannel source separation techniques

As discussed in the previous chapter, one popular approach in robust ASR is to perform some sort of speech enhancement prior to speech recognition. One promising way to separate speech from interfering noise is by using spatial information of the sound scene through the use of several spatially separated microphones, called an array. Microphone array processing allows the user a hands-free communication as it does not require wearing a close-talk microphone. More importantly, it can lead to substantial improvements regarding the signal-to-noise ratio (SNR) of the processed signal.

This section gives an overview of microphone array based source separation techniques, particularly those which are relevant to the topic of this dissertation. Source separation has been studied for over two decades and a wide range of different methods has been proposed for tackling the problem, ranging from beamforming over statistical approaches to auditory model inspired processing. Our review is by no means a comprehensive survey of all existing techniques, and the interested reader is directed to the manifold literature on the topic [Benesty et al., 2005; Divenyi, 2005; Hyvärinen, 1999; Hyvärinen and Oja, 2000; Lee, 1998; O’Grady et al., 2005; Omologo et al., 2001; van der Kouwe et al., 2001; van Veen and Buckley, 1988].

3.1 Introduction

When presented with a number of audio mixtures recorded using a set of microphones, the process of extracting the constituent source signals that make up the mixtures is called multi-channel acoustic source separation. Blind source separation (BSS) characterizes the extreme case where no additional information in regards to the mixing process or the involved source signals is required for separation. More formally, the problem can be written as follows [O’Grady et al., 2005]: Given $M$ observations of $N$ sources mixed via an unknown $M \times N$ mixing matrix $A$, the BSS objective is to estimate the underlying sources, $S$, solely from the mixtures, $X$.

Depending on the dimensionality of the mixing process, BSS problems are categorized into the following three cases. If $M = N$, the problem is called even-determined and a linear transformation of the data is sufficient for estimating the underlying sources. In the case of $M > N$, the problem is over-determined and, provided that $A$ is full
3. Review of multichannel source separation techniques

rank, can be solved by least-square optimization or linear transformation involving matrix pseudo-inversion [O’Grady et al., 2005]. If $M < N$, the problem becomes under-determined and a non-linear technique is usually required to achieve separation. For under-determined problems, source separation is not equivalent to mixing matrix estimation and demixing the sources from the mixture is not straightforward.

Another classification of source separation problems is related to the assumptions made about the mixing process itself. Microphone recordings from a natural environment may include a large number of acoustic sources as well as their signal reverberations. In this difficult scenario, the audio processor needs to realize that several signals arriving from multiple directions at different times may actually belong to one individual source [O’Grady et al., 2005]. In order to reduce the complexity of the problem it is common to make several simplifying assumptions about the surroundings in which the microphone observations are recorded.

Typically, three different types of mixing models are distinguished in the BSS literature. The instantaneous mixing model is the most restrictive case, where the source signals arrive instantly at the microphones without any delays but with different signal intensities. This type of mixing model is useful in certain biomedical applications but becomes highly unrealistic for real acoustic environments. The next more realistic mixing model is the anechoic case, which describes the propagation of source signals along single non-dispersive paths to each of the microphones. The anechoic mixing model can be further extended to the echoic case by allowing the signals to travel along multiple paths between each source and each microphone. This is the most realistic mixing model and is usually modeled mathematically as a convolution operation, i.e., the source signals are convolved with room impulse responses. The echoic mixing model is therefore sometimes referred to as convolutive mixing model.

3.2 Methods

3.2.1 Beamforming

Spatial filtering exploits the fact that the desired and interfering signals often originate from different spatial locations [van Veen and Buckley, 1988]. One example of spatial filtering is called beamforming, where a microphone array is used to steer the main beam of a spatial directivity pattern towards a specific direction of interest. The desired direction is also referred to as look direction. A beamformer forms a linear combination of the microphone observations allowing to reinforce signals propagating from the look direction, while attenuating energy impinging on the array from other angles. In its simplest form, this may only require the signals received by the array to be delayed in order to compensate for the different arrival times at the microphones. An example for a frequency
3.2. Methods

Beamformers are usually classified into several categories depending on how the spatial filter weights are chosen. In the following, we will briefly discuss fixed and adaptive beamforming techniques as well as comment on their benefits and shortcomings. Some literature references are also provided for further reading.

In fixed beamforming, the spatial weights are chosen independently from the particular signal-interference scenario and do not depend on the array data. The best-known technique is called delay-and-sum beamforming and simply averages the time-aligned microphone signals for enhancing the target signal. The task of finding the correct delays is related to the problem of source localization. Further details about source localization methods can be found in Malonakis et al. [2000]. The widespread appeal of delay-and-sum processing stems from its simplicity and ease of implementation. It has successfully been used as a preprocessing step in many speech recognition applications and is often used as a multi-channel baseline for more advanced beamforming techniques [Omologo et al., 2001; Seltzer, 2003]. However, the major drawback of delay-and-sum arrays is the large number of microphones that are required to increase the SNR. Under the idealistic assumption that the jammer signals are completely uncorrelated between the microphones and with the target signal, the SNR will at most increase by 3 dB for every doubling of the number of microphones in the array [Johnson and Dudgeon, 1993]. However, physical reality places limitations on the number of sensor elements that can be used in practice.

In adaptive beamforming, the weights are computed based on the data statistics received at the microphones in order to optimize the array response. In general, an adaptive beamformer places nulls in the direction of the interfering sources in an attempt to maximize the signal-to-interference-plus-noise ratio at the beamformer output [van Veen and Buckley, 1988]. In Article 3 of this thesis we use an approach, called linearly constraint minimum variance (LCMV) beamforming [Applebaum and Chapman, 1976], in order to enhance a desired signal while suppressing interfering sources from other directions. The weights are chosen to minimize the output signal power subject to a constraint response vector. The constraints prevent the trivial solution, where all of the filter weights are set to zero and offer added control over the spatial directivity pattern. For example, if the directions of the interfering sources are known it may be desirable to force zero gains in these directions while maintaining a unity response in the look direction. A special case of the LCMV beamformer is the minimum variance distortionless response (MVDR) beamformer, which imposes a single distortionless response constraint in the direction of the target signal. Other types of adaptive beamforming include eigenvector beamforming [Haimovich and Bar-Ness, 1988] and partially adaptive arrays [Morgan, 1978]. When compared to delay-and-sum arrays, the interference nulling allows adaptive beam-
Review of multichannel source separation techniques

formers to achieve much higher SNR improvements with fewer sensors. However, the
statistics required to compute the filter weights are usually unknown, and therefore need
to be estimated from the available data. This can be done using sequential adaptive al-
gorithms, such as least-mean-squares (LMS) [Frost, 1972] or recursive least squares (RLS)
[Yang and Böhme, 1992], which update the beamformer weights each time a new data
sample becomes available. Another option are block-adaptive methods, such as sam-
ple matrix inversion (SMI) [Capon, 1969], which collects a block of data and updates the
beamformer weights each time a new block of data is received.

The major problem in adaptive beamforming is signal cancelation, which refers to the
undesired case where the beamformer attenuates the target signal as if it were interference.
This may be caused through inaccurate estimates of the source statistics or directions. Another reason is correlation between target and interfering signals, as is often
the case in reverberant environments with multi-path propagation. Several approaches
have been suggested to reduce the problem of signal cancelation [Cox et al., 1987; Shan
and Kailath, 1985; Yang, 1987]. However, most of them require a large number of sen-
sor elements to give satisfactory performance in reverberant environments. A notable
exception is the approach by van Compernolle [1990], who avoids signal cancelation by
only adapting the filter weights during noise dominant portions of the incoming signal.
Using only four microphones, he obtains some improvement in ASR performance. Cer-
mak et al. [2006] and Araki et al. [2007a] use a similar concept for frequency domain
beamforming. They first detect the active time-frequency slots for each speech source
and then use this information to calculate the spatial filter weights for each frequency
band. The reported separation results indicate superior performance when compared
with conventional beamforming which utilizes all time-frequency points in the spatial
filter weight calculation. More recently, separation performance has also been evaluated
on ASR tasks. Using the maximum SNR beamforming approach of Araki et al. [2007a],
Kolossa et al. [2008] report significant gains in speech recognition accuracy for both over-
and even-determined BSS cases.

In conclusion, the ability to adjust the filter weights over time makes adaptive beam-
forming a promising approach for the separation of multiple speech sources (e.g., in a
meeting room) in time-varying environments.

3.2.2 Independent component analysis

Independent component analysis (ICA) aims to find a linear representation of non-gauss-
ian data (mixtures) such that the extracted components (sources) are statistically indepen-
dent, or as independent as possible [Hyvärinen and Oja, 2000]. To guarantee separation,
two assumptions must be satisfied: Firstly, the sources are assumed to be mutually in-
dependent. Secondly, at most one of the independent components has a Gaussian distri-
3.2. Methods

bution. ICA is generally implemented as an optimization problem (see Fig. 3.2), where the independent components are derived from maximizing some measure of independence, also called contrast function. Such contrast measures include mutual information [Comon, 1994], entropy [Bell and Sejnowski, 1995], non-gaussianity [Hyvärinen et al., 2002], and sparseness [Zibulevsky and Pearlmutter, 2001]. An overview of various ICA methods and their applications can be found in Hyvärinen [1999]; Hyvärinen and Oja [2000]; Lee [1998].

As with any technique, ICA has some drawbacks that limit its practicability. In general, ICA cannot determine the actual number of source signals, a uniquely correct ordering of the sources (permutation ambiguity) nor the proper scaling of the source signals (scaling ambiguity). Moreover, because the demixing process relies on matrix inversion, ICA can only be applied to even- or over-determined BSS problems. Araki et al. [2003] investigated the limitations of frequency-domain ICA for the convolutive mixture problem and found that the separation performance is actually upper bounded by that of adaptive beamforming. However, in contrast to adaptive beamforming, ICA methods do not require knowledge about the array geometry and source locations and can adapt in the presence of target and jammer signal. The separation quality is usually determined using some SNR related or perceptual quality criteria but has also been assessed in terms of speech recognition improvements (see Low et al. [2004] for an example).

3.2.3 Sparse source separation

An increasing number of BSS methods are based on the assumption that the underlying source signals are sparse. Several definitions for sparseness exist declaring signals to be sparse if they "contain as many zeros as possible" [Georgiev et al., 2005] or can be de-
scribed as "representations where a small percentage of the signal coefficients capture a large percentage of the signal energy" [Yilmaz and Rickard, 2004]. In order to quantify the sparseness of a signal, a variety of measures can be used. Such measures of sparsity include tanh-functions [Karvanen and Cichocki, 2003], the Gini index [Rickard and Fallon, 2004] or kurtosis [Li and Lutman, 2006]. A sparse representation can often be obtained by projecting the source on a Fourier, Gabor or Wavelet basis. In such a sparse domain, only a limited number of sensors are required to perform source separation because only a limited number of sources will be active at any given data point. Consequently, a large number of techniques exploit the sparse nature of most interesting real world signals when solving the difficult under-determined BSS problem.

In a first line of research, called sparse component analysis (SCA) [Cichocki et al., 2004; Georgiev et al., 2005; Zibulevsky and Pearlmutter, 2001], blind source separation is approached by seeking sparse representation in an over-complete basis, called a dictionary. The aim of these methods is the estimation of the coefficients of the sources, as represented in the dictionary, and not the source signals themselves. This can be formulated as a $L_0$ optimization problem and would theoretically require an exhaustive search across all possible representations. Fortunately, it has been shown [Donoho and Elad, 2003] that this combinatorial search problem can be avoided by minimizing a $L_1$ optimization criterion instead. $L_1$ minimization leads to a sparse solution but can be implemented very effectively using linear programming techniques (e.g., see the Basis Pursuit algorithm by Chen et al. [1998]).

Another research direction [Abrard and Deville, 2005; Li et al., 2006; Yilmaz and Rickard, 2004] exploits source sparsity as a pre-processing step and works under the assumption that source signals have disjointed support in the frequency domain, i.e. all sources will not be present at all time-frequency points. In this thesis, we focus on a class of sparse BSS methods that employ histograms or cluster algorithms for deriving the mixing parameter estimates. Of particular interest are methods that rely on time-frequency
3.2. Methods

masking for source demixing. The most popular method is arguably the DUET method [Yilmaz and Rickard, 2004], which exploits the sparsity of speech in the time-frequency domain to separate an arbitrary number of speech signals using only two anechoic mixtures. The speech mixtures are passed through a short-time Fourier transform designed to create a sparse representation, and for every time-frequency cell the relative attenuation and phase shift between the two microphones is estimated. These estimates are then used to compute a (weighted) histogram whose peaks are used to estimate the time-frequency masks required for demixing. Alternatively, in Araki et al. [2007b] the observations related to relative attenuation and phase shift are assigned directly to one of the source signals with the help of the k-means cluster algorithm.

The separation quality of these methods has been evaluated in real environments with weak reverberation. Although the output signals are generally of good quality, they can suffer from audible musical noise distortions as a consequence of the binary masking. Moreover, separation performance deteriorates quickly for higher reverberation times because the histogram peak detection and off-the-shelf cluster algorithms are highly vulnerable to outliers and noise in the data set. In Article 2 and 3, we take a closer look at this problem and propose some new measures to improve the robustness of cluster algorithms in reverberant environments. Due to the non-linear masking procedure, the demixed sources are not directly suitable for ASR evaluation with a conventional cepstral HMM recognizer. This is demonstrated in more detail in Article 1 along with a possible solution that combines the BSS outputs directly with a missing data recognizer. This combination was motivated by work in the field of computational auditory scene analysis (CASA) [Bregman, 1990; Cooke et al., 2001; Divenyi, 2005], where it has been quite common to evaluate the separation performance on ASR tasks. Some of these auditory segregation models are described in the next section.

3.2.4 Auditory model inspired processing

Based on everyday life experience, it is clear that our human auditory system is highly effective in performing the source separation task under extremely difficult conditions. In order to advance the separation capabilities of microphone array based systems, several approaches have been developed which mimic the processing that takes place in the human auditory system.

For example, cross-correlation type processing is thought to be an important part of the human auditory system and the first binaural model implementing such a mechanism for sound localization was that of Jeffress [1948]. Later, Sullivan [1996] investigated the use of cross-correlations between multiple microphones to improve not only sound localization but also the quality of the input speech signal fed to the speech recognizer. He reported moderate speech recognition improvements over simple delay-and-sum beam-
forming in real environments. In a more recent approach, called polyaural processing, Stern et al. [2007] revisited the idea of cross-correlation modeling. Auditory principles are implemented through non-linear processing of interaural filterbank channels and an optional second stage of cross-correlation across frequency. This extended method provided substantial gains in recognition accuracy when compared to simple delay-and-sum beamforming in the presence of interfering sources. As in other systems, performance degraded in reverberant environments but remained superior to delay-and-sum processing.

Other work focused on the amplitude modulation spectrogram (AMS), which is motivated from neurophysiological findings on amplitude modulation processing in higher stages of the auditory system in mammals. The AMS was shown to produce characteristic patterns for speech that are different from the representation of most noise types and reverberation. Amplitude modulation analysis was applied in Kollmeier and Koch [1994] for a binaural noise suppression scheme. However, only moderate improvements in speech intelligibility were reported and no ASR experiments were conducted.

Another example are binaural CASA systems, such as the ones proposed by Harding et al. [2006]; Roman et al. [2003] and Palomäki et al. [2004], which exploit interaural level and intensity differences as cues for source segregation. Unlike in spatial filtering, the localization information is used only for selecting the spectro-temporal areas that are representative of a particular source signal. No actual spatial filtering is performed for source separation. This is very similar to BSS techniques that rely on time-frequency masking for source separation. However, some key differences exist with respect to the particular mask estimation technique (supervised vs. unsupervised) and the statistical modeling of the localization cues. More importantly, while BSS systems aim to separate all source signals most binaural CASA models pursue the goal of extracting a single target signal from the acoustic background. Further details on the similarities and differences of BSS and binaural CASA systems are given in van der Kouwe et al. [2001] and Article 1. Binaural CASA methods have shown to produce significant improvements in terms of SNR and ASR performance, particularly in anechoic environments. The recognition accuracy is often evaluated with the help of a missing data recognizer [Harding et al., 2006; Palomäki et al., 2004; Roman et al., 2003] or more recently, with conventional cepstral decoding after data imputation [Roman et al., 2006] and uncertainty decoding [Srinivasan et al., 2007].

3.3 Summary

This chapter has presented a non-exhaustive review of the field of multi-channel source separation. After distinguishing between different types of mixing models, we briefly
described some of the main concepts and techniques along with their advantages and disadvantages.

With respect to our goals outlined in Section 1.3, we identified sparse BSS strategies that are based on time-frequency masking as ideal candidates for improving the link between preprocessing and automatic speech recognition. The DUET-BSS algorithm has been designed for speech enhancement scenarios and has not been evaluated in speech recognition applications. If combined with a missing data speech recognizer sparse BSS techniques may present an attractive alternative to binaural CASA front-ends. Such systems could be applied in scenarios where binaural processing is ineffective or not possible (e.g., mobile phones). Time-frequency masking and clustering strategies may also be combined with conventional spatial filtering techniques, like adaptive beamforming, which requires robust estimates of the source locations and jammer-only periods during the spatial filter design. Furthermore, the lack of robustness against reverberation in current cluster algorithms indicates a promising area for further research. An advantage of BSS systems is that they not only recover a target signal but also produce estimates of the interfering sources. Thus, BSS often results in a rich description of the acoustic scene, that could be exploited by passing on not only the target signal to the HMM decoder but also information regarding the strength of the interfering signals.

In the articles contained in this dissertation, we will build upon these ideas and implement them into new signal processing algorithms which connect preprocessing and recognition stages more effectively.

References


3. Review of multichannel source separation techniques


3. Review of multichannel source separation techniques


Part II

General discussion
Chapter 4

Discussion and conclusions

This chapter provides an integrated statement on the achieved outcomes of the research described in this work. We begin with a general discussion of the main findings and contributions of this thesis and establish the significance of the work in the context of previous research. We proceed with a description of the limitations of this thesis before concluding with suggestions for future research directions.

4.1 Major findings and contributions

In this section, we summarize the major findings and contributions of this thesis and discuss their significance in the context of other work.

In Article 1, we have investigated a straightforward combination of a two-sensor BSS technique, called DUET [Yilmaz and Rickard, 2004] with a missing data speech decoder. Our experimental evaluation for anechoic speech mixtures showed that recognition performance remained high, despite the presence of multiple competing speakers. It was found that the proposed median smoothing of the time-frequency masks can reduce spurious points that would otherwise lead to insertion errors in the speech decoder. In contrast to the supervised learning techniques employed in other approaches, DUET blindly estimates the separation masks based on a simple frequency independent classification of attenuation and delay parameters. The spatial cues are extracted from STFT ratios, which offer significant speedups over the computationally expensive cross-correlation functions used in most binaural CASA models. More importantly, DUET avoids recalibration for each new spatial source configuration. The approach can be advantageous for ASR scenarios, where microphone setups with large spacing or multiple sensors are not feasible (e.g., mobile phones). Furthermore, the avoidance of the waveform resynthesis step relaxes the requirement for the spectral analysis technique to be capable of perfect reconstruction and opens up new possibilities for choosing a suitable time-frequency transform. Alternative time-frequency decompositions, such as the discrete wavelet transform (DWT) or wavelet packets (WP), have been used in other BSS approaches [Kopriva and Seršić, 2007; Tan and Fèvotte, 2005] in order to exploit their improved sparseness properties. To develop this idea further, the STFT analysis could be replaced with a filterbank that directly simulates a perceptual frequency scale. For example, Kühne and
Togneri [2006] have successfully tested a continuous wavelet transform (CWT) for frame-synchronous speech feature extraction. The modified CWT directly approximates the mel-frequency scale and additionally offers superior time-frequency analysis through its multi-resolution feature. Such a filterbank front-end would eliminate the need for binning several STFT coefficients together in order to form mel-frequency scaled filterbank energy vectors. Similarly, the time-frequency masks used in the source separation could directly be passed on to the speech decoder without the frequency conversion step introduced in Article 1.

Following these initial investigations on anechoic sound mixtures, Article 2 dealt with the problem of robust source localization in reverberant environments. Successful source localization can be considered an important prerequisite for many spatial signal processing algorithms. As our main contribution, we investigated the use of a fuzzy c-means cluster algorithm which operates on location cues extracted from a linear microphone array. In order to increase the algorithm’s robustness against sound reflections we incorporated observation weights to emphasize reliable direction-of-arrival (DOA) cues over unreliable ones. The weights were computed from a priori learned local feature statistics around sound onsets. Experimental results illustrated the superiority of the new method in terms of DOA localization accuracy when compared with standard fuzzy clustering, especially for closely spaced sources in reverberant environments. Our motivation for the proposed weighting scheme stems from a number of previous studies [Faller and Merimaa, 2004; Huang et al., 1997; Palomäki et al., 2004] which have shown that in echoic enclosures only a small fraction of the location cues remain reliable indicators for the correct source locations. The proposed wFCM cluster algorithm offers an alternative implementation of this principle and may be useful in a number of BSS applications. For example, the developed wFCM algorithm may be of interest for unsupervised spatial filtering techniques that require robust estimates of the source locations for steering a microphone array. The algorithm could also be used in sparse BSS techniques that rely on clustering techniques for histogram peak detection [Aarabi and Mavandadi, 2003; Kim and Kil, 2007; Yilmaz and Rickard, 2004] or time-frequency mask estimation [Araki et al., 2005].

Article 3 improved the wFCM cluster algorithm further through the inclusion of context information and investigated its applicability for reverberant speech separation. The main contribution of this chapter was to establish the potential of neighborhood information as an efficient tool for increasing the robustness of standard cluster algorithms. To the best of our knowledge this is the first attempt in the field of blind source separation to directly incorporate neighborhood information into the parameter updates of a cluster algorithm. The approach is superior to the simple median smoothing, performed in Article 1, which exploits the neighborhood information solely during mask postprocessing.
We have shown that the new wCFCM cluster algorithm improves the source localization as well as the segmentation accuracy of conventional fuzzy clustering, particularly in reverberant conditions. Furthermore, wCFCM avoids the frequency permutation problem and is able to operate on data with short observation length. The developed methods may be of use for other BSS approaches, in particular for those that rely on clustering techniques for time-frequency mask estimation [Araki et al., 2007; Cermak et al., 2006]. Moreover, the inclusion of context information into the clustering opens up new possibilities for the integration of important principles from the field of auditory scene analysis [Bregman, 1990], such as common onset or common amplitude modulation. The implemented scheme can only be considered as a first step towards this goal and future contributions are required to fully exploit the potential of this idea. Other opportunities for the use of the new cluster algorithm may result from robust beamforming techniques [Bell et al., 2000; Chen and Vaidyanathan, 2006], which require an interval or a probability density function reflecting the level of DOA uncertainty. Estimating these DOA distribution could be one future area of application for the robust clustering strategies developed in this thesis.

In Article 4, we have evaluated the separation quality in terms of ASR performance by combining our multi-channel source separation technique (Article 3) with a missing data speech recognizer. Our main contribution is the introduction of the bounded-Gauss-Uniform PDF as a novel evidence mixture model for spectral missing data decoding. The new model enables a time-varying HMM variance compensation and properly reflects the bounded nature of spectral filterbank energies in both mixture components. Moreover, it includes several previously proposed evidence models, such as the Dirac-Uniform mixture or the univariate Gaussian PDF, as special cases. Among all tested evidence PDFs the bounded-Gauss-Uniform mixture model consistently achieved the best recognition results. Its main advantage is that it is better capable of propagating uncertainty information from the source separation front-end to the HMM speech decoder. The excellent performance of the model with a priori parameters has demonstrated its high potential to perform well in more realistic environments with additive and convolutive noise distortions. In practice, the success of the new model will depend on the accuracy of the estimated evidence PDF parameter set. Coupling the uncertainty propagation with the speech feature extraction process [Kolossa et al., 2005] may provide additional performance gains over the simple estimation techniques presented in this thesis. Furthermore, application of the new model is not limited to multi-channel cocktail-party processing. It is our hope that it will also find use in monaural missing data approaches, where the evidence PDF parameters need to be derived from single-channel enhancement techniques. Another area of interest are missing feature reconstruction methods. Whilst this thesis has focused exclusively on marginalization techniques, data impu-
4. Discussion and conclusions

...gestation strategies [Raj and Stern, 2005; Seltzer et al., 2004a; Srinivasan and Wang, 2007] have been shown to provide additional recognition gains by first estimating the missing spectrogram parts and then performing decoding in the cepstral feature domain. However, performance improvements may only be expected when the reconstruction of the missing feature parts is accomplished with sufficient accuracy. In this respect, it seems possible that the data imputation techniques developed in Raj [2000] can be modified to incorporate the additional information provided by the bounded-Gauss-Uniform model for increased reconstruction accuracy.

Lastly, a large amount of research is devoted to the search of noise resistant features or feature combinations for robust ASR. An important finding of this thesis suggests that the fusion of different feature sets may only be beneficial when an appropriate evidence model is used for the combination. As demonstrated in Article 4, a simple combination of filterbank energy and frequency filtered features did not result in improved recognition performance. Only when feature reliability was taken into consideration through an appropriate evidence model the recognition performance of the combined feature streams was superior. So far, relatively few studies have investigated feature combinations within the framework of missing data (see Palomäki et al. [2006] and Jančovič and Köküer [2008] for two examples). In light of these reports and the results presented here, further research into fusion of feature sets is clearly warranted. If applicable, future studies may also revisit previously unsuccessful combination attempts in order to verify that the reported conclusions hold when feature combination is performed within the missing data framework.

4.2 Limitations

As with all research the findings of this thesis need to be interpreted with regards to the limitations of the study. The following points summarize the restrictions applicable to the work presented in this thesis:

1. With respect to the cocktail party problems investigated in this thesis several unrealistic assumptions about the environment were made in order to make the problem tractable for an algorithmic solution. First of all, we assumed that the number of sources present in the scene is known to the fuzzy cluster algorithms developed in Article 2 and 3. The spatial feature extraction also requires knowledge about the distances between each microphone pair. Only linear arrays with two or six sensors were used in this study. We believe that these relatively mild assumptions do not question the validity of our findings and we have discussed some strategies for relaxing them in Article 2 and 3.
2. We also made a number of more serious assumptions that certainly limit the generalizability of the results. One of them is concerned with the speaker positions in the room. Although the developed source localization algorithm in Article 2 is capable of determining the source azimuths automatically, it requires the sources to remain at fixed room positions throughout the course of an experiment. Tracking multiple moving sources is a considerably more difficult problem that falls outside the scope of this thesis but needs to be addressed in the future.

Another limitation related to this issue is the fact that the current implementation is only capable of batch processing. Dealing with moving sources or aiming for real-time implementations would certainly impose challenging restrictions on the amount of data available for clustering and estimation of the spatial filter weights.

Lastly, we acknowledge that the data used in this study was synthetically mixed using a room image model for simulating sound reflections. Although quite a challenging setup the simple 'shoe box' model does not fully represent real world environments, where many more interfering sources are scattered in the acoustic space and sound reflections are far more complex than any computational procedure can model.

3. Another limiting factor of this study is the type of cocktail party mixtures investigated. We mainly focused on sound mixtures, where a male target speaker was corrupted either by another male or female speaker or by rock music. More realistic mixtures should also include diffuse background noises resulting, for instance, from air-conditioning or computer fan-ventilation. In order to deal with these non-directional distortions further adjustments to the source separation algorithm are necessary.

4. The automatic speech recognition experiments conducted in this work were performed with the help of a missing data speech recognizer. It is important to point out that the achieved results only hold for a low vocabulary task, such as connected digit recognition. The observed gains in recognition accuracy were achieved purely through an improved acoustic modeling framework. Other areas of natural language processing, like language modeling, will gain increasingly in importance for more complex recognition tasks.

4.3 Suggestions for future research

For the presented system to be deployed in more realistic scenarios several issues remain to be addressed in future research. Some starting points for improving the particular techniques developed in this work have already been suggested in the relevant publications.
in the last part of this thesis. Rather than repeating these suggestions here, we comment on two fundamental extensions to our system before adopting a broader perspective that looks at the future of ASR technology in the context of human-machine interaction.

**Statistical cocktail-party processing**

One key argument in this thesis was to model data uncertainty in order to improve the accuracy of the speech recognition process. We resorted to the concept of evidence modeling in order to provide the HMM speech decoder with a probabilistic description of the incoming data streams. The evidence PDFs were estimated here with rather ad-hoc methods based on the outcomes of a deterministic source separation algorithm. A theoretically more satisfying approach can be envisaged through the application of statistical cocktail-party processors as recently investigated by Nix [2005]. In his study, he applied multidimensional statistical estimation techniques to model the spectro-temporal dynamics of speech in a stochastic state-space framework. This approach naturally deals with incomplete or uncertain data and by providing a probabilistic description as separation result establishes a direct link between source separation and evidence PDF estimation. Furthermore, statistical source separation would easily facilitate the integration of inhomogeneous multidimensional feature sets. Despite the appealing properties of multidimensional statistical approaches a word of caution is in order. Computational demands are extremely high for some techniques and quickly become prohibitive for higher feature dimensions [Nix, 2005].

**Fusion of information sources**

Another promising avenue for future research is to consider the fusion of bottom-up and top-down information. There is ample evidence in the psychoacoustic and neurophysiological literature that supports the use of higher-level expectations in human auditory processing [McAdams, 1993]. Although McAdams acknowledges that existing auditory models differ in their degree of feedback, he also emphasizes the role of top-down information in separating speech from background noise. Yet, the overwhelming majority of systems tackle the problem of sound-separation independently from later recognition stages\(^1\). Speech is separated from interfering noise based on primitive low-level features without taking into account information from learned experience and other knowledge sources.

The proposed system in this thesis is no exception. It solely relies on the source separation and the subsequent evidence PDF estimation stage to identify and accurately

---

\(^1\)A more general critique on treating the problem of sound-separation isolated from later recognition stages can be found in Slaney [1995] and Ellis [1996].
describe the noise corruption within a given feature vector. Despite the progress made in recent missing data techniques, the HMM decoder remains highly vulnerable to mismatched components wrongfully declared as reliable by the preprocessing. In a sense, our system still suffers from the same shortcoming as traditional enhancement-recognition schemes. The robustness issue is still being "outsourced" to the preprocessing and the speech decoder itself has no means to detect or correct any residual mismatch present in the incoming data. In search of alternatives, it seems possible that the decoder itself is able to contribute significant information about the feature reliability. Contrary to the source separation stage, the speech decoder has detailed access to clean speech HMMs. Especially for missing data decoding with spectral features this top-down knowledge could be exploited during the marginalization process for downweighting low energy [Cranen and de Veth, 2004] or non-discriminative data components [Pullella et al., 2008], multi-pass hypothesis testing [Srinivasan and Wang, 2005] or actively searching for speech fragments [Barker et al., 2005] or parts of the spectrum that best match the trained models [Ming, 2006]. Alternatively, and in accordance with McAdams model, it is advisable to make overall observations of the data and then let that knowledge flow back top-down for the purpose of adapting the low-level signal analysis. A successful example of such a technique can be found in [Seltzer et al., 2004b], where an adaptive microphone array processor is optimized based on the most likely word hypothesis of a speech recognizer.

**Human-Machine communication**

Ultimately, speech recognition and source separation should work together under a unified framework enabling the overall system to exploit all available knowledge sources for optimal performance in a variety of environmental conditions. Problem solving in complex environments is generally characterized by unexpected situations outside of the machine’s knowledge base and requires the ability to detect and correct mistakes during individual information processing stages. Only when we depart from the passive statistical pattern matching framework that is dominant in current ASR technology and begin the transition to more active decision making paradigms can we expect to succeed in our quest for truly robust speech recognition systems. Put in a broader context, robust speech recognition may only be one part in an automated human-machine communication process involving other modalities such as image and video streams. The human part of this interface is well adept at integrating and fusing multi-modal sensory inputs. On the other hand, the process of emulating the human ability to communicate by sight, sound, and touch with machines is still in its infancy and extensive research efforts worldwide are under way to close this gap.
4. Discussion and conclusions

References


D. Kolossa, A. Klimas, and R. Orglmeister. Separation and robust recognition of noisy,
REFERENCES

convolutive speech mixtures using time-frequency masking and missing data techniques, 2005.


Part III

Publications
Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition

1 Introduction

In order to deploy automatic speech recognition (ASR) effectively in real world scenarios it is necessary to handle hostile environments with multiple speech and noise sources. One classical example is the so-called "cocktail-party problem" [Cherry, 1953], where a number of people are talking simultaneously in a room and the ASR task is to recognize the speech content of one or more target speakers amidst other interfering sources. Although the human brain and auditory system can handle this everyday problem with ease it is very hard to solve with computational algorithms. Current state-of-the-art ASR systems are trained on clean single talker speech and therefore inevitably have serious difficulties when confronted with noisy multi-talker environments.

One promising approach for noise robust speech recognition is based on the missing data automatic speech recognition (MD-ASR) paradigm [Cooke et al., 2001]. MD-ASR requires a time-frequency (T-F) mask indicating the reliability of each feature component. The classification of a partly corrupted feature vector can then be performed on the reliable parts only, thus effectively ignoring the components dominated by noise. If the decision about the reliability of the spectral components can be made with absolute certainty, missing data systems can achieve recognition performance close to clean conditions even under highly adverse signal-to-noise ratios (SNRs) [Cooke et al., 2001; Raj and Stern, 2005; Wang, 2005].

The most critical part in the missing data framework is the blind estimation of the feature reliability mask for arbitrary noise corruptions. The remarkable robustness of the human auditory system inspired researchers in the field of computational auditory scene analysis (CASA) to attempt auditory-like source separation by using an approach based on human hearing. CASA systems first decompose a given signal mixture into a highly redundant T-F representation consisting of individual sound elements/atoms. These elementary atoms are subsequently arranged into separate sound streams by applying a number of grouping cues such as proximity in frequency and time, harmonicity or common location [Bregman, 1990; Brown and Cooke, 1994; Wang, 2005]. The output of these grouping mechanisms can often be represented as a T-F mask which separates the target
from the acoustic background. Essentially, T-F masking provides a link between speech separation and speech recognition [Cooke et al., 2001; Wang, 2005].

Most previous work related to missing data mask estimation is based on single-channel data (see Cerisara et al. [2007] for a review) and relies on SNR criteria [Barker et al., 2000; Cooke et al., 2001; El-Maliki and Drygajlo, 1999], harmonicity cues [Hu and Wang, 2004; van Hamme, 2004] or cue combinations [Seltzer et al., 2004a]. Alternatively, binaural CASA models [Harding et al., 2006; Kim and Kil, 2007; Roman et al., 2003] exploit interaural time and intensity differences (ITD)/(IID) between two ears for missing data mask estimation. While used in the CASA community for quite some time, the concept of T-F masking has recently attracted some interest in the field of blind signal separation (BSS) [Araki et al., 2005; Yilmaz and Rickard, 2004]. Similar to CASA, these methods exploit the potential of T-F masking to separate mixtures with more sources than sensors. However, the BSS problem is tackled from a signal processing oriented rather than psychoacoustic perspective. This, for instance includes the use of multiple sensor pairs [Araki et al., 2007] and statistical approaches such as Independent Component Analysis [Hyvärinen, 1999; Kolossa et al., 2006].

This chapter presents a scheme which combines BSS with robust ASR through the systematic application of T-F masking for both speech separation and speech recognition (Fig. 1). The outlined approach summarizes our previous work reported in Kühne et al. [2007a,b]. In particular, we investigate the performance of a recently proposed BSS method called DUET [Yilmaz and Rickard, 2004] as front-end for missing data speech recognition. Since DUET relies on T-F masking for source demixing, this combination

![Diagram](image)

**Figure 1:** Flowchart for proposed combination of DUET source separation and missing data speech recognition.
arises as a natural choice and is straightforward to implement. In Kühne et al. [2007a] an approach was presented that avoids DUET’s source reconstruction step and directly uses the mask together with the spectral mixture as input for the speech decoder. In subsequent work [Kühne et al., 2007b], a simple but effective mask post-processing step was introduced in order to remove spurious T-F points that can cause insertion errors during decoding. Our proposed combination fits seamlessly into standard feature extraction schemes [Young et al., 2006], but requires a modification of the decoding algorithm to account for missing feature components. It is particularly attractive for ASR scenarios where only limited space and resources for multi-channel processing are available (e.g., mobile phones).

The effectiveness of the proposed BSS-ASR combination is evaluated for a simulated cocktail-party situation with multiple speakers. Experimental results are reported for a connected digits recognition task. Our evaluation shows that, when the assumptions made by DUET hold, the estimated feature reliability masks are comparable in terms of speech recognition accuracy to the oracle masks obtained with a prior knowledge of the sources. We further demonstrate that a conventional speech recognizer fails to operate successfully on DUET’s resynthesized waveforms, which clearly shows the merit of the proposed approach.

The remainder of this chapter is organized as follows: Section 2 briefly reviews the DUET source separation method and outlines its main assumptions. Section 3 explains the methods used for feature extraction and missing data mask generation in more detail. Section 4 presents the experimental evaluation of the system. Section 5 gives a general discussion and illustrates the differences and similarities with a related binaural CASA segregation model. The section further comments on some of the shortcomings in the proposed approach. Finally, the chapter concludes in Section 6 with an outlook on future research.

2 Source separation

This section presents a short review of the DUET-BSS algorithm used in this study for blind separation of multiple concurrent talkers. We start with an introduction of the BSS problem for anechoic mixtures and highlight the main assumptions made by the DUET algorithm. After briefly outlining the main steps of the algorithm, the section closes with a short discussion on why the reconstructed waveform signals are not directly suitable for conventional speech recognition. For a more detailed review of DUET the reader is referred to Yilmaz and Rickard [2004] and Rickard [2008].
2.1 Anechoic mixing model

The considered scenario uses two microphone signals \( x_1(t) \) and \( x_2(t) \) to capture \( N \geq 2 \) speech sources \( s_1(t), \ldots, s_N(t) \) assuming the following anechoic mixing model

\[
x_m(t) = \sum_{j=1}^{N} a_{mj} s_j(t - \delta_{mj}), \quad m = 1, 2
\]

(1)

where \( a_{mj} \) and \( \delta_{mj} \) are the attenuation and delay parameters of source \( s_j \) at microphone \( x_m \). The goal of any BSS algorithm is to recover the source signals \( s_j(t), j = 1, \ldots, N \) using only the mixture observations \( x_m(t), m = 1, 2 \). The mixing model can be approximated in the Short-Time-Fourier-Transform (STFT) domain as an instantaneous mixture at each frequency bin \( l \) through

\[
X_m(k, l) \approx \sum_{j=1}^{N} a_{mj} e^{-i\omega_0 \delta_{mj}} S_j(k, l).
\]

(2)

The STFT transform \( S(k, l) \) for a time domain signal \( s(t) \) is defined as

\[
S(k, l) := \sum_{\tau=-T/2}^{T/2-1} w(\tau) s(\tau + k\tau_0) e^{-i\omega_0 \tau},
\]

(3)

where \( \tau_0 \) and \( \omega_0 \) specify the time-frequency grid resolution and \( w(\tau) \) is a window function (e.g., Hamming, Hanning) of size which attenuates discontinuities at the frame edges. The instantaneous BSS problem can be solved quite elegantly in the frequency domain due to the sparsity of time-frequency representations of speech signals. DUET proceeds by considering the following STFT ratio

\[
\frac{X_2(k, l)}{X_1(k, l)} = \sum_{j=1}^{N} \frac{a_{2j} e^{-i\omega_0 \delta_{2j}} S_j(k, l)}{\sum_{j=1}^{N} a_{1j} e^{-i\omega_0 \delta_{1j}} S_j(k, l)},
\]

(4)

where the nominator and denominator are weighted sums of complex exponentials representing the delay and attenuation of the source spectra at the two microphones.

2.2 Assumptions

The key assumption in DUET is that speech signals satisfy the so-called W-disjoint orthogonality (W-DO) requirement

\[
S_i(k, l) S_j(k, l) = 0, \forall i \neq j, \forall k, l.
\]

(5)
also known as "sparseness" or "disjointness" condition with the support of a source $S_j$ in the T-F plane being denoted as $\Omega_j := \{(k, l) : S_j(k, l) \neq 0\}$. The sparseness condition (5) implies that the supports of two W-DO sources are disjoint, e. g., $\Omega_i \cap \Omega_j = \emptyset$. This motivates a demixing approach based on time-frequency masks, where the mask for source $S_j$ corresponds to the indicator function for the support of this source:

$$M_j(k, l) = \begin{cases} 1, & \text{if } (k, l) \in \Omega_j \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

It has previously been shown [Roman et al., 2003; Wang, 2005; Yilmaz and Rickard, 2004] that binary time-frequency masks exist that are capable of demixing speech sources from just one mixture with high speech fidelity. For example, Wang [2005] proposed the notion of an ideal/oracle binary mask

$$O_j(k, l) := \begin{cases} 1, & \text{if } 20 \log_{10} \frac{|S_j(k, l)|}{\sum_{i \neq j} |S_i(k, l)|} \text{dB} \geq 0 \text{ dB} \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

which determines all time-frequency points where the power of the source exceeds or equals the power of the sum of all interfering sources (see Wang [2005] for a more detailed motivation of the ideal binary masks). Note that these masks can only be constructed if the source signals are known prior to the mixing process as they are defined by means of a SNR criterion. Instead, DUET relies on spatial cues extracted from two microphones to estimate the ideal binary mask. It solely depends on relative attenuation and delays of a sensor pair and assumes an anechoic environment where these cues are most effective. An additional assumption requires that the attenuation and delay mixing pairs for each source are unambiguous.

### 2.3 Estimation of relative mixing parameters using DUET

Due to (5) it follows, that only one arbitrary source $S_j$ will be active at any T-F point such that (4) simplifies to

$$\frac{X_2(k, l)}{X_1(k, l)} = \frac{a_{2j}}{a_{1j}} e^{-il\omega_0(\delta_{2j} - \delta_{1j})} = a_j e^{-il\omega_0 \delta_j}, \quad \forall (k, l) \in \Omega_j \quad (8)$$

with $a_j$ and $\delta_j$ denoting relative attenuation and delay parameters between both microphones and ($\delta_j \neq \delta_k$, $\forall j \neq k$). The goal is now to estimate for each source $S_j$ the corresponding mixing parameter pair $(a_j, \delta_j)$ and use this estimate to construct a time-frequency mask that separates $S_j$ from all other sources. An estimate of the attenuation and delay parameter at each T-F point is obtained by applying the magnitude and
Linking Blind Source Separation and Robust Speech Recognition

Phase operator to (8) leading to

$$\tilde{a}(k,l) := \left| \frac{X_2(k,l)}{X_1(k,l)} \right|, \quad \tilde{\delta}(k,l) := -\frac{1}{l\omega_0} \arg\left( \frac{X_2(k,l)}{X_1(k,l)} \right). \quad (9)$$

If the sources are truly W-DO then accumulating the instantaneous mixing parameter estimates in (9) over all T-F points will yield exactly $N$ distinct $(\tilde{a}, \tilde{\delta})$ pairs equal to the true mixing parameters:

$$\bigcup_{(k,l)} \{(\tilde{a}(k,l), \tilde{\delta}(k,l))\} = \{(a_j, \delta_j) : j = 1, \ldots, N\} \quad (10)$$

The demixing mask for each source is then easily constructed using the following binary decision

$$M_j(k,l) := \begin{cases} 1, & \text{if } (\tilde{a}(k,l), \tilde{\delta}(k,l)) = (a_j, \delta_j) \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

However, in practice the W-DO assumption holds only approximately and it will no longer be possible to observe the true mixing parameters directly through inspection of the instantaneous estimates in (9). Nevertheless, one can expect that the values will be scattered around the true mixing parameters in the attenuation-delay parameter space. Indeed, it was shown in Yilmaz and Rickard [2004] that T-F points with high power possess instantaneous attenuation-delay estimates close to the true mixing parameters. The number of sources and their corresponding attenuation-delay mixing parameters are then estimated by locating the peaks in a power weighted $(\tilde{\alpha}, \tilde{\delta})$-histogram (see Fig. 2), where $\tilde{\alpha}(k,l) := \tilde{a}(k,l) - (\tilde{a}(k,l))^{-1}$ is the so-called symmetric attenuation [Yilmaz and Rickard, 2004]. The peak detection was implemented using a weighted k-means algorithm as suggested in Harte et al. [2005].

### 2.4 Time-Frequency mask construction and demixing

Once the peak locations $(\tilde{\alpha}_j, \tilde{\delta}_j), j = 1, \ldots, N$ have been determined, a second pass over the raw data set is required to assign each observation to one of the detected source locations. We used simple minimum distance classification to construct the binary T-F mask for source $S_j$ as

$$\hat{M}_j(k,l) := \begin{cases} 1, & \text{if } j = \arg\min_z d_z^2(k,l) \\ 0, & \text{otherwise}, \end{cases} \quad (12)$$
2. Source separation

Symmetric attenuation $\alpha$
Relative delay $\delta$

(a)

Figure 2: Power weighted attenuation-delay histogram (a) for a mixture of three sources with mixing parameters $\{(\alpha_1; \delta_1), (\alpha_2; \delta_2), (\alpha_3; \delta_3)\} = \{(-0.03; 0.94), (0; 0), (0.03; -0.94)\}$ and (b) the estimated time-frequency masks with selected points marked in black.

where $d_z^2$ is the squared Euclidean distance

$$d_z^2(k, l) = (\hat{\alpha}(k, l) - \tilde{\alpha}_z)^2 + (\hat{\delta}(k, l) - \tilde{\delta}_z)^2$$

between the instantaneous mixing parameter estimate $(\hat{\alpha}(k, l), \hat{\delta}(k, l))$ and the histogram peak $(\tilde{\alpha}_z, \tilde{\delta}_z)$. The demixing then proceeds by masking the maximum likelihood combination $X_{ML}(k, l)$ of both mixtures [Yilmaz and Rickard, 2004] to obtain the source estimate as

$$\hat{S}_j(k, l) = \hat{M}_j(k, l) \left( \frac{X_1(k, l) + \hat{a}_j e^{i\omega_0 \hat{\delta}_j} X_2(k, l)}{1 + \hat{a}_j^2} \right).$$

Note that for the maximum likelihood combination of both mixtures the symmetric attenuation parameter was converted back to the relative attenuation parameter $\hat{a}_j$. $\hat{S}_j(k, l)$ can then be converted back into the time domain by means of an inverse STFT transformation. However, here we are interested in evaluating the DUET demixing performance using an automatic speech recognizer. The reconstructed time domain signal $\hat{s}_j(t)$ will not be directly applicable for conventional speech recognition systems because non-linear masking effects due to $\hat{M}_j$ are introduced during waveform resynthesis. Conventional speech recognizers perform decoding on complete spectra and can not deal with partial spectral representations. Therefore, additional processing steps, either in the form of data imputation to reconstruct missing spectrogram parts [Raj and Stern, 2005] or missing data marginalization schemes Cooke et al. [2001] that can handle partial data during decoding, are required before speech recognition can be attempted.
In this work the latter option was chosen allowing us to avoid source reconstruction and directly exploit the spectrographic masks for missing data decoding. After source separation the missing data recognizer was informed which mask corresponded to the target speaker by comparing the detected histogram peaks with the true mixing parameters. However, the high STFT resolution is usually not suitable for statistical pattern recognition as it would lead to very high-dimensional feature vectors. The following section explains how the results of the DUET separation can be integrated into standard feature extraction schemes and be utilized for missing data speech recognition.

3 Automatic speech recognition with missing data

A Hidden Markov Model (HMM) based missing data speech recognizer [Cooke et al., 2001] was used for all speech recognition experiments reported in this study. While the HMMs are trained on clean speech in exactly the same manner as in conventional ASR the decoding is treated differently in missing data recognition. Additionally to the feature vector sequence a mask is required to declare each feature component as reliable or unreliable using a hard or soft decision [Barker et al., 2000; Morris et al., 2001]. This section starts with a detailed description of the extracted acoustic features and how the DUET masks can be utilized for missing data recognition. A mask post-processing step is introduced in order to remove isolated mask points that can cause insertion errors in the speech decoding process. We then proceed with the missing data decoding and explain how observation likelihoods are computed in the presence of missing feature components.

3.1 Feature extraction

It is known that the human ear resolves frequencies by grouping several adjacent frequency channels into so-called critical bands [Moore, 2003]. For speech recognition purposes the linear STFT frequency resolution is usually converted to a perceptual frequency scale, such as the bark or mel scale [Moore, 2003; Young et al., 2006]. A widely used approximation of the non-linear frequency resolution of the human auditory system is the mel-frequency scale

\[ f_m(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right), \]

where \( f \) denotes the linear frequency in Hz and \( f_m \) is the corresponding non-linear frequency scale in mel. The grouping of individual frequency channels into critical bands can be accomplished by applying a triangular mel-filterbank to the magnitude or power
FFT spectrum [Young et al., 2006]. The triangular filters

\[
\lambda_b(l) = \begin{cases} 
0 & \omega_0 \leq \omega_{c(b-1)}^b, \\
\frac{\omega_0 - \omega_{c(b-1)}^b}{\omega_{c(b)}^b - \omega_{c(b)}^b} & \omega_{c(b-1)}^b \leq \omega_0 \leq \omega_{c(b)}^b, \\
\frac{\omega_{c(b)}^b - \omega_0}{\omega_{c(b+1)}^b - \omega_{c(b-1)}^b} & \omega_{c(b)}^b \leq \omega_0 \leq \omega_{c(b+1)}^b, \\
0 & \omega_0 > \omega_{c(b+1)}^b, 
\end{cases}
\]  

(16)

with

\[
\omega_{c(b)}^b = 2\pi \cdot 700 \left(10^{f_{cb}/2595} - 1\right)
\]  

(17)

are equally spaced along the mel-frequency scale through

\[
f_{cb} = f_l + b \cdot \frac{f_h - f_l}{B + 1}, \quad b = 1, \ldots, B.
\]  

(18)

Here \(B\) is the number of mel-frequency channels and \(f_l, f_h\) are the lower and higher cut-offs of the mel-frequency axis.

**Acoustic feature extraction**

The preferred acoustic features employed in missing data speech recognition are based on spectral representations rather than the more common mel-frequency-cepstral-coefficients (MFCCs). This is due to the fact that a spectrographic mask contains localized information about the reliability of each spectral component, a concept not compatible with orthogonalized features, such as cepstral coefficients (see also de Veth et al. [2001] for a further discussion). For the scope of this study the extracted spectral features for missing data recognition followed the FBANK feature implementation of the widely accepted Hidden Markov Model Toolkit [Young et al., 2006]. Let \(o_k = (o_{k1}, \ldots, o_{kn})^T\) be the \(n\)-dimensional spectral feature vector at time frame \(k\). The static log-spectral feature components (see Fig. 3(b)) are computed as

\[
o_{kb} = \log \left(\max_l \left\{ \lambda_b(l) |X_{ML}(k,l)|, 1 \right\} \right), \quad b = 1, \ldots, B,
\]  

(19)

where \(\lambda_b\) are the triangular mel-filterbank weights defined in (16) and \(X_{ML}\) is the maximum likelihood combination of both mixture observations as specified in (14). It is common to append time derivatives to the static coefficients in order to model their evolution over a short time period. These dynamic parameters were determined here via the stan-
Figure 3: Spectrograms for the TIDIGITS utterance "3o33951" mixed with three interfering speakers in anechoic condition. (a) linear FFT frequency scale; (b) non-linear mel-frequency scale

dard regression formula

\[ \Delta o_{kb} := o_k(\frac{\theta}{2} + b) = \frac{\sum_{\theta=1}^{\Theta} \theta (o_{(k+\theta)b} - o_{(k-\theta)b})}{2 \sum_{\theta=1}^{\Theta} \theta^2}, \quad b = 1, \ldots, B, \]  

where \( \Delta o_{kb} \) is the regression coefficient at time frame \( k \) and mel-frequency subband \( b \), computed over the corresponding static features using a temporal integration window of size \( \Theta \) [Young et al., 2006]. For this study, only first-order regression coefficients were used, thus producing a feature vector of dimension \( n = 2B \).

Missing data reliability masks

The reliability of each feature component is indicated by a corresponding missing feature mask provided here by the source separation stage. Before converting the mask to the mel-frequency scale we introduce a mask post-processing step to eliminate spurious points in the mask. One important aspect that has not been considered so far is the high correlation of neighboring time-frequency points. That is, if a time-frequency point \((k,l)\) is assigned to speaker \( S_j \) then it is very likely that points in the neighborhood of \((k,l)\) are also belonging to \( S_j \) (see Fig. 5(a)). The DUET method solely relies on the mask assignment in the attenuation-delay parameter space and does not take neighborhood information into account. We observed that for mixtures with more than two sources the time-frequency masks were overlaid by some scattered isolated "noise" points (compare Fig. 5(a), (c)). This type of noise is similar to "shot-noise" known in the image processing community and can be dealt with effectively by means of a non-linear median filter [Russ, 1999]. Similar smoothing techniques have been used previously for missing data mask post-processing [Harding et al., 2005]. For this study, the median filter was preferred.
over other linear filters as it preserves the edges in the mask while removing outliers and smoothing the homogenous regions. The basic operation of a two-dimensional median filter consists in sorting the mask values of a T-F point \((k, l)\) and its neighborhood and replacing the mask value with the computed median \(\bar{M}(k, l)\). Several different neighborhood patterns exist in the literature ranging from 4-nearest neighbors over \(3 \times 3\) or \(5 \times 5\) square neighborhoods to octagonal regions [Russ, 1999]. Here, we used a \(5 \times 5\) plus sign-shaped median filter

\[
\bar{M}(k, l) = \text{median}\{M(u, v) : (u, v) \in N_{(k,l)}\}
\]

with the neighborhood pattern (Fig. 4) defined as

\[
N_{(k,l)} := \{(n, m) : \max\{|n - k|, |m - l|\} \leq 2 \land \min\{|n - k|, |m - l|\} = 0\}.
\]

The filter is able to preserve vertical or horizontal lines that would otherwise be deleted by square neighborhoods. This is important in our application as these lines are often found at sound onsets (vertical, constant time) or formant frequency ridges (horizontal, constant frequency). Other more sophisticated rank filters like the hybrid median filter or cascaded median filters have not been considered here but can be found in Russ [1999].

The effect of the median filtering can be observed in Fig. 5(e), where most of the isolated points have been removed while still preserving the main characteristics of the oracle mask (Fig. 5(a)).

The final missing data mask is then obtained by converting the high STFT resolution to the mel-frequency domain. Similar to (19), we apply the triangular mel-weighting
function $\lambda_b$ to obtain a soft mel-frequency mask

$$w_{kb} = \frac{\sum_l \lambda_b(l) M(k, l)}{\sum_l \lambda_b(l)}.$$  \hfill (23)

While the mask (23) is valid for static features only, a reliability mask is also required for the dynamic feature coefficients in (20). The corresponding mask for $\Delta o_{kb}$ was determined based on the static mask values as

$$\Delta w_{kb} := w_{k(\frac{a}{2}+b)} = \prod_{\theta = \Theta_0, \theta \neq 0} w_{i(k+\theta)b}.$$ \hfill (24)

### 3.2 HMM observation likelihoods with missing features

In this study a HMM based missing data recognizer was used for scoring the n-dimensional spectro-temporal feature vectors described in Section 3.1. The HMM state output distributions were modeled via Gaussian mixture models (GMMs) with diagonal covariance matrices. Let the GMM model parameters for a particular HMM state $q$ be denoted as $\Lambda_q = \{c_q, \mu_q, \sigma^2_q\}$, where the three components represent the mixture weights, mean and variance vectors of the Gaussian mixture probability density function. For a GMM with $R$ mixtures the emission likelihood of $o_k$ for HMM state $q$ is given by

$$p(o_k|\Lambda_q) = \sum_{r=1}^R c_{qr} \prod_{i=1}^n p(o_{ki}|\mu_{qri}, \sigma^2_{qri}),$$ \hfill (25)

where in the case of missing or uncertain features $p(o_{ki}|\mu_{qri}, \sigma^2_{qri})$ is evaluated as

$$p(o_{ki}|\mu_{qri}, \sigma^2_{qri}) = w_{ki} G(o_{ki}; \mu_{qri}, \sigma^2_{qri}) + (1 - w_{ki}) \int_{b_{ki}}^{b_{ki}} G(\tilde{o}_{ki}; \mu_{qri}, \sigma^2_{qri}) d\tilde{o}_{ki},$$ \hfill (26)

with $w_{ki}$ denoting the value of the missing data mask at T-F point $(k, i)$, $a_{ki}$ and $b_{ki}$ being the lower and upper integration bound and $G(o_{ki}; \mu_{qri}, \sigma^2_{qri})$ being a univariate Gaussian

$$G(o_{ki}; \mu_{qri}, \sigma^2_{qri}) = \frac{1}{\sqrt{2\pi} \sigma^2_{qri}} \exp \left[-\frac{1}{2} \frac{(o_{ki} - \mu_{qri})^2}{\sigma^2_{qri}} \right],$$ \hfill (27)

with mean $\mu_{qri}$ and variance $\sigma^2_{qri}$. The value of the missing data mask $w_{ki}$ weights the present and missing data contributions with a soft “probability” between 0 and 1 [Barker et al., 2000; Harding et al., 2006]. The likelihood contribution in (26) for the missing static features is evaluated as a bounded integral over the clean static feature probability density by exploiting the knowledge that the true clean speech value is confined to the
4. Experimental evaluation

interval between zero and the observed noisy spectral energy, e.g. \( a_{ki}^{\text{clean}} \in [a_{ki}, b_{ki}] = [0, o_{ki}], \forall i = 1, \ldots, n \). Past research [Cooke et al., 2001; Morris et al., 2001] has shown that bounding the integral in (26) is beneficial as it provides an effective mechanism to incorporate counter-evidence by penalizing models with insufficient spectral energy. However, no bounds on dynamic feature components were utilized here, thus \( a_{ki} \to -\infty \) and \( b_{ki} \to \infty, \forall i = \frac{n}{2} + 1, \ldots, n \).

4 Experimental evaluation

4.1 Setup

Recognizer architecture and HMM model training

The proposed system was evaluated via connected digit experiments on the TIDIGITS database [Leonard, 1984] with a sample frequency of 20 kHz. The training set for the recognizer consisted of 4235 utterances spoken by 55 male speakers. The HTK toolkit [Young et al., 2006] was used to train 11 word HMMs (‘1’-‘9’, ‘oh’, ‘zero’) each with eight emitting states and two silence models (‘sil’, ‘sp’) with three and one state. All HMMs followed standard left-to-right models without skips using continuous Gaussian densities with diagonal covariance matrices and 10 mixture components. Two different sets of acoustic models were created. Both used 20 ms Hamming-windows with 10 ms frame shifts for the STFT analysis. Note that Yilmaz and Rickard [2004] recommend a Hamming window size of 64 ms for a sampling frequency of 16 kHz in order to maximize the W-DO measure for speech signals. However, for the ASR application considered here, the chosen settings are commonly accepted for feature extraction purposes. The first set of HMMs was used as the cepstral baseline system with 13 MFCCs derived from a 32-channel HTK mel-filterbank plus delta and acceleration coefficients and cepstral mean normalization. This kind of baseline has been widely used in missing data ASR evaluations [Cooke et al., 2001; Harding et al., 2006; Morris et al., 2001]. The second model set was used for the missing data recognizer and used spectral rather than cepstral features as described in Section 3.1. In particular, acoustic features were extracted from a HTK mel-filterbank with \( B = 64 \) channels and first order delta coefficients were appended to the static features according to (19) and (20).

Test data set and room layout

The test set consisted of 166 utterances of seven male speakers containing at least four digits mixed with several masking utterances taken from the TIMIT database [Garofolo et al., 1993, see Table 1].
Figure 5: Example of localization masks for the TIDIGITS target source (black) "303951" in a mixture of three competing speakers (white). (a) oracle mask on linear FFT frequency scale; (b) oracle mask on non-linear mel-frequency scale; (c) DUET mask on linear FFT frequency scale; (d) DUET mask converted to non-linear mel-frequency scale; (e) median filtered mask of (c); (f) median filtered DUET mask from (e) converted to non-linear mel-frequency scale
4. Experimental evaluation

<table>
<thead>
<tr>
<th>Table 1: Transcription for six utterances taken from the test section of the TIMIT database.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialect</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>DR5</td>
</tr>
<tr>
<td>DR5</td>
</tr>
<tr>
<td>DR2</td>
</tr>
<tr>
<td>DR8</td>
</tr>
<tr>
<td>DR3</td>
</tr>
<tr>
<td>DR4</td>
</tr>
</tbody>
</table>

The signal-to-interferer ratio (SIR) for each masker was approximately 0 dB. Stereo mixtures were created by using an anechoic room impulse response of a simulated room of size $4 \text{ m} \times 6 \text{ m} \times 3 \text{ m}$ (length x width x height). Two microphones were positioned in the center of the room, 2 m above the ground, with an interelement distance of $d_{\text{mic}} = 1.72 \text{ cm}$ to guarantee accurate phase parameter estimates [Yilmaz and Rickard, 2004]. Fig. 6(a) shows the setup for a single masker scenario and Fig. 6(b) for a multi-speaker scenario with up to six different speech maskers (three male, three female) placed at a distance of $d_{\text{spk}} = 1 \text{ m}$ to the microphones. For testing, the HTK decoder (HVite) was modified according to (26) to incorporate the missing data marginalization framework.

4.2 Results

A number of experiments were conducted to investigate the DUET separation in terms of speech recognition performance. The cepstral baseline measured the decoder’s robustness against speech intrusions by scoring directly on the speech mixture. The missing data system reported the improvements over this baseline obtained by ignoring the spectral parts that are dominated by interfering speakers as indicated by the missing data

![Figure 6](image-url)
The performance in clean conditions (zero maskers) was 99.16% for the cepstral baseline and 98.54% for the spectral missing data system using the unity mask.

**Angular separation between target and masker**

The first experiment used a female TIMIT speech masker to corrupt the target speech signal. The speaker of interest remained stationary at the 0° location while the speech masker was placed at different angles but identical distance to the microphone pair (see Fig. 6(a)). The recognition performance was evaluated for a conventional recognizer and the missing data system using the oracle and estimated soft masks (Fig. 7). Not surprisingly, the oracle mask performed best marking the upper performance bound for the missing data system while the conventional recognizer represented the lower bound. When the speech masker was placed between 45° to 180° angle relative to the target speaker, the estimated mask almost perfectly matched the oracle mask and hence achieved very high recognition accuracy. However, once the spatial separation between masker and target fell below 30° the accuracy score rapidly started to deteriorate falling below that of the cepstral baseline at the lowest separation angles (0° – 5°). The correctness score followed the same trend as the accuracy score but performed better than the baseline for closely spaced sources. For these small angular separations the assumption that the sources possess distinct spatial signatures becomes increasingly violated and the DUET histogram localization starts to fail. The more the sources move together the less spatial information is available to estimate the oracle mask leading to large mask estimation errors. Nevertheless, the oracle masks still exist even when target and masker are placed at identical positions because they depend on the local SNR rather than spatial locations.
4. Experimental evaluation

![Graph showing accuracy and correctness score for different numbers of speech maskers.](image)

**Figure 8:** Speech recognition performance in terms of (a) accuracy and (b) correctness score for different numbers of concurrent speech maskers. A conventional decoder using MFCC features was used to score on the speech mixtures and DUET’s reconstructed target signal. The spectral missing data system performed decoding with the proposed soft reliability mask (DUET+post-processing+mel-scale conversion) and the binary oracle mask.

**Number of concurrent speech maskers**

The second experiment recorded the recognition performance when the target speaker was corrupted by up to six simultaneous TIMIT maskers (Fig. 8). Accuracy and correctness score were measured for the conventional recognizer using as input the speech mixture or the demixed target speaker as generated by DUET. As before, the missing data recognizer used the oracle and estimated soft masks. The number of simultaneously active speech maskers was increased by successively adding one masker after another according to the order shown in Fig. 6(b).

As expected, the conventional recognizer performed very poorly when scoring on the speech mixture. Performance dropped from 99% in clean conditions to 13% for the single speech masker case. Clearly, state-of-the-art cepstral feature extraction alone provides no protection against additive noise intrusions. For all but the single masker case, it also failed to produce significant improvements for the demixed DUET speech signal. In fact, for most conditions scoring on the speech mixture was better than decoding with the demixed DUET output. As discussed in Section 2.4 and 3.1, conventional speech recognizers require complete data and can not deal with masked spectra such as produced by DUET. In contrast, the missing data system is able to handle missing feature components and provided the upper performance bound when using the oracle mask. Performance degraded very gradually with only a 6% decrease between clean conditions and corruption with six speech maskers. The estimated soft missing data masks closely matched the performance of the oracle masks for up to three simultaneously active speech maskers before starting to fall behind. The more speakers are present in the mixture the more the sparseness assumption (5) becomes invalid making an accurate peak detection in the
attenuation-delay histogram increasingly difficult. Indeed, closer inspection of the 5 & 6 masker scenarios revealed that often peaks were overlapping and the peak detection algorithm failed to identify the locations correctly. For example, once the fifth masker was added, we observed in some cases that the histogram showed only four distinct peaks instead of five. This occasionally led the peak detection algorithm to place the fifth peak near the target speaker location. Due to DUET’s minimum distance classification the wrongly detected speaker location absorbed some of the T-F points actually belonging to the target speaker. Consequently, performance dropped significantly for the 5 & 6 masker configurations, as evident from Fig. 8. Results can be improved somewhat by using soft assignments [Araki et al., 2006b; Kühne et al., 2007b] instead of the winner-takes-it-all concept utilized for the mask construction in (12).

Mask post-processing

The last experiment investigated the influence of the proposed mask post-processing for a four speaker configuration (three maskers). To underline the importance of the mask smoothing the recognition performance with and without the proposed two-dimensional median filtering was measured (see Table 2). In order to eliminate the effect of the histogram peak detection the true mixing parameters were directly passed to the mask construction and no source localization was performed.

Clearly, if no median smoothing is applied to the DUET masks the recognized digit hypotheses contained a high number of insertion and substitution errors. Over 70% of the observed insertions were caused by the digit models "oh" and "eight". With the proposed median smoothing technique both the insertion and substitution errors were dramatically reduced resulting in an improved recognition performance.

Table 2: Recognition results in terms of HTK correctness (COR) and accuracy (ACC) score for missing data masks with and without median smoothing. The number of insertions (INS), deletions (DEL) and substitutions (SUB) is also given.

<table>
<thead>
<tr>
<th>Mask type</th>
<th>COR %</th>
<th>ACC %</th>
<th>DEL</th>
<th>SUB</th>
<th>INS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without mask smoothing</td>
<td>88.62</td>
<td>75.37</td>
<td>17</td>
<td>92</td>
<td>127</td>
</tr>
<tr>
<td>With mask smoothing</td>
<td>94.57</td>
<td>93.53</td>
<td>12</td>
<td>40</td>
<td>10</td>
</tr>
</tbody>
</table>

5 Discussion

The experimental results reported here suggest that DUET might be used as an effective front-end for missing data speech recognition. Its simplicity, robustness and easy integration into existing ASR architecture are the main compelling arguments for the proposed
model. It also fundamentally differs from other multi-channel approaches in the way it makes use of spatial information. Instead of filtering the corrupted signal to retrieve the sources [Low et al., 2004; McCowan et al., 2000; Seltzer et al., 2004b] the time-frequency plane is partitioned into disjoint regions each assigned to a particular source.

A key aspect of the model is the histogram peak detection. Here, we assumed prior knowledge about the number of speakers which should equal the number of peaks in the histogram. However, for a high number of simultaneous speakers the sparseness assumption becomes increasingly unrealistic and as a consequence sometimes histogram peaks are not pronounced enough in the data set. Forcing the peak detection algorithm to find an inadequate number of peaks will produce false localization results. Ultimately, the algorithm should be able to automatically detect the number of sources visible in the data which is usually denoted as unsupervised clustering. This would indeed make the source separation more autonomous and truly blind. However, unsupervised clustering is a considerably more difficult problem and is still an active field of research [Grira et al., 2004]. Other attempts to directly cluster the attenuation and delay distributions using a statistical framework have been reported elsewhere [Araki et al., 2007; Mandel et al., 2006] and would lead to probabilistic mask interpretations.

A point of concern is the microphone distance $d_{mic}$ that was kept very small to avoid phase ambiguities [Yilmaz and Rickard, 2004]. Clearly, this limits the influence of the attenuation parameter (see Fig. 2(a)). Rickard [2008] has offered two extensions to overcome the small sensor spacing by using phase differentials or tiled histograms. Another option to consider is the use of multiple microphone pairs or sensor arrays allowing for full three-dimensional source localization [Araki et al., 2006a, 2007].

While the proposed median smoothing was highly successful in reducing spurious points in the time-frequency masks the filter was applied as a post-processing step only. Other more sophisticated methods that incorporate neighborhood information already into the mask assignment or the peak detection itself might be more appropriate. In particular, Markov Random Fields [Li, 2001] have been quite successful in the field of image processing but tend to be more complex and demanding in terms of computational resources. Other schemes for incorporating neighborhood information into clustering or mixture model learning are also readily available [Ambroise et al., 1997; Chuang et al., 2006]. The advantage of the proposed post-processing scheme lies in its simplicity and relatively fast computation. Nevertheless, careful selection of the size of the median filter is required as otherwise the filter tends to remove too much energy of the target signal.

In regards to related work the overall architecture of our system is in line with previously proposed binaural CASA models. However, the DUET separation framework differs in some key aspects as it models human hearing mechanisms to a much lesser degree. Whereas Harding et al. [2006] and Roman et al. [2003] perform mask estima-
tion for each critical band using supervised learning techniques, DUET blindly estimates these masks based on a simple frequency independent classification of attenuation and delay parameters. The spatial cues are extracted from STFT ratios which offer significant speedups over computationally expensive cross-correlation functions commonly used to compute binaural ITDs [see also Kim and Kil, 2007, for an efficient method of binaural ITD estimation using zero-crossings]. More importantly, Roman et al. [2003] need to recalibrate their system for each new spatial source configuration which is not required in our model. DUET also directly operates on the mixture signals and does not employ Head-Related-Transfer-Functions (HRTFs) or gammatone filterbanks for spectral analysis. However, we expect supervised source localization schemes to outperform DUET’s simple histogram peak detection when angular separation angles between sources are small (0° – 15°).

In terms of ASR performance we achieved comparable results to Roman et al. [2003], in that the estimated masks matched the performance of the oracle masks. Recognition accuracy remained close to the upper bound for up to three simultaneous speech maskers. While other studies [Mandel et al., 2006; Roman et al., 2003] have reported inferior localization performance of DUET even for anechoic, two or three source configurations we can not confirm these observations based on the experimental results discussed here. Mandel et al. [2006] offer a possible explanation for this discrepancy by stating that DUET was designed for a closely spaced omni-directional microphone pair and not the dummy head recordings used in binaural models.

Finally, we acknowledge that the results presented here were obtained under ideal conditions that met most of the requirements of the DUET algorithm. In particular the noise-free and anechoic environment can be considered as strong simplifications of real acoustic scenes and it is expected that under more realistic conditions the parameter estimation using DUET will fail. Future work is required to make the estimators more robust in hostile environments. To this extent, it is also tempting to combine the DUET parameters with other localization methods [Kim and Kil, 2007] or non-spatial features such as harmonicity cues [Hu and Wang, 2004]. However, the integration of additional cues into the framework outlined here remains a topic for future research.

6 Conclusion

This chapter has investigated the DUET blind source separation technique as a front-end for missing data speech recognition in anechoic multi-talker environments. Using the DUET attenuation and delay estimators time-frequency masks were constructed by exploiting the sparseness property of speech in the frequency domain. The obtained masks were then smoothed with a median filter to remove spurious points that can cause
insertion errors in the speech decoder. Finally, the frequency resolution was reduced by applying a triangular mel-filter weighting which makes the masks more suitable for speech recognition purposes. The experimental evaluation showed that the proposed model is able to retain high recognition performance in the presence of multiple competing speakers. For up to three simultaneous speech maskers the estimated soft masks closely matched the recognition performance of the oracle masks designed with a priori knowledge of the source spectra. In our future work we plan to extend the system to handle reverberant environments through the use of multiple sensor pairs and by combining the T-F masking framework with spatial filtering techniques that can enhance the speech signal prior to recognition.

7 Acknowledgments

This work was supported in part by The University of Western Australia, Australia and in part by National ICT Australia (NICTA). NICTA is funded through the Australian Government’s Backing Australia’s Ability Initiative, in part through the Australian Research Council.

References


J. Barker, L. Josifovski, M. Cooke, and P. Green. Soft decisions in missing data techniques


H. Hyvärinen. Survey on independent component analysis. Neural Computing Surveys, 2:


Robust Source Localization in Reverberant Environments Based on Weighted Fuzzy Clustering

Abstract

Successful localization of sound sources in reverberant enclosures is an important prerequisite for many spatial signal processing algorithms. We investigate the use of a weighted fuzzy c-means cluster algorithm for robust source localization using location cues extracted from a microphone array. In order to increase the algorithm’s robustness against sound reflections we incorporate observation weights to emphasize reliable cues over unreliable ones. The weights are computed from local feature statistics around sound onsets because it is known that these regions are least affected by reverberation. Experimental results illustrate the superiority of the method when compared with standard fuzzy clustering. The proposed algorithm successfully located two speech sources for a range of angular separations in room environments with reverberation times of up to 600 ms.

1 Introduction

Acoustic source localization by means of a microphone array is still an active field of research. It is the task of extracting the localization information of one or several sound sources by sampling the sound field through a number of spatially distinct microphones. One important application for source localization algorithms can be found in the field of blind source separation (BSS). Specifically, under-determined BSS strategies that rely on source sparseness frequently exploit localization information, such as directions-of-arrival (DOA), for segmenting the time-frequency (T-F) plane into disjoint regions each assigned to a particular source [Araki et al., 2007; Yilmaz and Rickard, 2004]. For example, Araki et al. [2007] use the k-means algorithm to perform the partitioning of the T-F plane into a number of clusters, which is assumed to be known a priori. For stationary sources, each of the cluster centers corresponds to the location parameter of a particular source. Each cluster membership matrix indicates to which degree a T-F point belongs to a particular source and can be interpreted as a T-F mask. Individual sources are separated from the mixture by selecting all T-F points belonging to the source’s cluster as indicated by its membership mask. Location cues are most reliable in anechoic environments, where almost all T-F points contribute observations that are effective for
Robust Source Localization in Reverberant Environments

clustering. However, for reverberant mixtures the location cues become increasingly corrupted. This often leads to incorrect localization and partitioning results, mostly due to falsely detected cluster centers, e.g., source locations.

In this letter, we address this issue by presenting a robust source localization technique based on a weighted fuzzy $c$-means algorithm. Contrary to Araki et al. [2007], where every observation is treated as equally important, we deal with imperfect location cues by indicating their usability for the clustering process. The observation weights are obtained prior to the clustering by scanning the T-F plane around sound onsets for regions with low DOA fluctuations. This was motivated by a number of previous studies which have shown that in echoic enclosures only a small fraction of the location cues correspond to the correct source locations. For example, Faller and Merimaa [2004] showed that binaural source localization remains feasible even in highly reverberant conditions by selecting cues consistent with an interaural coherence measure. Huang et al. [1997] emulated the precedence effect by concentrating on location cues extracted from sound onsets, a concept that is known to be exploited by the human auditory system [Litovsky et al., 1999]. Through computer simulations, we show that the proposed algorithm is successful in locating two speech sources for a range of angular separations in room environments with reverberation times ($\text{RT}_{60}$) of up to 600 ms.

The remainder of this letter is organized as follows. Section 2 describes the proposed localization algorithm in more detail and illustrates the importance of observation weighting for reverberant mixtures. Section 3 describes the experimental protocol and presents the results for a number of simulated source localization experiments. Finally, the letter concludes with a short summary in Section 4.

2 Robust Source Localization algorithm

2.1 Mixing model and sparseness assumption

Consider $N$ sources in a reverberant enclosure impinging on a uniform linear microphone array (ULA) made up of $M$ identical, omnidirectional sensors with inter-element spacing $d$ (Fig. 1). The source positions are assumed to be stationary in the median plane at azimuth angles $\theta_1, \ldots, \theta_N$. It is further assumed that each microphone observation can be represented in the frequency domain as an instantaneous mixing model

$$X_m(k, l) \approx \sum_{n=1}^{N} H_{mn}(l) S_n(k, l), \quad m = 1, \ldots, M$$

where $k$ represents a time index, $l$ is a frequency index and $H_{mn}(l)$ is the room impulse response from source $n$ to sensor $m$. $X_m(k, l)$ and $S_n(k, l)$ are the Short-Time Fourier
2. Robust Source Localization algorithm

Transforms (STFTs) of the $m$-th microphone observation and $n$-th source defined on a T-F grid by the lattice spacing parameters $(\tau_0, \omega_0)$. A common assumption for speech signals [Araki et al., 2007; Yilmaz and Rickard, 2004] is that for each T-F point only one arbitrary source $S_n$ will be active, such that (1) simplifies to

$$X_m(k,l) \approx H_{mn}(l) S_n(k,l).$$

(2)

Note that assumptions (1) and (2) become increasingly unrealistic for short STFT window lengths and long reverberation times due to strong reflections from preceding sound events.

2.2 Spatial feature extraction

The most commonly adopted location feature is based on the estimation of the time delay between two microphones using the generalized cross-correlation or cross-power spectrum phase [Knapp and Carter, 1976]. However, within the BSS framework of T-F masking it is more common to use level ratios and/or phase differences which produce instantaneous location estimates for each T-F cell [Araki et al., 2007; Yilmaz and Rickard, 2004].

According to Togami et al. [2007], for sparse sources in echoic environments the longer the distance $d_j$ is between a sensor pair $(X_{j,1}, X_{j,2})$ (Fig. 1) the better the DOA localization performance will be. Hence, the instantaneous DOA at T-F point $(k,l)$ is computed as the normalized phase difference

$$\psi(k,l) = -\frac{1}{l\omega_0 d_{j_{\text{max}}} c^{-1}} \text{arg} \left[ \frac{X_{j_{\text{max}},1}(k,l)}{X_{j_{\text{max}},2}(k,l)} \right],$$

(3)

where $j_{\text{max}}$ denotes the index of the sensor pair with the longest distance $d_{j_{\text{max}}}$ and $c$ is the propagation velocity of sound [Araki et al., 2007; Togami et al., 2007]. However, when
$d_{\text{max}} > \frac{c}{2f_{\text{max}}}$, with $f_{\text{max}}$ being the signal’s maximum frequency, the sensor pair violates the spatial aliasing theorem and the phase values in (3) become ambiguous. To avoid this problem, we employ the SPIRE algorithm [Togami et al., 2007] which utilizes the smaller non-aliased distance pairs to restore the aliased values of the longer distance pairs. SPIRE is applicable when multiple microphone pairs are available and at least one sensor-pair distance is shorter than the aliasing distance. Note that $\psi$ can be converted to its equivalent azimuth angle via $\theta = \arcsin(\psi)$. The frequency normalization in (3) avoids the permutation problem usually encountered in frequency domain BSS and ensures that for short data enough DOA measurements are available for clustering [Araki et al., 2007].

### 2.3 Fuzzy clustering of DOA values

A weighted fuzzy $c$-means (wFCM) algorithm [Miyamoto et al., 2006] is then used for grouping the extracted features $\psi$ into $N$ clusters. In wFCM clustering is achieved by minimizing the cost function

$$J = \sum_{\forall (k,l)} \sum_{n=1}^{N} u_{n}^q(k,l)w(k,l) \| \psi(k,l) - \hat{\psi}_n \|^2, \quad (4)$$

where $u_{n}(k,l) \in [0, 1]$ represents the membership of $\psi(k,l)$ in the $n$th cluster, $w(k,l)$ is the observation weight for $\psi(k,l)$, $\hat{\psi}_n$ is the $n$th cluster center and $\| \cdot \|$ is a distance metric, such as the $L_2$ norm. The parameter $q > 1$ controls the softness of the clustering and is fixed here to $q = 2$. Starting from a random partitioning, the cost function (4) is iteratively minimized by alternating the updates for centers and memberships

$$\hat{\psi}_n = \frac{\sum_{\forall (k,l)} u_{n}^q(k,l)w(k,l)\psi(k,l)}{\sum_{\forall (k,l)} u_{n}^q(k,l)w(k,l)}, \quad (5)$$

$$u_{n}(k,l) = \left[ \sum_{j=1}^{N} \left( \frac{\| \psi(k,l) - \hat{\psi}_n \|}{\| \psi(k,l) - \hat{\psi}_j \|} \right)^{\frac{2}{q-1}} \right]^{-1}, \quad (6)$$

until an appropriate termination criterion is met. While the final centroids correspond to the DOA estimates the membership matrices can be used as soft T-F masks in BSS applications [Araki et al., 2007; Yilmaz and Rickard, 2004]. Note that wFCM defaults to the standard FCM clustering if the weights are chosen to be unity.

### 2.4 Observation weights

The observation weights $w$ are used to emphasize sound onsets and regions with low DOA fluctuations. The steeper an onset and the lower the local variance for a DOA
measurement, the more weight should be given to this observation during clustering. In
particular, the weights are computed as

\[ w(k, l) = (l\omega_0)^2 w_{\text{ons}}(k, l) w_{\text{var}}(k, l), \]  

where \( w_{\text{ons}} \) denotes the onset component and \( w_{\text{var}} \) are the weights based on the local
DOA variance around \( \psi(k, l) \).

The motivation for (7) is that reliable DOA measurements are often found at echo-free
sound onsets [Huang et al., 1997] and single source areas with low local DOA variances
[Abrard and Deville, 2005]. High variances, however, indicate regions where sources
overlap or where reflections contaminate the DOA measurements. The additional term
\( (l\omega_0)^2 \) gives more weight to high frequencies because the localization accuracy in low
frequencies is often severely degraded by reverberation. The weighting mechanism is
particularly helpful in reverberant conditions because it favors T-F points that better sat-
ify the sparseness assumption (2).

To determine \( w_{\text{ons}} \) a simple onset weighting scheme is implemented. Let the instan-
taneous power be defined as in Yilmaz and Rickard [2004]

\[ E(k, l) = |X_{j_{\text{max}},1}(k, l)X_{j_{\text{max}},2}(k, l)| \]  

and let its first-order time difference on the log-scale be

\[ o(k, l) = \log \left[ \frac{E(k, l)}{E(k-1, l)} \right] = \log[E(k, l)] - \log[E(k-1, l)]. \]  

After smoothing \( o \) with a 5 × 5 median filter, the onset weights are determined via the
sigmoid compression

\[ w_{\text{ons}}(k, l) = \frac{1}{1 + \exp\{\alpha_1 [o(k, l) - \beta_1]\}}, \]  

where \( \alpha_1 \) is the sigmoid slope and \( \beta_1 \) is the sigmoid center. Both parameters can be tuned,
such that sound offsets with small or negative \( o(k, l) \) values are suppressed and onsets
with large positive \( o(k, l) \) are emphasized. The second weight component is derived from
local DOA statistics gathered in a small neighborhood \( N_{(k, l)} \) around each DOA measure-
ment. Let the local DOA mean \( \mu_{\psi}(k, l) \) around \( \psi(k, l) \) be

\[ \mu_{\psi}(k, l) = \frac{1}{|N_{(k, l)}|} \sum_{\psi(k', l') \in N_{(k, l)}} \psi(k', l'), \]  

where
and the local DOA variance $\sigma^2_{\psi}(k, l)$ around $\psi(k, l)$ be

$$\sigma^2_{\psi}(k, l) = \frac{1}{|N(k,l)| - 1} \sum_{\forall (k',l') \in N(k,l)} \left| \psi(k', l') - \mu_{\psi}(k, l) \right|^2,$$  

(12)

where the neighborhood $N(k,l) := \{(k', l') : k' = k \land |l' - l| \leq P \}$ is chosen in this study as a nine point window of adjacent frequency bins, e.g., $P = 4$. The variances are then mapped to the $[0, 1]$ interval using the sigmoid function

$$w_{\text{var}}(k, l) = \frac{1}{1 + \exp\left\{\alpha_2 \left(\log[\sigma^2_{\psi}(k, l)] - \beta_2\right)\right\}},$$  

(13)

where $\alpha_2$ and $\beta_2$ are the sigmoid slope and center parameter respectively. Again, both parameters can be tuned, such that T-F points with large DOA variances are suppressed and areas with low DOA fluctuations are emphasized. Optimal selection of the sigmoid parameters is beyond the scope of this paper, and therefore all parameters are empirically derived through a series of tuning experiments (see Section 3).

Fig. 2 illustrates the importance of the proposed observation weighting for a two-source configuration. For anechoic conditions, as shown in Fig. 2(a), almost all of the observations $\psi(k, l)$ in the T-F plane are reliable for clustering. The corresponding azimuth histogram with unity weights in Fig. 2(b) clearly shows two distinctive peaks close to the true DOA angles. However, as evident from Fig. 2(c), in reverberant conditions only a small fraction of the DOA observations remain within a localization error of $5^\circ$. Consequently, the azimuth histogram with unity weights in Fig. 2(d) fails to identify the two sources. Only when the observations are weighted according to their reliability $w(k, l)$ the two sources become visible again (Fig. 2(e) and (f)).

### 3 Experimental Evaluation

**Setup**

Multi-path sound propagation was simulated for a small rectangular room of dimensions $6 \text{ m} \times 4 \text{ m} \times 3 \text{ m}$ (length x width x height). Wall reflections were estimated using the image model method for simulating small-room acoustics [Lehmann and Johansson, 2008]. Room impulse responses for different reverberation times were generated for each sensor of a six-channel ULA with inter-element spacing of $d = 4.28 \text{ cm}$ at a sampling frequency of $8 \text{ kHz}$. The array was positioned in the middle of the room at a height of $2 \text{ m}$. Two speech sources with equal gain were placed at different horizontal angles facing array broadside and a distance of $1.5 \text{ m}$ from the array center. A total of 240 different speech mixtures were constructed for testing with utterances from the TIMIT and
3. Experimental Evaluation

Figure 2: Example of reliable localization information for two sources located at $-10^\circ$ and $10^\circ$ under (a)-(b) anechoic and (c)-(f) reverberant conditions with $RT_{60} = 600$ ms. (a) Anechoic DOA observations $\psi(k, l)$ with max. of $5^\circ$ localization error (black points); (b) Anechoic azimuth histogram with unity weights; (c) Reverberant DOA observations $\psi(k, l)$ with max. of $5^\circ$ localization error (black points); (d) Reverberant azimuth histogram with unity weights; (e) Estimated DOA weights $w(k, l)$; (f) Reverberant azimuth histogram with weights from (e).
Robust Source Localization in Reverberant Environments

TIDIGIT databases. Simulations were run for three different DOA configurations with azimuths of \((\theta_1, \theta_2) \in \{(-20^\circ, 20^\circ), (-10^\circ, 10^\circ), (-5^\circ, 5^\circ)\}\) and three room reverberation times \(RT_{60} \in \{0 \text{ ms}, 300 \text{ ms}, 600 \text{ ms}\}\). The STFT frame size was 25 ms with a shift of 10 ms. Following a range of tuning experiments on a cross-validation set the sigmoid slope parameters were fixed to \(\alpha_1 = -3\) and \(\alpha_2 = 8\). For each utterance, the sigmoid center \(\beta_1\) was fixed to be the 98th percentile of the set \(\bigcup_{(k,l)} \{o(k,l)\}\). Similarly, the center parameter \(\beta_2\) was set to be the \(p\)-th percentile of \(\bigcup_{(k,l)} \{\log[\sigma^2_{\psi}(k,l)]\}\). The value of \(p\) was tuned for each configuration. The stronger the reverberation and the smaller the azimuth separation between sources the lower the percentile was chosen. The localization performance of the fuzzy \(c\)-means algorithm was then measured with and without the proposed observation weighting.

Results and discussion

As expected, for anechoic conditions (Fig. 3, left column) the localization performance using the standard fuzzy clustering was sufficient for all azimuth separation angles. The observation weighting had little effect in these cases as almost all T-F observations produced DOA values close to the true azimuth angles. However, for 300 ms and 600 ms reverberation times (Fig. 3, middle and right column) the standard clustering failed to locate the two sources correctly. Too many observations have become unreliable and have consequently started to bias the clustering towards incorrect solutions. On the other hand, the proposed weighting scheme successfully located both sources for most tested configurations, even for the challenging case of only 10\(^\circ\) angular separation and 600 ms reverberation time (Fig. 3, right column, bottom row).

In terms of limitations, our current implementation is based on a linear array which is restricted to azimuth angle estimation and is subject to front-back confusions. However, for full three-dimensional localization the outlined approach can easily be extended to non-linear array geometries [Araki et al., 2007; Togami et al., 2007]. The major drawback of the current approach is the rather ad-hoc determination of the weighting factors which requires prior knowledge about the environment. Ideally, the sigmoid parameters should be automatically adapted by the algorithm itself. Further research is needed to address these issues and to extend the method to cases with moving speakers and sources with smooth or no onsets.

4 Conclusion

We have presented a weighted fuzzy clustering algorithm for tackling multi-source localization in reverberant environments. In order to increase the algorithm’s robustness observation weights were incorporated to emphasize reliable over unreliable DOA cues.
The weights were derived from local DOA statistics around sound onsets because it is known that these regions are least affected by reverberation. Our experimental evaluation showed that the proposed method produces superior localization results when compared with standard fuzzy clustering, particularly in reverberant conditions. As a consequence, it is expected that the resulting cluster partitions will also lead to better T-F separation masks. In subsequent work, we therefore intend to investigate reverberant BSS problems by integrating our robust source localization scheme into T-F masking and spatial filtering techniques.
References


A Novel Fuzzy Clustering Algorithm using Observation Weighting and Context Information for Reverberant Blind Speech Separation

Abstract

Time-frequency masking has evolved as a powerful tool for tackling blind source separation problems. In previous work, mask estimation was performed with the help of well-known standard cluster algorithms. Spatial observation vectors, extracted from a set of microphones, were grouped into separate clusters, each representing a particular source. However, most off-the-shelf clustering methods are not very robust to outliers or noise in the data. This lack of robustness often leads to incorrect localization and partitioning results, particularly for reverberant speech mixtures. To address this issue, we investigate the use of observation weights and context information as means to improve the clustering performance under reverberant conditions. While the observation weights improve the localization accuracy by ignoring noisy observations, context information smooths the cluster membership levels by exploiting the highly structured nature of speech signals in the time-frequency domain. In a number of experiments, we demonstrate the superiority of the proposed method over conventional fuzzy clustering, both in terms of localization accuracy as well as speech separation performance.

1 Introduction

The goal of any source separation method is the recovery of the sources from a given set of mixture observations. When this problem is tackled without prior knowledge about the mixing process or the original source signals it is usually referred to as blind source separation (BSS). The algorithmic solution to the BSS problem is of great importance in a number of different fields, such as speech processing, seismology, remote sensing, econometrics, medical imaging and communication systems. The classical example for acoustic signals is the so-called "cocktail party problem" [Cherry, 1953], where a number of people talk simultaneously in a room and the task is to extract one or more target speakers amidst other interfering speakers and background noise.

In recent years, the concept of time-frequency (TF) masking has evolved as a popular tool for tackling the BSS problem [Araki et al., 2006c; Mandel et al., 2006; Peterson and
Kadambe, 2003; Sawada et al., 2007; Weiss et al., 2008; Yilmaz and Rickard, 2004]. Separation of the sources is achieved by exploiting a specific property of the sources, called sparsity or sparseness [Yilmaz and Rickard, 2004]. For example, it has been shown that speech signals hardly overlap in their short-time-Fourier transform (STFT) representation, or more formally stated, their STFT supports are approximately disjoint [Jourjine et al., 2000; Yilmaz and Rickard, 2004]. This discovery has motivated a demixing approach, where a particular source is separated from the mixture simply by masking all coefficients not belonging to its STFT support. Because this procedure does not depend on matrix inversion it can be applied even for the under-determined BSS case, e.g., when there are more sources than mixtures.

Several practical algorithms that implement the TF masking concept have been proposed in the literature. One of the first was the degenerate unmixing estimation technique (DUET) for anechoic mixtures [Jourjine et al., 2000; Yilmaz and Rickard, 2004]. DUET operates on stereo data and employs a histogram technique for mask estimation. Later, an extension, called DUET-ESPRIT (DESPRIT) [Melia and Rickard, 2007], was developed in order to handle the echoic mixing case by combining DUET with the estimation of signal parameters via rotational invariance technique (ESPRIT) [Roy and Kailath, 1989]. DESPRIT operates on multi-channel data recorded by a uniform linear microphone array but localization performance is still subject to front-back confusions. The MENUET (Multiple sENsor dUET) algorithm [Araki et al., 2007b] further extended the sensor arrangement to arbitrarily non-linear array geometries allowing for full three-dimensional source localization. In order to fully automate the process of TF mask estimation, the application of well-known cluster algorithms, such as k-means [Hartigan and Wong, 1979], was introduced in Araki et al. [2006a, 2007b]. In this line of research, the observation vectors are grouped into separate clusters, each representing a particular source. If the observation vectors embody spatial information each cluster center represents a location estimate of the source’s position. The cluster memberships, on the other hand, can be interpreted as TF localization masks marking the dominant points of each source in the TF plane. While a hard clustering algorithm, like k-means, produces binary membership masks other research has concentrated on fuzzy clustering methods [Kühne et al., 2009] or the probabilistic expectation-maximization (EM) algorithm for soft TF mask estimation [Mandel et al., 2006; O’Grady and Pearlmutter, 2004; Sawada et al., 2007; Weiss et al., 2008].

Although the introduction of standard clustering techniques has certainly advanced the field of TF masking it is not without its shortcomings. Unfortunately, most off-the-shelf methods, such as k-means and fuzzy c-means (FCM), are not very robust to outliers or noise in the data. This lack of robustness often leads to incorrect localization and partitioning results under reverberant conditions. In previous work [Kühne et al., 2009], we have partly addressed this issue by extending the standard FCM to cope with unreliable
data through the use of observation weights. This weighted fuzzy c-means (wFCM) algorithm improves the localization performance by down-weighting unreliable data points during the cluster centroid computation. However, like FCM and k-means, the wFCM technique completely ignores the highly structured nature of speech signals in the TF domain. The cluster membership value at a particular TF point is still assigned in isolation of its context or surroundings making the TF mask estimation extremely vulnerable to noise.

Context or neighborhood information has long been established as a suitable tool for increasing the noise robustness of image segmentation algorithms [Ambroise et al., 1997; Chuang et al., 2006; Li, 2001; Liew et al., 2000; Pham, 2001; Xia et al., 2007]. Natural images often exhibit a high correlation between neighboring points due to the fact that objects are formed through patches of connected pixels [Russ, 1999]. By the same argument, it is reasonable to assume that the dominant parts of speech signals also form patches and are not randomly scattered across the TF plane. Given the lack of robustness in conventional clustering techniques it seems promising to integrate such a structural constraint into the TF mask estimation procedure for improving the separation performance under noisy and reverberant conditions.

This paper therefore presents a novel weighted contextual fuzzy c-means (wCFCM) clustering technique for the problem of acoustic source separation. Motivated by the success of neighborhood information in medical image segmentation [Pham, 2001], we introduce a novel regularization term into the wFCM objective function in order to model the context information around a TF point. The term "context information" thereby refers to available information about the cluster memberships of adjacent observations gathered from a local TF neighborhood. Using the same technique as in Pham [2001], the strength of the regularization term is then automatically determined by employing a pseudo-cross-validation scheme. The proposed contextually constrained wCFCM algorithm biases the clustering solution towards homogenous TF masks and is shown to be more robust to reverberation than traditional approaches. To the best of the authors' knowledge, this is the first work studying the effect of context information within the framework of clustering-based acoustic source separation. The remainder of this paper is organized as follows. Section 2 starts with a short description of the convolutive BSS problem. Section 3 presents an overview of the system architecture and briefly explains the main signal processing steps involved. In Section 4, we briefly review the FCM and wFCM clustering methods before describing the proposed wCFCM algorithm in more detail. Section 5 reports on our experimental evaluation and demonstrates that in comparison with conventional fuzzy clustering the wCFCM algorithm leads to superior speech separation performance, particularly in reverberant conditions. The section also comments on several limitations in our approach and points out some potential extensions for future work. The paper concludes in Section 6 with a short summary.
2 Problem statement

Consider $N$ sources in a reverberant enclosure impinging on a uniform linear microphone array (ULA) made up of $M$ identical, omnidirectional sensors with inter-element spacing $d$. The sources are positioned stationary in the median plane (Fig. 1) at unknown azimuth angles $\theta_1, \ldots, \theta_N$. It is further assumed that each microphone observation can be modeled as a convolutive sum

$$x_m(t) = \sum_{n=1}^{N} \sum_{p} h_{mn}(p) s_n(t-p), \quad m = 1, \ldots, M$$  (1)

where $x_m(t)$ is the mixture observation at sensor $m$, $s_n(t)$ is the $n$-th source signal and $h_{mn}(p)$ denotes the room impulse response from source $s_n$ to microphone $x_m$. We assume that $N$ and $M$ as well as the sensor spacing $d$ are known and that $d$ is chosen such that no spatial aliasing occurs. The goal is to recover an estimate $\hat{s}_n(t)$ for each source $i \in \{1, \ldots, N\}$ from the $M$ mixture observations $x_m(t)$.

![Figure 1: A uniform linear microphone array with $j \in \{1, 2, 3\}$ sensor pairs $(X_{j,1}, X_{j,2})$ and two sources $S_1$, $S_2$ located at azimuth angles $\theta_1$ and $\theta_2.$](image1)

3 System Overview

This section presents an overview about the source separation system utilized in our study (Fig. 2). We briefly discuss each step of the model before describing the clustering stage in detail in the subsequent section.

![Figure 2: Basic scheme of the time-frequency masking approach for blind source separation.](image2)
Throughout the rest of the paper, the following notations and definitions are adopted:

- \( \arg[\cdot] \) phase of a complex number;
- \( (\cdot)^T \) transpose;
- \( (\cdot)^H \) Hermitian transpose;
- \( (\cdot)^* \) optimal value;
- \( ||\cdot|| \) Euclidean norm;
- \( |\cdot| \) absolute value or cardinality;
- \( \hat{\cdot} \) estimated quantity;
- \( (\cdot) \leftarrow (\cdot) \) replacement of left hand side by right hand side;
- \( \mathbf{j} \) imaginary unit.

### Step 1 - Short time spectral analysis

The first step converts the time domain signals \( x_m(t) \), sampled at frequency \( f_s \), into their STFT representation

\[
X_m(k,l) = \sum_{\tau = -L/2}^{L/2-1} \text{win}(\tau) x_m(\tau + k\tau_0) e^{-j\omega_0\tau}, \tag{2}
\]

where \( k \in \{0, \ldots, K-1\} \) is a time frame index, \( l \in \{0, \ldots, L-1\} \) is a frequency bin index, \( \text{win}(\tau) \) is a window function and \( \tau_0 \) and \( \omega_0 \) are appropriately chosen TF grid resolution parameters. A \( L \)-point Hanning window

\[
\text{win}(\tau) = 0.5 - 0.5 \cos\left(\frac{2\pi \tau}{L}\right), \quad \tau = 0, \ldots, L-1 \tag{3}
\]

was utilized in this paper for attenuating signal discontinuities at the frame edges. By transforming (1) via (2) into the frequency domain, the convolutive BSS problem can be approximated as an instantaneous mixing model

\[
X_m(k,l) \approx \sum_{n=1}^{N} H_{mn}(l) S_n(k,l), \tag{4}
\]

where \( H_{mn}(l) \) is the room impulse response from source \( S_n \) to sensor \( X_m \) at frequency bin \( l \) and \( X_m(k,l) \) and \( S_n(k,l) \) are the STFTs of the \( m \)-th microphone observation and the \( n \)-th source signal, respectively. Another advantage of working in the STFT domain is the ability to exploit the source sparseness by approximating the sum in (4) with

\[
X_m(k,l) \approx H_{mn}(l) S_n(k,l), \quad \forall n \in \{1, \ldots, N\} \tag{5}
\]

where \( S_n(k,l) \) is the dominant source at TF cell \( (k,l) \). While this assumption holds well for anechoic speech mixtures [Yilmaz and Rickard, 2004], it becomes increasingly unrealistic for long reverberation times due to strong reflections from preceding sound events.
Step 2 - Spatial feature extraction  
In the second step, instantaneous location features are extracted for each TF point. For that purpose past research has identified a number of location cues such as directions of arrival (DOA) as well as level ratios and/or phase differences (see Araki et al. [2007b] for a review).

According to Togami et al. [2007], for sparse sources in echoic environments, the longer the distance $d_j$ is between a sensor pair $(X_{j,1}, X_{j,2})$ the better the DOA localization performance will be. Hence, the instantaneous DOA at TF point $(k, l)$ is computed here as

$$\psi(k, l) = -\frac{1}{l_0 d_{j_{\text{max}}} c^{-1}} \arg \left[ \frac{X_{j_{\text{max}},1}(k, l)}{X_{j_{\text{max}},2}(k, l)} \right],$$

(6)

where $j_{\text{max}}$ denotes the index of the sensor pair with the biggest spacing $d_{j_{\text{max}}}$ and $c$ is the propagation velocity of sound [Araki et al., 2007b; Togami et al., 2007].

In order to avoid spatial aliasing when $d_{j_{\text{max}}} > c/f_s$, we employ the SPIRE algorithm [Togami et al., 2007], which utilizes the smaller non-aliased distance pairs to restore the aliased values of the longer distance pairs (Fig. 1).

Note that without the normalization term $l_0 d_{j_{\text{max}}} c^{-1}$ in (6) the features remain frequency dependent and clustering must be performed for each frequency bin separately. Bin-wise classification strategies, such as Sawada et al. [2007], usually require longer data observations in order to guarantee accurate clustering results at each frequency bin. More importantly, the order in which the clusters are determined may be different from one frequency bin to another and a reordering is generally required to ensure that the same cluster index corresponds to the same source across all frequencies. As proposed by Yilmaz and Rickard [2004] and Jourjine et al. [2000], the frequency normalization avoids this so-called permutation problem [Mitianoudis and Davies, 2003] by utilizing all frequency bins in one single clustering step and allows the algorithm to operate on observations with short data length.

Step 3  
The DOA data set $\Psi = \{\psi(k, l) | \psi(k, l) \in \mathbb{R}, (k, l) \in \Omega\}$ is then divided into $N$ clusters, where $\Omega = \{(k, l) : 0 \leq k \leq K-1, 0 \leq l \leq L-1\}$ denotes the set of TF points in the STFT plane. Each cluster is represented by a set of prototype vectors, called centroids or centers $V = [v_n]$ with $v_n \in \mathbb{R}$ and $V \in \mathbb{R}^N$, and a partition matrix $U = [u_n(k, l)] \in \mathbb{R}^{N \times K \times L}$ indicating the degree $u_n(k, l)$ to which a data point $\psi(k, l)$ belongs to the $n$-th cluster.

While in hard clustering, such as Araki et al. [2007b], each data element belongs to exactly one cluster (binary membership values) in fuzzy clustering data points can belong to more than one cluster (continuous membership values). Here, fuzzy clustering is employed in order to reflect the localization uncertainty in a reverberant data set through a soft partitioning. More formally, let the space of all possible fuzzy partitions be defined
as

\[
P = \left\{ U = [u_n(k,l)] \mid \forall n \in \{1, \ldots, N\}, \forall (k,l) \in \Omega : \right. \\
\left. u_n(k,l) \in [0, 1]; \sum_{n=1}^{N} u_n(k,l) = 1; 0 < \sum_{\forall (k,l) \in \Omega} u_n(k,l) \right\}. 
\] (7)

Given a particular data set \( \Psi \), the search for the best fuzzy partition \( U^* \) in \( P \) is a constrained non-linear optimization problem. An algorithmic solution is usually implemented as an alternating optimization scheme, which iterates between updates for \( V \) and \( U \) until a convergence criterion is met [Theodoridis and Koutroumbas, 2006]. The final cluster centroids \( \hat{\psi}_n := v_n^* \) represent estimates of the DOA source locations and the corresponding partition matrix can be interpreted as a collection of \( N \) fuzzy TF masks

\[
\hat{M}_n(k,l) := u_n^*(k,l), \quad n = 1, \ldots, N.
\] (8)

Alternatively, binary masks may be obtained through a defuzzification process that converts the fuzzy partitioning \( U^* \) into a hard or crisp segmentation. One popular defuzzification method is to simply assign the TF point to the cluster of highest membership, e.g.,

\[
\hat{M}_n(k,l) := \begin{cases} 
1, & \text{if } n = \arg\max_j \left\{ u_j^*(k,l) \right\} \\
0, & \text{otherwise.}
\end{cases}
\] (9)

**Step 4 - Time-Frequency masking** Next, we obtain the separated signals \( \hat{S}_n(k,l) \) by applying the estimated localization masks \( \hat{M}_n(k,l) \) to one of the mixture observations:

\[
\hat{S}_n(k,l) = \hat{M}_n(k,l)X_J(k,l), \quad n = 1, \ldots, N
\] (10)

where \( J \) is a selected microphone index. Note that TF masking is prone to musical noise artifacts caused by zero-padding of spectral components in \( \hat{S}_n(k,l) \). Our use of contextually constrained fuzzy masks may ameliorate this problem somewhat by providing a smoother separation result with fewer spectral discontinuities in the extracted signals.

**Step 5 - Source resynthesis** Finally, the estimated source signals are reconstructed in the time-domain by applying the overlap-and-add method [Rabiner and Schafer, 1978] onto the masked spectra. We follow Araki et al. [2007b] and denote the reconstructed source estimate as

\[
\hat{s}_n(t) = \frac{1}{C_{\text{win}}} \sum_{k'=0}^{L/\tau_0-1} \hat{s}_n^{k+k'}(t),
\] (11)
\[ C_{\text{win}} = \frac{0.5}{m} L \]

is a constant for the Hanning window function and individual segments are obtained by an inverse STFT

\[ \hat{s}_n^k(t) = \begin{cases} 
\sum_{l=0}^{L-1} \hat{S}_n(k, l)e^{j\omega_0(t-k\tau_0)} & \text{if } (k\tau_0 \leq t \leq k\tau_0 + L - 1), \\
0 & \text{otherwise.} 
\end{cases} \]

\[ (12) \]

\section{Fuzzy clustering with observation weighting and context information}

In this section, we describe three fuzzy cluster algorithms which can be used to estimate the TF separation masks as defined in Section 3.

We start by giving a brief review of the FCM algorithm [Bezdek, 1981] and its implementation as an alternating optimization scheme. We continue with the wFCM algorithm [Kühne et al., 2009], which is able to cope with unreliable data points through the use of observation weights. The weighting scheme allows for accurate centroid determination even if the data set is contaminated by noise. Next, we adopt a recently proposed clustering technique [Pham, 2001] from the field of medical image segmentation for the problem of acoustic source separation. The new wCFCM method can produce more accurate separation masks under reverberant conditions through the use of context information during the membership updates.

The section closes with an example illustrating the estimated TF masks by each cluster algorithm under anechoic and reverberant conditions.

\subsection{Fuzzy c-means clustering}

Optimization problem and cost function

Generally, the problem of finding the best fuzzy partition \( U^* \) given the data set \( \Psi \), can then be written as a constrained non-linear optimization problem

\[ (U_{\text{FCM}}^*, V_{\text{FCM}}^*) = \arg \min_{(U, V) \in \mathcal{P}} \left\{ J_{\text{FCM}} \right\} \text{ subject to } (7). \]

\[ (13) \]

For the FCM algorithm [Bezdek, 1981], the cost function \( J_{\text{FCM}} \) takes on the form

\[ J_{\text{FCM}} = \sum_{n=1}^{N} \sum_{\forall (k, l) \in \Omega} u_n(k, l)^q D_n(k, l), \]

\[ (14) \]
where \( q \in (1, \infty) \) is a fuzzification parameter controlling the softness of the memberships and

\[
D_n(k, l) := \| \psi(k, l) - v_n \|^2
\]

is the squared Euclidean distance between the observation \( \psi(k, l) \) and the centroid \( v_n \) of the \( n \)-th cluster.

### Cluster prototype and membership updating

The minimization problem in (13) can be solved by means of Lagrange multipliers and is usually implemented as an alternating optimization scheme due to the absence of a closed form solution [Bezdek, 1981; Theodoridis and Koutroumbas, 2006].

Starting from a random partitioning, the cost function (14) is iteratively minimized by alternating the updates for the centroids and memberships

\[
v^*_n = \frac{\sum_{(k, l) \in \Omega} u_n(k, l)^q \psi(k, l)}{\sum_{(k, l) \in \Omega} u_n(k, l)^q} , \quad \forall n
\]

\[
u^*_n(k, l) = \left[ \sum_{j=1}^{N} \left( \frac{D_n(k, l)}{D_j(k, l)} \right)^{\frac{1}{q-1}} \right]^{-1} , \quad \forall n, k, l
\]

until an appropriate convergence criterion is met. Convergence is considered to be obtained when the difference between successive partition or prototype matrices is less than some predefined threshold \( \epsilon \) [Bezdek, 1981]. Although convergence is guaranteed, the alternating optimization scheme may only converge to a local rather than global optimum. It is therefore recommended to execute several runs of the algorithm and pick the best result.

The clustering procedure for the standard FCM algorithm is summarized in Alg. 1. While this scheme is computational efficient it lacks robustness against noise and outliers. In the context of clustering, outliers are usually defined as observations that are far away from all cluster centers [Rousseeuw and Leroy, 1987]. Consequently, these points should be represented by low membership weights during the centroid computation (16). However, because of the constraints in (7), FCM assigns outliers rather high membership values close to \( 1/N \). It is this inability to deal with corrupted observations in (16) that often leads the FCM algorithm to produce incorrect localization results under reverberant conditions [Kühne et al., 2009]. We, therefore, conclude that FCM is suitable only for anechoic data sets that contain few outliers or noisy observations.
Algorithm 1: FCM - The fuzzy c-means clustering algorithm.

**input**: \( \Psi, N, q, \epsilon \)

**output**: \( U^*_{\text{FCM}}, V^*_{\text{FCM}} \)

1. initialize partition matrix \( U^{(0)} \in P \) randomly
2. repeat for \( j = 1, 2, \ldots \)
3. update centroids \( V^{(j)} \) with \( U^{(j-1)} \) using (16)
4. compute distances \( D^{(j)} \) with \( V^{(j)} \) via (15)
5. update partition matrix \( U^{(j)} \) with \( D^{(j)} \) using (17)
6. until \( \| U^{(j)} - U^{(j-1)} \| < \epsilon \)
7. return \( U^*_{\text{FCM}} \leftarrow U^{(j)} \) and \( V^*_{\text{FCM}} \leftarrow V^{(j)} \)

4.2 Weighted fuzzy c-means clustering

Optimization problem and cost function

In weighted fuzzy c-means (wFCM) clustering [Kühne et al., 2009; Miyamoto et al., 2006] the reliability of each datum \( \psi(k, l) \) is indicated by an observation weight \( w(k, l) \). Let \( W = \{ w(k, l) | w(k, l) \in \mathbb{R}^+, (k, l) \in \Omega \} \) be the corresponding set of observation weights for \( \Psi \). The constrained optimization problem with observation weighting then becomes

\[
(U^*_{\text{wFCM}}, V^*_{\text{wFCM}}) = \text{argmin}_{(U, V) \in \mathcal{P}} \left\{ J_{\text{wFCM}} \right\} \text{ subject to (7),}
\]

with the cost function \( J_{\text{wFCM}} \) defined as

\[
J_{\text{wFCM}} = \sum_{n=1}^{N} \sum_{(k,l) \in \Omega} u_n(k, l)^q w(k, l) D_n(k, l).
\]

Cluster prototype and membership updating

This minimization problem can again be solved by Lagrange multipliers leading to the wFCM update equations:

\[
u^*_n = \frac{\sum_{(k,l) \in \Omega} u_n(k, l)^q w(k, l) \psi(k, l)}{\sum_{(k,l) \in \Omega} u_n(k, l)^q w(k, l)}, \quad \forall n
\]

\[
u^*_n(k, l) = \left[ \sum_{j=1}^{N} \left( \frac{D_n(k, l)}{D_j(k, l)} \right) \right]^{-1}, \quad \forall n, k, l
\]

The additional weighting factor \( w(k, l) \) in (20) allows the wFCM algorithm to incorporate prior knowledge about the reliability of each observation during the centroid updates. Its main purpose is to reduce the influence of unreliable data points while increasing
the weight of reliable observations. If the weights are chosen appropriately, the centroid estimation in wFCM becomes much less susceptible to outliers and noisy data points. Note, however, that wFCM uses the same update equation for the membership values as FCM in (17).

Selection of observation weights

For the purpose of robust source localization, a good choice of the observation weights is crucial. A number of previous studies [Faller and Merimaa, 2004; Huang et al., 1997; Litovsky et al., 1999] have shown that in echoic enclosures, only a small fraction of the location cues correspond to the correct source locations.

Based on our previous work [Kühne et al., 2009], the observation weights are estimated here prior to the clustering by scanning the TF plane for regions with low DOA fluctuations. It is thereby assumed that TF regions with low DOA fluctuations are not affected by sound reflections [Faller and Merimaa, 2004], possess a high SNR [Kim et al., 2006] and are indicative of single source zones [Abrard and Deville, 2005]. High variances, on the other hand, indicate regions where sources overlap or reflections contaminate the DOA measurements.

The local DOA variance \( \sigma^2_{\psi}(k, l) \) is computed over a small neighborhood \( N_{(k,l)} \) as

\[
\sigma^2_{\psi}(k, l) = \frac{1}{|N_{(k,l)}| - 1} \sum_{\forall (k', l') \in N_{(k,l)}} [\psi(k', l') - \mu_{\psi}(k, l)]^2,
\]

where \( \mu_{\psi}(k, l) \) is the local DOA mean

\[
\mu_{\psi}(k, l) = \frac{1}{|N_{(k,l)}|} \sum_{\forall (k', l') \in N_{(k,l)}} \psi(k', l').
\]

The neighborhood \( N_{(k,l)} \) was chosen as a 11-point window of adjacent frequency bins. The lower the local variance for a DOA measurement, the more weight should be given to this observation during clustering. We found, that a good choice for \( w(k, l) \) is the following empirically determined function

\[
w(k, l) = 1 + \frac{1}{\max\{\sigma^2_{\psi}(k, l), \kappa\}},
\]

which assigns large weights \( w(k, l) \gg 1 \) to regions with low DOA fluctuations while penalizing areas with high variances through unity weights, \( w(k, l) \approx 1 \). The constant \( \kappa \) prevents a division by zero and controls the upper limit of the weights. In our implementation, \( \kappa \) was set to \( 10^{-3} \). Note that wFCM defaults to the standard FCM if the weights are chosen to be unity for all TF points.
Figure 3: Example of observation weighting in noisy DOA feature sets. (a) Observation weights $w$ in time-frequency plane; lighter areas have lower weights; darker areas higher weights. (b) DOA histogram with unity weights (light gray bars) and with weights from (a) (dark gray bars). The true DOA angles of the two speech sources are $\pm 20^\circ$.

Fig. 3 shows an example of the weights $w(k, l)$ and illustrates the impact of the weighting scheme on the clustering structure. The iterative clustering procedure for the weighted FCM algorithm is summarized in Alg. 2.

**Algorithm 2: wFCM** - The weighted fuzzy c-means clustering algorithm.

**input**: $\Psi, W, N, q, \epsilon$

**output**: $U^{*}_{wFCM}, V^{*}_{wFCM}$

1. initialize partition matrix $U^{(0)} \in \mathcal{P}$ randomly
2. repeat for $j = 1, 2, \ldots$
3. update centroids $V^{(j)}$ using $W & U^{(j-1)}$ via (20)
4. compute distances $D^{(j)}$ with $V^{(j)}$ via (15)
5. update partition matrix $U^{(j)}$ using (21)
6. until $\| U^{(j)} - U^{(j-1)} \| < \epsilon$
7. return $U^{*}_{wFCM} \leftarrow U^{(j)}$ and $V^{*}_{wFCM} \leftarrow V^{(j)}$

### 4.3 Weighted contextual fuzzy c-means clustering

A common drawback of FCM and wFCM is their lack of robustness when confronted with reverberant speech mixtures. Typically, the estimated membership functions contain many misclassified points which often appear as speckled patterns in the TF masks. This is not surprising, given that no particular structure is imposed on the speech spectrum and membership classification of a datum depends solely on the Euclidean distances of a single TF point in the DOA feature space. Assigning a TF point independently from its context or surroundings ignores the highly structured nature of speech in the TF domain. Human speech sounds are formed through continuous movements of
articulatory organs inside the vocal tract and therefore display a smooth and continuous appearance when contemplated in the TF domain. The new wCFCM cluster algorithm incorporates such a homogeneity assumption on the speech spectra in form of contextually constrained membership functions.

**Optimization problem and cost function**

We consider the following constrained optimization problem

\[
(U_w^{*\text{wCFCM}}, V_w^{*\text{wCFCM}}) = \underset{(U,V) \in \mathcal{P}}{\text{argmin}} \left\{ J_{\text{wCFCM}} \right\} \text{ subject to } (7) (25)
\]

and follow Pham [2001] in defining the wCFCM cost function as

\[
J_{\text{wCFCM}} = \sum_{n=1}^{N} \sum_{\forall (k,l) \in \Omega} u_n(k,l)^q w(k,l) D_n(k,l) + \frac{\beta}{2} \sum_{n=1}^{N} \sum_{\forall (k,l) \in \Omega} u_n(k,l)^q \sum_{\forall (k',l') \in \mathcal{N}(k,l)} \sum_{n' \neq n} u_{n'}(k',l')^q. 
\]

Note that the first term in (26) is identical to the wFCM objective function while the second term acts as a regularization term forcing TF points from a neighborhood \( \mathcal{N}(k,l) \) to have similar membership values in the same cluster. This penalty is minimized when the membership value for a particular cluster is large and the membership values for the other clusters in a local TF neighborhood are small [Pham, 2001]. The parameter \( \beta \) controls the trade-off between minimizing the wFCM objective function and biasing the solution towards homogenous membership masks. For ease of notation, we define

\[
C_n(k,l) := \sum_{\forall (k',l') \in \mathcal{N}(k,l)} \sum_{n' \neq n} u_{n'}(k',l')^q
\]

and write (26) as

\[
J_{\text{wCFCM}} = \sum_{n=1}^{N} \sum_{\forall (k,l) \in \Omega} u_n(k,l)^q \left[ w(k,l) D_n(k,l) + \frac{\beta}{2} C_n(k,l) \right]. 
\]

**Cluster prototype and membership updating**

The constrained minimization problem (25) is again solved by Lagrange multipliers and implemented as an alternating optimization scheme. It can be shown (see Appendix) that
the wCFCM update equations are given by

\[
    v_n^* = \frac{\sum_{\forall (k,l) \in \Omega} u_n(k,l) q w(k,l) \psi(k,l)}{\sum_{\forall (k,l) \in \Omega} u_n(k,l) q w(k,l)}, \quad \forall n \tag{29}
\]

\[
    u_n^*(k,l) = \left[ \frac{\sum_{j=1}^{N} \left( \frac{w(k,l) D_n(k,l) + \beta C_n(k,l)}{w(k,l) D_j(k,l) + \beta C_j(k,l)} \right)^{\frac{q-1}{q}}} \right]^{-1}, \quad \forall n, k, l. \tag{30}
\]

As is evident from (29) the wCFCM algorithm inherits the robust centroid estimation of the wFCM algorithm through the use of the observation weights \( w(k,l) \). However, the memberships \( U^* \) are computed differently from wFCM depending on the value of the contextual weighting parameter \( \beta \). For \( \beta = 0 \), no context information is utilized and (30) becomes identical to the wFCM update equation. When \( \beta > 0 \), the value at \( u_n(k,l) \) is influenced by the membership values \( u_{n',k',l'} \) at neighboring TF points \( (k',l') \in N(k,l) \) in other clusters \( n' \neq n \). The result is a smoothing effect that causes neighboring TF points to have similar memberships in the same cluster.

The main steps of the wCFCM algorithm are summarized in Alg. 3.

**Algorithm 3**: wCFCM - The weighted contextual fuzzy c-means clustering algorithm.

**input**: \( \Psi, W, N, q, \beta, \epsilon \)

**output**: \( U^*_wCFCM, V^*_wCFCM \)

1. initialize partition matrix \( U^{(0)} \in \mathcal{P} \)
2. repeat for \( j = 1, 2, \ldots \)
3. update centroids \( V^{(j)} \) with \( U^{(j-1)} \) via (29)
4. compute distances \( D^{(j)} \) with \( V^{(j)} \) via (15)
5. compute context \( C^{(j)} \) with \( U^{(j-1)} \) via (27)
6. update partition matrix \( U^{(j)} \) using (30)
7. until \( \| U^{(j)} - U^{(j-1)} \| < \epsilon \)
8. return \( U^*_wCFCM \leftarrow U^{(j)} \) and \( V^*_wCFCM \leftarrow V^{(j)} \)

Selection of regularization parameter \( \beta \)

Proper selection of \( \beta \) is crucial to obtain near-optimal performance under varying environmental conditions. In general, the stronger the room reverberation the higher the degree of smoothing required to obtain a satisfactory clustering result. On the other hand, if reverberation is mild and there is very little noise in the feature set, then too much regularization will result in degraded performance due to over-smoothing.
In practice, generally only limited information about the mixing process or the room environment is available preventing us from choosing an optimal $\beta$ a priori. It is therefore highly desirable to obtain appropriate estimates for $\beta$ directly from the data without having to rely on trial-and-error methods or unrealistic assumptions about the noise characteristics of the input data.

Cross-validation is a well-established technique for determining a near-optimal regularization parameter without any a priori knowledge of either the amount of noise or its distribution [Reeves, 1992; Reeves and Mersereau, 1990]. One iteration of cross-validation involves partitioning a data set into complementary subsets, executing the algorithm under study with a fixed regularization parameter on one subset, and validating the outcome on the other subset. To reduce variability, the validation results are normally averaged over multiple iterations of cross-validation using different choices for the subsets.

In our application, true cross-validation with multiple data partitions is computationally prohibitive because of the large number of data points. Instead, we have resorted to a suboptimal procedure of the true cross-validation scheme, called the holdout method [Reeves, 1992]. In holdout, the data set is divided into one estimation set and one validation set. The algorithm of interest is first applied to the estimation set using a fixed value for $\beta$. The points of the validation set are assumed to be missing in this step. The outcome of the estimation step is then used to test the appropriateness of $\beta$ by computing a cross-validation error on the validation set. This process is repeated for different values of $\beta$, and the value $\beta^*$ that results in the lowest cross-validation error is considered to be optimal. Contrary to true cross-validation, this offers the advantage that the cross-validation error is only computed once using the left-out data from the validation set. Although the holdout method is not as reliable as true cross-validation, it usually yields reasonable estimates for $\beta$ with substantial computational savings [Reeves, 1992].

For our wCFCM algorithm, we mainly adopt the holdout scheme presented in Pham [2001], which has been shown to perform well in clustering problems related to medical image segmentation. For a particular choice of $\beta$, the centroids $v_n^{(\beta)}$ and cluster membership values $u_n^{(\beta)}(k,l)$ obtained from the estimation set are validated for their clustering performance using the following wFCM criterion-based cross-validation error

$$E_{cv}(\beta) = \sum_{n=1}^{N} \sum_{(k,l) \in \Omega_v} u_n^{(\beta)}(k,l)^q w(k,l) \| \psi(k,l) - v_n^{(\beta)} \|^2,$$

(31)

where $\Omega_v$ denotes the indices of all TF points in the validation set. The choice of $\Omega_v$ is not very critical as long as each source is sufficiently represented in the set. As recommended in Reeves and Mersereau [1990], we chose the validation points in $\Omega_v$ randomly with $|\Omega_v| = \frac{1}{10} |\Omega|$.

Fig. 4 shows a typical plot of the cross-validation error $E_{cv}(\beta)$ computed for various
values of \( \beta \) using the described holdout strategy. For most cases, the cross-validation error function is of convex shape and shows a clear global minimum. We follow Pham [2001] and stop the search once the first local minimum of \( E_{cv}(\beta) \) has been found. The complete description of the steps involved for the wCFCM algorithm using pseudo-cross-validation is given in Alg. 4.

**Algorithm 4**: wCFCM using the holdout method for parameter selection (adopted from Pham [2001]).

\[
\begin{align*}
\textbf{input} & : \Psi, \Omega_v, W, N, q, \epsilon \\
\textbf{output} & : U_{wCFCM}^*, V_{wCFCM}^*
\end{align*}
\]

1. run wFCM to determine \( J_{wFCM} \)
2. compute \( \beta_{inc} = 0.1 \frac{J_{wFCM}}{(J_{wCFCM} - J_{wFCM})/\beta} \)
3. set \( \beta = \beta_{inc} \)
4. run wCFCM on estimation set \( \Omega \setminus \Omega_v \)
5. compute \( E_{cv}(\beta) \) on validation set \( \Omega_v \)
6. if \( E_{cv}(\beta) \) is not at a local minimum set \( \beta = \beta + \beta_{inc} \) and go to Step 4, using the current clustering result as an initialization of the next application of wCFCM; otherwise set \( \beta^* = \arg\min_{\beta} \{ E_{cv}(\beta) \} \) and go to Step 7.
7. apply wCFCM to entire TF plane \( \Omega \) using \( \beta^* \) as regularization parameter

4.4 Example

The following example provides an illustration of the TF masks produced by the three cluster algorithms FCM, wFCM and wCFCM under anechoic and reverberant conditions. For comparison purposes, the estimated fuzzy membership masks are presented alongside binary a priori masks [Wang, 2005; Yilmaz and Rickard, 2004]. These a priori masks are obtained using the premixed source signals and serve here as "ground truth" reference for judging the quality of the partitioning result of each cluster algorithm.
Our example consists of a speech mixture with two sources of equal gain. Clustering is performed in anechoic ($RT_{60}^1 = 0\, \text{ms}$) and reverberant ($RT_{60} = 300\, \text{ms}$) conditions. A fuzzy exponent $q = 2$ was used for all cluster algorithms. The holdout method was employed to automatically select the amount of smoothing performed in wCFCM. Note that the data length was relatively short with 2.8 s.

Fig. 5 shows the a priori mask as well as the fuzzy membership masks generated by FCM, wFCM and wCFCM with $q = 2$ for the anechoic speech mixture. All three cluster algorithms produced very accurate results when compared to the a priori mask. In the absence of any sound reflections almost all extracted features are reliable indicators of the source locations making it possible for FCM and wFCM to perform well in this situation. The wCFCM result was similarly successful although it exhibited some minor loss in detail in the low frequency regions due to its neighborhood smoothing.

Fig. 6 shows the a priori mask as well as the fuzzy membership masks generated by FCM, wFCM and wCFCM for the reverberant speech mixture. Due to the adverse impact of sound reflections on the localization measurements many of the extracted location features deviated significantly from their true value. As is evident from Fig. 6(b) and 6(c), the isolated membership assignment of each TF point in FCM and wFCM is highly vulnerable to noise in the feature set. When compared to the “ground truth” in Fig. 6(a) the FCM and wFCM masks are more speckled and contain many misclassifications. In contrast, the wCFCM result in Fig. 6(d) is much smoother and less speckled due to the inclusion of context information. This suggests that wCFCM is more robust than conventional clustering and may have the potential to improve the separation performance of current BSS systems for reverberant speech mixtures.

5 Experimental evaluation

In this section, we present some results for the application of the three fuzzy cluster algorithms in a BSS framework with synthetic speech mixtures. The first experiment examined the clustering performance on an over-determined BSS task using conventional TF masking for demixing the sources from the mixture. In the second experiment, we studied the impact of using a fuzzy mask as opposed to a binary mask when performing the demixing in reverberant conditions. The third experiment reports on the source localization accuracy of the three cluster algorithms when deployed in anechoic and reverberant environments. The last experiment investigated the performance of the cluster algorithms in a BSS application that combines TF masking with beamforming.

$RT_{60}^1$ is defined as the time required for reflections of a direct sound to decay by 60 dB following sound offset [Lehmann and Johansson, 2008].
Figure 5: Comparison of a priori masks with fuzzy membership masks generated by FCM, wFCM and wCFCM in anechoic conditions. Lighter areas indicate lower membership values; darker areas represent higher membership levels. RT<sub>60</sub> = 0 ms, SIR = 0 dB, M = 6, d = 4.28 cm, f<sub>s</sub> = 8 kHz, q = 2.

5.1 Experimental setup

Multipath sound propagation was simulated for a small rectangular room with dimensions 6 m x 4 m x 3 m (length x width x height). Wall reflections were estimated using the image model method for simulating small-room acoustics [Lehmann and Johansson, 2008]. Room impulse responses for different reverberation times were generated for each sensor of a six-channel ULA with inter-element spacing of d = 4.28 cm and a sampling frequency of 8 kHz. The array was positioned in the middle of the room at a height of 2 m. Facing array broadside, two sources with equal gain were placed in the horizontal plane at azimuth angles of θ<sub>1</sub> = −20° and θ<sub>2</sub> = 20° and a distance of 1.5 m from the array center.

The sound mixtures consisted of two speech sources, one from the TIDIGIT [Leonard, 1984] and the other from the TIMIT [Garofolo et al., 1993] database. For evaluation
5. Experimental evaluation

Figure 6: Comparison of *a priori* TF masks with fuzzy membership masks generated by FCM, wFCM and wCFCM in reverberant conditions. Lighter areas indicate lower membership values; darker areas represent higher membership levels. $\text{RT}_{60} = 300 \text{ ms}$, SIR = 0 dB, $M = 6$, $d = 4.28 \text{ cm}$, $f_s = 8 \text{ kHz}$, $q = 2$.

purposes, a total of 240 different mixtures were constructed. The average utterance length was around 2.5 s. Simulations were run for three room reverberation times $\text{RT}_{60} \in \{0 \text{ ms}, 300 \text{ ms}, 600 \text{ ms}\}$. The STFT frame size was 64 ms with a shift of 10 ms.

It is widely known that the performance of fuzzy clustering strongly depends on the initialization of the algorithm. For FCM and wFCM, the best solution among 50 runs was selected as final result in order to minimize the risk of finding a local rather than global optimum. The wCFCM algorithm was initialized with the best wFCM result and the regularization parameter was determined using the cross-validation method. A rectangular neighborhood of size $15 \times 9$ (frequency $\times$ time) TF points was used for the contextual regularization term in wCFCM.

For the purpose of quantifying the separation performance, we resorted to the measures provided by the freely available BSS_EVAL toolbox [Févotte et al., 2005]. The toolbox operates on the assumption that a given source estimate $\hat{s}(t)$ can be modeled as the
following sum
\[ \hat{s}(t) = s(t) + e_i(t) + e_n(t) + e_a(t), \] 
(32)

where \( s(t) \) is an allowed deformation of the target source, \( e_i(t) \) accounts for distortions due to unwanted interfering sources, \( e_n(t) \) is perturbing noise and \( e_a(t) \) characterizes all other artifacts introduced by the separation algorithm, e.g., musical noise. The decomposition of the estimated sources was performed using the toolbox function `bss_decomp_filt`, which allows for time-invariant filter distortions of the target source. The filter length was set to 256 taps as recommended in Vincent et al. [2006]. The following three global performance measures were computed. Firstly, the source-to-distortion ratio (SDR)

\[ SDR := 10 \log_{10} \left[ \frac{\sum_t |s(t)|^2}{\sum_t |e_i(t) + e_n(t) + e_a(t)|^2} \right] \text{dB} \] 
(33)

is an overall quality measure for the separation results. Secondly, the sources-to-interferences ratio (SIR)

\[ SIR := 10 \log_{10} \left[ \frac{\sum_t |s(t)|^2}{\sum_t |e_i(t)|^2} \right] \text{dB} \] 
(34)

quantifies the strength of interfering sources in the target source estimate. Lastly, the sources-to-artifacts ratio (SAR)

\[ SAR := 10 \log_{10} \left[ \frac{\sum_t |s(t) + e_i(t) + e_n(t)|^2}{\sum_t |e_a(t)|^2} \right] \text{dB} \] 
(35)

measures the amount of artifacts in the source estimates. For the experiments considered here, we assumed ideal omni-directional microphones so that \( e_n(t) \) can be omitted in the above definitions. In order to express the SIR and SDR improvements between the speech mixture input and the processed BSS output, we also computed the corresponding gains, e.g., \( SIR_{\text{gain}} = SIR_{\text{output}} - SIR_{\text{input}} \). All performance criteria are expressed in dB and the higher the ratios are the better the quality of the separation result is.

5.2 Results

Separation performance for conventional TF masking

First, we tested the clustering performance using conventional TF masking for demixing the two speakers from the mixtures. In all cases, clustering was performed with \( q = 2 \) and binary TF masks were estimated using the maximum membership assignment in (9).

Fig. 7(a)-(c) show the separation results for the three cluster algorithms in anechoic and reverberant test scenarios. These figures demonstrate the superiority of wCFCM over FCM for the reverberant test cases. For example, wCFCM achieved substantial SIR gains of up to 5 dB over conventional fuzzy clustering (Fig. 7(b)) while at the same time
producing similar artifacts in the output signals (Fig. 7(c)). We also note that wCFCM performed slightly worse than FCM and wFCM in anechoic conditions. This was caused by the cross-validation method used to determine the optimal smoothing parameter. It was found that the method overestimated the strength of $\beta$ in some cases, which led to performance degradations due to over-smoothing.

**Soft vs. hard masking**

Next, we studied the impact of using a fuzzy mask as opposed to a binary mask when performing the demixing. The wCFCM cluster algorithm was run several times with a different fuzzy exponent $q \in \{1.1, 1.3, 1.5, 1.8, 2.0, 2.3, 2.6, 3.0\}$. The choice of $q$ controls the softness of the generated TF masks and the closer this parameter is to unity the more binary the membership levels become. The separation performance was recorded for both fuzzy and binary masks, as defined in Eq. (8) and (9). The reverberation time $\text{RT}_{60}$ was 300 ms.

From Fig. 8, we observe that the binary masks outperformed the fuzzy masks significantly in terms of interference suppression for values of $q > 2$. For smaller values of $q$ the performance of the fuzzy masks approached those of the binary masks. This is expected, because for $q \in (1, 1.5]$ the fuzzy clustering effectively turns into a hard clustering with almost binary membership values. Note also that the highest SIR gains were achieved for $q = 2.0$ and $q = 2.3$, which suggests that for reverberant mixtures fuzzy clustering techniques may perform better than hard clustering approaches, such as k-means.

From Fig. 9, we see that according to the SAR measure the fuzzy masks caused fewer artifacts in the output spectra than their binary counterparts. This agrees with a previous study [Araki et al., 2006b], which found that soft TF masks can significantly reduce musical noise by preventing excessive zero-padding in the BSS outputs. The question whether musical noise distortions are acceptable often depends on the target application.
Figure 8: Separation performance of fuzzy and binary TF masking in terms of SIR improvements when \( w\text{CFCM} \) clustering was performed with different values for the fuzzy exponent \( q \).

Figure 9: Separation performance of fuzzy and binary TF masking in terms of SAR when \( w\text{FCM} \) clustering was performed with different values for the fuzzy exponent \( q \).

For example, speech recognition systems are usually more interested in suppressing energy from interfering sources than reducing musical noise. On the other hand, this may be different for audio applications intended for human listeners.

In conclusion, our results suggest that for striking a balance between SIR and SAR, a good choice for the fuzzy exponent is \( q \approx 2 \). The proposed fuzzy clustering also provides the user with the option to apply hard or soft TF masking, depending on the application at hand.

Source DOA localization accuracy

In this experiment, the localization accuracy of the three cluster algorithms \( \text{FCM} \), \( \text{wFCM} \) and \( w\text{CFCM} \) was determined under different room reverberation times. Performance
was quantified in terms of the root-mean-square error (RMSE)

\[
\text{RMSE} := \sqrt{\frac{1}{N} \sum_{n=1}^{N} (\hat{\theta}_n - \theta_n)^2},
\]

which measures the difference between the estimated \(\hat{\theta}_n\) and the true DOA angle \(\theta_n\) averaged over all sources. The lower the RMSE value the better the localization accuracy of the cluster algorithm. The fuzzifier parameter was set to \(q = 2\).

Fig. 10 shows the obtained results for each cluster algorithm in terms of the RMSE localization error. Not surprisingly, for anechoic conditions the localization performance of all three cluster algorithms was very accurate. With only two sources and no reverberation most DOA observations contributed reliable measurements for the clustering process. For reverberant data, strategies with observation weighting (wFCM, wCFCM) clearly outperformed conventional FCM. This is consistent with previous studies [Aarabi and Mavandadi, 2003; Faller and Merimaa, 2004; Kim et al., 2006; Yilmaz and Rickard, 2004], which found that the accuracy of histogram based source localization strategies greatly benefits from assigning reliability weights to data points.

![Figure 10: Localization accuracy in terms of RMSE for the three different cluster algorithms FCM, wFCM and wCFCM in different reverberant environments. The error bars show the standard deviation computed over all outputs.](image)

**Separation performance for combined TF masking and beamforming**

In our last experiment, we investigated the performance of each fuzzy cluster algorithm when deployed in a BSS application that combines TF masking with spatial filtering. In Cermak et al. [2006], it was shown that the outcome of the clustering stage can be used to blindly design the spatial filter weights of \(N\) adaptive beamformers in the frequency
domain. The beamformed sources are then estimated as

$$\hat{S}_n(k,l) = b_n^*(l)^H X(k,l), \quad n = 1, \ldots, N$$  \hfill (37)

where $X(k,l) = [X_1(k,l), \ldots, X_M(k,l)]^T$ is an observation vector and $b_n^*(l) = [b_{1n}^*(l), \ldots, b_{Mn}^*(l)]^T$ denotes the optimal beamformer weight at frequency bin $l$ according to some design criterion. In our implementation, we employed linear constrained minimum variance (LCMV) beamforming, where the filter weights are given by [Malonakis et al., 2000]

$$b_n^*(l) = R_n^{-1}(l)A(l)\left(A(l)^H R_n^{-1}(l)A(l)\right)^{-1}\delta_n.$$  \hfill (38)

$R_n(l)$ is the noise-plus-interference correlation matrix, $A(l) = [a_1(l), \ldots, a_N(l)]$ is the constraint matrix containing the steering vectors

$$a_n(l) = \left[e^{-j\omega_0 d_{m1} c^{-1}\psi_n}, \ldots, e^{-j\omega_0 d_{M1} c^{-1}\psi_n}\right]^T$$  \hfill (39)

and $\delta_n = (\delta_{n1}, \ldots, \delta_{nN})^T$ is the constraint response vector with

$$\delta_{ni} = \begin{cases} 1, & \text{if } i = n \\ 0, & \text{otherwise.} \end{cases}$$  \hfill (40)

Essentially, the spatial filter weights $b_n^*(l)$ are designed to let pass all signals from DOA $\psi_n$ while rejecting all energy received from interfering DOAs $\psi_{i \neq n}$. However, in practice the true $R_n(l)$ and $A(l)$ are unknown and need to be derived from the available data. As proposed in Cermak et al. [2006], we determined suitable estimates for both quantities by utilizing the outcome of the clustering. The constraint matrix $\hat{A}(l)$ was obtained by replacing $\psi_n$ in (39) with the estimated cluster centroid $\hat{\psi}_n$. The jammer correlation matrix $\hat{R}_n(l)$ was estimated through a weighted mean

$$\hat{R}_n(l) = \frac{\sum_{k=0}^{K-1} \rho_n(k,l)X(k,l)X(k,l)^H}{\sum_{k=0}^{K-1} \rho_n(k,l)},$$  \hfill (41)

where the weights $\rho_n(k,l) = 1 - \hat{M}_n(k,l)$ specify the jammer dominant TF slots for source $S_n$ as indicated by the corresponding TF mask. For a more in-depth discussion on issues related to beamforming and TF masking, we refer the reader to the relevant references [Araki et al., 2007a; Cermak et al., 2006, 2007]. For this experiment, we used the binary maximum memberships masks (9) and performed clustering with $q = 2$ in all cases.

Fig. 11 shows the separation performance of the LCMV beamformer when the filter weights were estimated blindly using either the FCM, wFCM or wCFCM cluster algorithm. Among these three methods, wCFCM achieved the best outcome with FCM.
5. Experimental evaluation

and wFCM producing nearly identical separation results. In general, the LCMV beamformer performed very well on anechoic mixtures outperforming the separation results achieved with conventional TF masking (see Fig. 7). On the other hand, the separation capabilities of a small microphone array are limited in an echoic environment. As evident from Fig. 11(b), the SIR measure dropped from 27 dB in anechoic settings to around 3 dB for the most reverberant test scenario. Similar performance deteriorations for adaptive beamforming have also been observed in previous studies [van Hoesel and Clark, 1995; Weiss, 1987]. We also note that the poor localization accuracy of FCM had very little impact on the separation performance. This can be explained by the small array aperture of $\approx 21 \text{ cm}$, which resulted in broad beams with high side-lobes at low frequencies ($< 1 \text{ kHz}$) making the LCMV beamformer particularly vulnerable to sound reflections in a multi-path environment.

Because of these deficiencies it is common to post-process the beamformer outputs further using some sort of post-filtering. Fig. 12 shows the separation performance when TF masking was additionally applied to the LCMV beamformer outputs. We observed that the non-linear masking operation improved the SIR measure (Fig. 12(b)) considerably by further suppressing the signal energies in jammer dominated TF cells. In particular, the TF masks produced by wCFCM resulted in substantial SIR gains of up to 5 dB compared to LCMV beamforming alone. These findings are in line with Kolossa and Orglmeister [2004], who reach similar conclusions regarding the use of TF masking as a postprocessing step in frequency domain BSS.

However, as noted previously, the downside of such an operation is the introduction of non-linear distortions (musical noise) in the output signals. The SAR criterion, which measures this type of distortion, indicated an increase in artifacts only for the anechoic but not for the reverberant cases (compare Fig. 11(c) and Fig. 12(c)). We conducted some informal listening tests to assess the audio quality of the separated signals, because it has been reported [Dmour and Davies, 2008; Fèvotte and Godsill, 2006] that the SAR measure may not always accurately represent the amount of musical noise. These tests confirmed that the post-processed source estimates suffered from stronger musical noise than the LCMV outputs without additional TF masking.

5.3 General discussion

Overall, our study has demonstrated that observation weighting and context information can improve the source separation performance of a conventional fuzzy cluster algorithm under reverberant conditions. The proposed wCFCM algorithm achieved substantial gains in terms of SIR and SDR improvements over conventional FCM in a number of test scenarios.
Fuzzy Clustering using Observation Weighting and Context Information

Figure 11: Source separation performance of LCMV beamforming in terms of (a) SDR gain, (b) SIR gain and (c) SAR when beamformer weights were estimated with FCM, wFCM or wCFCM. The error bars show the standard deviation computed over all outputs.

Figure 12: Source separation performance of combined TF masking and LCMV beamforming in terms of (a) SDR gain, (b) SIR gain and (c) SAR when masks and beamformer weights were estimated with FCM, wFCM or wCFCM. The error bars show the standard deviation computed over all outputs.

However, no comparisons between wCFCM and other BSS algorithms were presented in this paper. The main objective of this study was to provide a proof of concept and establish the potential of context information for increasing the robustness of standard cluster algorithms. While in the field of image segmentation the importance of context information has been demonstrated before [Chuang et al., 2006; Li, 2001; Liew et al., 2000], we were unable to find any previous reports on this topic for the acoustic BSS problem. It is our hope that this paper will encourage other researchers in the BSS community to explore similar strategies in order to advance the paradigm of TF masking. We also expect wCFCM to prove very useful in related research areas, such as musical source separation [Li and Wang, 2008] or missing data speech recognition [Cooke et al., 2001].

Before concluding this paper, we would like to comment on several limitations in our approach and point out some extensions likely to result in further performance improvements.
5. Experimental evaluation

Firstly, it is known that the Euclidean $L_2$-norm distance is not robust to outliers or strong noise in the data set. Here, we have addressed this issue by reducing the influence of outliers and noisy DOA observations through the use of observation weights. Alternatively, the non-robust Euclidean $L_2$-norm distance could be replaced with more robust $L_p$-norm distances [Hathaway et al., 2000], such as the $L_1$-norm [Kersten, 1999] or kernel-based distance measures [Zhang and Chen, 2003].

One important point that we have not addressed so far is the question of computational complexity. For this study, all three fuzzy cluster algorithms were implemented in MATLAB™ 7.5 on a 3 GHz Intel® Core™ 2 Duo machine running Linux. Execution times varied according to the dimension of the data set and the amount of noise. While the size of the data set is directly linked to the resolution parameters of the short-time spectral analysis, the amount of noise is influenced by environmental factors, such as the room reverberation time and the number of sources.

Table 1 shows the CPU times for the three fuzzy cluster algorithms averaged over 50 runs. Clearly, wCFCM requires by far the longest CPU-time among the three cluster algorithms. This is mostly due to the additional computation of the context regularization term in each iteration and the need to select the smoothing parameter $\beta$ via cross-validation. Reducing the computational burden can therefore be considered an important topic for future research.

Another point of concern is the size and shape of the neighborhood system $N_{(k,l)}$ used to collect the context information around a TF point. Although the simple rectangular context window was successful in improving the wCFCM membership estimation it remained fixed throughout the entire TF plane and did not adapt to the structure of the speech sources. In order to preserve the local homogeneity of the underlying speech characteristics and avoid loss of detail in the TF masks the use of adaptive neighborhood models [Andreadis et al., 1996] needs to be investigated. Ideally, the size as well as the shape of the context window should be tailored to the local source characteristics in the TF plane. In this regard, it seems also worthwhile to investigate strategies in which the smoothing parameter $\beta$ is allowed to vary with the local characteristics of the speech spectra.

Extending the algorithm to multi-dimensional feature sets is another topic for further research. In our current implementation, only one-dimensional spatial cues extracted from the sensor pair with the biggest spacing are utilized during clustering. The extension of the wCFCM cluster algorithm to higher feature dimensions is straightforward. This includes the use of additional delay estimates from other sensor pairs, for example when using a non-linear array geometry as in MENUET [Araki et al., 2007b], and the use of level ratios as in DUET [Yilmaz and Rickard, 2004]. However, spatial cues become less effective the stronger the reverberation and the smaller the angular separation be-
Table 1: Average CPU time (±1 standard deviation over 50 trials, in seconds) for separating a 3 s mixture of two speech sources in different reverberant environments.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reverberation time RT$_{60}$ in ms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>FCM</td>
<td>0.4 ± 0.1</td>
</tr>
<tr>
<td>wFCM</td>
<td>0.4 ± 0.1</td>
</tr>
<tr>
<td>wCFCM</td>
<td>2.8 ± 0.6</td>
</tr>
<tr>
<td></td>
<td>(17.4 ± 8.4)$^a$</td>
</tr>
</tbody>
</table>

$^a$ Numbers in braces indicate the extra time spent for selecting the regularization parameter $\beta^*$ using cross-validation.

...tween the speakers. Augmenting the location features with pitch or harmonicity cues may provide another important source of information for the cluster algorithm in these challenging conditions.

Lastly, like in most related work [Araki et al., 2007b; Cermak et al., 2006; Sawada et al., 2007; Yilmaz and Rickard, 2004], the number of source signals $N$ needs to be supplied by the user. For our algorithm to operate in a fully unsupervised way it is necessary to automatically detect the number of (speech) sources present in the scene. Although this is in itself a challenging task the problem is well studied in the pattern classification literature and a large number of suboptimal solutions exist [Gath and Geva, 1987].

6 Conclusions

In this paper, we presented a novel fuzzy cluster algorithm to blindly separate reverberant mixtures of speech signals using the concept of TF masking. In order to better deal with noisy data sets, the proposed wCFCM technique incorporates observation weights and context information directly into the clustering procedure. The former helps to improve the source localization accuracy by ignoring noisy observations during the centroid updates. The latter smooths the cluster membership levels by exploiting the highly structured nature of speech signals in the TF domain. Moreover, wCFCM avoids the frequency permutation problem and is able to operate on observations with short data length.

In a number of experiments with anechoic and reverberant speech mixtures, wCFCM was found to be superior to conventional fuzzy clustering, both in terms of DOA localization accuracy as well as source separation performance.

Future work needs to validate the method on real data and compare the separation performance against other competing state-of-the-art BSS algorithms.
Acknowledgments

The authors would like to thank Dr. Eric Lehmann for providing the MATLAB® code for the image model used to generate the room impulse responses. We also express our gratitude to the three anonymous reviewers for their constructive suggestions and criticisms.

Appendix

Derivation of wCFCM update equations

This section presents the derivation of the necessary conditions for the centroids and membership functions to be at a local minimum of the wCFCM objective function. Due to the conditions on the fuzzy memberships in (7), this is a constrained minimization problem which can be solved by the method of Lagrange multipliers. We remark that the derivation for the membership levels is similar to the one in Pham [2001] and is given here for completeness. For the following discussion we assume that $q > 1$.

We begin by defining the Lagrangian function as

$$
L(U, V) := \sum_{n=1}^{N} \sum_{(k,l) \in \Omega} u_n(k, l)^q w(k, l) \| \psi(k, l) - v_n \|^2 \\
+ \frac{\beta}{2} \sum_{n=1}^{N} \sum_{(k,l) \in \Omega} u_n(k, l)^q \sum_{(k',l') \in N(k,l)} \sum_{n' = 1, n' \neq n}^{N} u_{n'}(k', l')^q \\
+ \sum_{(k,l) \in \Omega} \lambda(k, l) \left( 1 - \sum_{n=1}^{N} u_n(k, l) \right),
$$

where the $\lambda(k, l)$ are the Lagrange multipliers enforcing the membership constraint in (7).

Taking the partial derivative of (42) with respect to $v_n$ and setting the result to zero, we have

$$
\left[ \frac{\partial L}{\partial v_n} = -2 \sum_{(k,l) \in \Omega} u_n(k, l)^q w(k, l)(\psi(k, l) - v_n) \right]_{v_n = v_n^*} = 0.
$$

Solving for $v_n^*$ we directly obtain the update equation for the centroids

$$
v_n^* = \frac{\sum_{(k,l) \in \Omega} u_n(k, l)^q w(k, l) \psi(k, l)}{\sum_{(k,l) \in \Omega} u_n(k, l)^q w(k, l)}.
$$
Similarly, by taking the derivative of (42) with respect to $u_n(k, l)$ and setting the result to zero, we obtain

$$\frac{\partial L}{\partial u_n(k, l)} = qu_n(k, l)^{q-1} \left( w(k, l) \| \psi(k, l) - v_n \|^2 + \beta \sum_{\gamma(k', l') \in \mathcal{N}(k, l)} \sum_{n' \neq n} u_{n'}(k', l')^q \right) - \lambda(k, l) = 0. \quad (45)$$

Solving for $u_n^*(k, l)$ leads to

$$u_n^*(k, l) = \left[ q \left( w(k, l) \| \psi(k, l) - v_n \|^2 + \beta \sum_{\gamma(k', l') \in \mathcal{N}(k, l)} \sum_{n' \neq n} u_{n'}(k', l')^q \right) \right]^{-\frac{1}{q-1}} \lambda(k, l), \quad (46)$$

which still depends on $\lambda(k, l)$. Because the solution in (46) must also satisfy the membership constraint

$$\sum_{n=1}^{N} u_n^*(k, l) = 1, \quad (47)$$

we can substitute (46) in (47) and solve for $\lambda(k, l)$, which gives

$$\lambda(k, l)^{-\frac{1}{q-1}} = \sum_{n=1}^{N} \left[ q \left( w(k, l) \| \psi(k, l) - v_n \|^2 + \beta \sum_{\gamma(k', l') \in \mathcal{N}(k, l)} \sum_{n' \neq n} u_{n'}(k', l')^q \right) \right]^{-\frac{1}{q-1}}. \quad (48)$$

Finally, by combining (46) and (48), we obtain the update equation for the membership levels

$$u_n^*(k, l) = \frac{\left( w(k, l) \| \psi(k, l) - v_n \|^2 + \beta \sum_{\gamma(k', l') \in \mathcal{N}(k, l)} \sum_{n' \neq n} u_{n'}(k', l')^q \right)^{-\frac{1}{q-1}}}{\sum_{j=1}^{N} \left( w(k, l) \| \psi(k, l) - v_j \|^2 + \beta \sum_{\gamma(k', l') \in \mathcal{N}(k, l)} \sum_{n' \neq j} u_{n'}(k', l')^q \right)^{-\frac{1}{q-1}}}. \quad (49)$$
With the help of (15) and (27), this can also be written as

$$u^*_n(k,l) = \left( \frac{w(k,l)D_n(k,l) + \beta C_n(k,l)}{\sum_{j=1}^{N} (w(k,l)D_j(k,l) + \beta C_j(k,l))} \right)^{-\frac{1}{q-1}} \quad (50)$$

$$= \left[ \sum_{j=1}^{N} \left( \frac{w(k,l)D_n(k,l) + \beta C_n(k,l)}{w(k,l)D_j(k,l) + \beta C_j(k,l)} \right)^{\frac{1}{q-1}} \right]^{-1} \quad (51)$$

Note that by using the results of Pham [2001] and Bezdek [1980], it can be shown that the wCFCM update equations also guarantee to decrease the wCFCM objective function in each iteration.

**References**


Fuzzy Clustering using Observation Weighting and Context Information


REFERENCES


R. van Hoesel and G. Clark. Evaluation of a portable two-microphone adaptive beamforming speech processor with cochlear implant patients. *Journal of the Acoustical Soci-


A Novel Evidence Model for Missing Data Speech Recognition with Applications in Reverberant Multisource Environments

Abstract

Conventional hidden Markov model (HMM) decoders often experience severe performance degradations in practice due to their inability to cope with uncertain data in time-varying environments. In order to address this issue, we propose the bounded-Gauss-Uniform mixture probability density function (PDF) as a new class of evidence model for missing data speech recognition. The new mixture model offers a time-varying, frame-by-frame compensation of the HMM state variances while at the same time properly reflecting the bounded nature of spectral filterbank energy features in both mixture components. Exemplary for a reverberant multi-source scenario, we illustrate how the parameters of this new mixture PDF can be estimated in the spectral feature domain with the help of a multi-channel source separation front-end. The superiority of the proposed model over conventional evidence PDFs is demonstrated for a connected digits recognition task under varying test conditions.

1 Introduction

The capability to deal with multiple concurrent acoustic sources in reverberant enclosures is a key requirement for automatic speech recognition (ASR) technology to be deployed in real life situations. Despite several decades of intensive research ASR performance still falls short in comparison with human listeners [Lippmann, 1997]. The inability of conventional decoding strategies to cope with uncertain data in time-varying environments is often seen as the main reason for this shortcoming.

In Morris et al. [1998] and Cooke et al. [2001], missing data ASR has been proposed as a new decoding paradigm with the prospect of achieving a higher noise robustness under dynamically changing environments. At the heart of this method lies the observation that under noisy conditions only parts of a spectral feature vector become corrupted whilst the remaining components stay relatively unaffected by the noise. The classification of such a partly corrupted feature vector is then performed by ignoring its noise corrupted elements when calculating the HMM state emission likelihoods. By utilizing a priori noise knowledge, several previous studies [Cooke et al., 2001; Harding et al., 2006; Roman
et al., 2003] have demonstrated the great potential of missing data decoders for retaining high recognition rates in environments with low signal-to-noise-ratios (SNRs). However, the crucial problem with missing data techniques remains the identification of the noise corrupted feature components under practical conditions. Considering the unlimited types of noise distortions that can be encountered in practice, it becomes clear that any such decision cannot be made with absolute certainty.

In order to quantify the level of uncertainty when dealing with indeterministic data, Morris et al. [2001] developed the concept of evidence modeling for missing data decoding. Evidence models are statistical descriptors in the form of probability density functions (PDFs) and express the belief to which each possible feature datum represents the true clean data value. The great diversity of statistical PDF types makes it possible to tailor the shape of the evidence model to the information available at hand. With their probabilistic framework, Morris et al. provided the theoretical basis for the heuristically motivated marginalization schemes that had been employed thus far in missing data techniques. More importantly, they point out that some form of speech enhancement needs to be incorporated into the observation PDF estimation procedure for truly exploiting the potential of this concept.

In this spirit, this paper presents a combination of multi-channel speech enhancement and missing data speech recognition. As our main contribution, we present with the bounded-Gauss-Uniform mixture PDF a new class of evidence PDF that offers some important advantages over previously proposed models. Firstly, it realistically represents the outcome of a speech enhancement process by modeling not only the enhanced feature value but also the imperfections of the enhancement process. Secondly, it takes into consideration that for some spectro-temporal regions the distortions may be too severe to be corrected by the pre-processing. Thirdly, it properly reflects the bounded nature of spectral features in both mixture components. In comparison with previous models, the new evidence PDF retains a fuller description of the available data, thereby providing a more effective link between speech enhancement and recognition.

The remainder of this article is organized as follows: Section 2 describes the HMM decoding with certain and uncertain data and briefly reviews the concept of evidence modeling. We also present an overview about the most common types of evidence PDFs found in the missing data related literature. Section 3 proceeds by introducing the novel bounded-Gauss-Uniform mixture model and demonstrates how its parameters can be estimated with the aid of a multi-channel blind source separation (BSS) front-end. Section 4 reports on our evaluation and presents results for a number of ASR experiments conducted in reverberant multi-source environments. The paper closes in Section 5 with a general discussion and gives an outlook on future work.
2 Evidence modeling for HMM state observation likelihoods

In this section, we briefly review HMM based decoding strategies with certain and uncertain data. The concept of evidence modeling is explained and an overview over the most common PDF types is provided.

According to the maximum a posteriori (MAP) criterion, the goal of the Viterbi speech decoder is to find the word sequence \( \hat{W} = w_1, w_2, \ldots, w_N \), which maximizes

\[
\hat{W} = \arg\max_W \max_{Q \in Q_W} P(W)P(Q|W)P(O|Q),
\]

where \( Q = q_1, q_2, \ldots, q_K \) represents a particular state sequence through the HMM network and \( O = o_1, o_2, \ldots, o_K \) is a given sequence of acoustic feature vectors. In the following, we are only concerned with the acoustic score \( P(O|Q) \) because the language model \( P(W) \) as well as the transition probabilities \( P(Q|W) \) are not affected by uncertain or missing features. Using the Markovian independence assumption the acoustic score can be written as

\[
P(O|Q) \equiv \prod_{k=1}^{K} p(o_k|q(k), \Lambda_{q(k)}),
\]

where \( p(o_k|q(k), \Lambda_{q(k)}) \) is obtained through the output distribution of HMM state \( q(k) \) with its corresponding model parameter set \( \Lambda_{q(k)} \). Next, we describe how these state likelihoods are computed with certain and uncertain data.

2.1 State likelihood computation with certain data

The HMM output probability distributions are usually modeled as Gaussian mixture models (GMMs) with the common assumption that each mixture component has a diagonal covariance matrix. Let the GMM model parameters for a particular HMM state \( j \) be denoted as \( \Lambda_j = \{c_j, \mu_j, \sigma_j\} \), where the three components represent the mixture weights, mean and variance vectors of the Gaussian PDFs. These parameters are learned during model training with clean data and are assumed to be free of any uncertainty. If also the \( n \)-dimensional feature vector \( o_k = (o_{k1}, \ldots, o_{kn})^T \) at time frame \( k \) is free of any uncertainty its emission likelihood for HMM state \( j \) is given by

\[
p(o_k|\Lambda_j) = \sum_{r=1}^{R} c_{jr} p(o_k|\Lambda_{jr}),
\]

where \( R \) denotes the number of mixtures for the GMM. Due to the diagonal covariance assumption, each mixture component score \( p(o_k|\Lambda_{jr}) \) can be evaluated as a product over
the individual feature vector components:

$$p(o_k | \Lambda_{jr}) = \prod_{i=1}^{n} p(o_{ki} | \Lambda_{jri}) = \prod_{i=1}^{n} G(o_{ki}; \mu_{jri}, \sigma^2_{jri}), \quad (4)$$

with the univariate Gaussian $G(o; \mu, \sigma^2)$ defined as

$$G(o; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(o - \mu)^2}{2\sigma^2}\right). \quad (5)$$

2.2 State likelihood computation with uncertain data

The computation of the likelihood score as outlined above leads to severe performance degradation if the feature vector $o_k$ is observed in an environment different to the one used for learning the speech models $\Lambda_j$. In a noisy environment, some of the components in $o_k$ may be uncertain or be even completely missing such that (5) cannot or should not be evaluated. For such uncertain data, a framework has been developed by Morris et al. [2001] that allows to model the degree of uncertainty in the data values by treating the feature vector $o_k$ itself as a stochastic random variable. The uncertainty is specified quantitatively by modeling each vector component $o_{ki}$ with a PDF $o_{ki} \sim \varepsilon(o_{ki}|\Theta_{ki})$, where $\varepsilon(o_{ki}|\Theta_{ki})$ is called the data evidence model [Morris et al., 2001]. The evidence parameter set $\Theta_{ki}$ represents all available information about the clean data, such as the noisy and enhanced feature value as well as any knowledge about the range of possible feature values.

The emission likelihood for an uncertain feature $o_{ki}$ is then determined with the help of the evidence model $\varepsilon(o_{ki}|\Theta_{ki})$ by replacing $p(o_{ki} | \Lambda_{jri})$ in (4) with its expected value

$$E[p(o_{ki} | \Lambda_{jri}) | \varepsilon(o_{ki}|\Theta_{ki})] = \int_{-\infty}^{\infty} p(o_{ki} | \Lambda_{jri}) \varepsilon(o_{ki}|\Theta_{ki}) do_{ki}$$

$$= \int_{-\infty}^{\infty} G(o_{ki}; \mu_{jri}, \sigma^2_{jri}) \varepsilon(o_{ki}|\Theta_{ki}) do_{ki}. \quad (6)$$

This amounts to evaluating $G(o_{ki}; \mu_{jri}, \sigma^2_{jri})$ over all possible feature values $o_{ki}$, weighted by the corresponding data evidence $\varepsilon(o_{ki}|\Theta_{ki})$. A wide range of statistical distributions, such as Gaussian or Uniform PDFs, are available to express the confidence in each possible data value on the basis of the existing information. In its most general form, $\varepsilon(o_{ki}|\Theta_{ki})$ can be represented as a mixture PDF provided that the integral in (6) yields a closed form solution [Morris et al., 2001].
2. Evidence modeling for HMM state observation likelihoods

2.3 Review of previously proposed evidence PDFs

Several models for $\varepsilon(o_{ki}|\Theta_{ki})$ have been proposed in the literature related to missing data speech recognition. The following discussion reviews four types of evidence PDFs and presents their implementation as a decoding rule with respect to Equation (6).

Dirac-delta PDF

The most simple evidence PDF is the Dirac-delta function

$$\varepsilon_1(o_{ki}|\Theta_{ki}) = \delta(o_{ki} - \mu_{ki}; \mu_{ki})$$  \hspace{1cm} (7)

with $\Theta_{ki} = \{\mu_{ki}\}$ and

$$\delta(o_{ki} - \mu_{ki}; \mu_{ki}) = \begin{cases} +\infty, & o_{ki} = \mu_{ki}, \\ 0, & o_{ki} \neq \mu_{ki}, \end{cases}$$  \hspace{1cm} (8)

where $\mu_{ki}$ specifies the estimated clean feature value and $\int_{-\infty}^{\infty} \delta(o_{ki} - \mu_{ki}) \, do_{ki} = 1$. The application of the Dirac-delta evidence model in (6) results in the following decoding rule

$$\mathbb{E}[p(o_{ki}|\Lambda_{jri})|\varepsilon_1(o_{ki}|\Theta_{ki})] = \mathcal{G}(\mu_{ki}; \mu_{jri}, \sigma_{jri}^2),$$  \hspace{1cm} (9)

which is identical to the standard GMM likelihood computation with certain or complete data. The Dirac-delta PDF models the assumption that any acoustic mismatch between training and testing environment has been removed by the data pre-processing. The shape of the PDF indicates full confidence in $\mu_{ki}$ to be identical with the true clean data value. Although this level of clean speech restoration is impossible to achieve under practical conditions, the Dirac-delta model is often used without explicit declaration in most traditional speech enhancement-recognition schemes.

Gaussian PDF

A natural extension of the Dirac-delta model is the univariate Gaussian PDF

$$\varepsilon_2(o_{ki}|\Theta_{ki}) = \mathcal{G}(o_{ki}; \mu_{ki}, \sigma_{ki}^2),$$  \hspace{1cm} (10)

which explicitly takes imperfections of the speech enhancement process into consideration. Rather than just providing a point estimate $\mu_{ki}$ for the clean speech value, this model expresses the uncertainty associated with the speech enhancement process through its additional scatter parameter $\sigma_{ki}^2$. Application of $\varepsilon_2(o_{ki}|\Theta_{ki})$ in (6) leads to the following
decoding rule
\[
E \left[ p(o_{ki} | \Lambda_{jri}) | \varepsilon_{2} (o_{ki} | \Theta_{ki}) \right] = G(\mu_{ki}; \mu_{jri}, \sigma_{jri}^{2} + \sigma_{ki}^{2}),
\]
(11)
which is only slightly different from (9) but allows for an additional time-varying HMM state variance compensation. Despite its appeal the Gaussian evidence model has not found widespread application in spectral missing data recognition and has foremost been used in the related field of uncertainty decoding [Arrowood, 2003; Benítez et al., 2004; Deng et al., 2005; Kolossa et al., 2006; Stouten et al., 2004].

Dirac-delta-Uniform mixture PDF

Among the missing data literature, the most popular evidence model is still the Dirac-Uniform mixture
\[
\varepsilon_{3} (o_{ki} | \Theta_{ki}) = w_{ki} \delta (o_{ki} - \mu_{ki}; \mu_{ki}) + (1 - w_{ki}) U (o_{ki}; a_{ki}, b_{ki}),
\]
(12)
where \( \Theta_{ki} = \{ \mu_{ki}, w_{ki}, a_{ki}, b_{ki} \} \) [Barker et al., 2000; Harding et al., 2006; Morris et al., 2001; Palomäki et al., 2004a]. The idea behind this model is to represent each feature component as either clean or noisy. While the first case is modeled by the Dirac-delta component \( \delta (o_{ki} - \mu_{ki}; \mu_{ki}) \), the latter is realized by a Uniform distribution
\[
U (o_{ki}; a_{ki}, b_{ki}) = \frac{1}{b_{ki} - a_{ki}} I_{[a_{ki}, b_{ki}]} (o_{ki}),
\]
(13)
where \( a_{ki}, b_{ki} \) with \( b_{ki} > a_{ki} \) specify the distribution boundaries and \( I_{[a_{ki}, b_{ki}]} (o_{ki}) \) is the usual indicator function equal to 1, if \( o_{ki} \in [a_{ki}, b_{ki}] \), and 0 otherwise. The mixture weight \( w_{ki} \in [0, 1] \) controls the contribution of each mixture component and is usually estimated as a hard or soft time-frequency mask. Using \( \varepsilon_{3} (o_{ki} | \Theta_{ki}) \) in (6) results in the following decoding rule
\[
E \left[ p(o_{ki} | \Lambda_{jri}) | \varepsilon_{3} (o_{ki} | \Theta_{ki}) \right] = w_{ki} G(\mu_{ki}; \mu_{jri}, \sigma_{jri}^{2}) + (1 - w_{ki}) \frac{1}{b_{ki} - a_{ki}} \int_{a_{ki}}^{b_{ki}} G(o_{ki}; \mu_{jri}, \sigma_{jri}^{2}) do_{ki},
\]
(14)
which is known as bounded marginalization [Cooke et al., 2001]. Past research has shown that bounding the integral in (14) is especially beneficial for static filterbank energies because it provides an effective mechanism to include counter-evidence by penalizing all speech models \( \Lambda_{jri} \) with insufficient spectral energy [Cooke et al., 2001].

A large number of applications of \( \varepsilon_{3} (o_{ki} | \Theta_{ki}) \) can be found in both monaural [Barker et al., 2000; Cooke et al., 2001; Josifovski et al., 1999; Morris et al., 1998; Seltzer et al., 2004] and binaural [Harding et al., 2006; Kühne et al., 2007b; Palomäki et al., 2004b; Roman
et al., 2003] missing data systems. The latter, in particular, have achieved very promising results in dealing with multiple speakers in anechoic conditions. The Dirac-Uniform mixture is perfectly suited for this kind of data because the dominant points in the spectra of two speech sources hardly overlap in an echo-free mixture [Yilmaz and Rickard, 2004].

However, despite its success, recent studies have questioned the model’s practicality in more realistic scenarios that contain additive and reverberative noise distortions. For example, Palomäki et al. [2004b] and Roman et al. [2006] report serious performance degradations when the missing data decoder was trained on anechoic data but testing was performed with reverberant speech mixtures. In this case, reverberation not only affects the source localization cues used for estimating the mixture weight $w_{ki}$ but also the feature values $\mu_{ki}$ itself, making the assumption that data values are either clean or noisy rather implausible.

**Gauss-Uniform mixture PDF**

More recently, the Gauss-Uniform mixture PDF

$$\varepsilon_4(o_{ki}|\Theta_{ki}) = w_{ki}g(o_{ki};\mu_{ki},\sigma^2_{ki}) + (1 - w_{ki})u(a_{ki}; b_{ki}, c_{ki})$$

was proposed in Küehne et al. [2008b] as a simple extension to the Dirac-Uniform mixture. In Küehne et al. [2008b], the Dirac-delta mixture component was replaced by a univariate Gaussian in order to capitalize on the variance adaptation capabilities provided by the additional model parameter $\sigma^2_{ki}$. The resulting decoding rule is given by

$$E[p(o_{ki}|\Lambda_{jri})|\varepsilon_4(o_{ki}|\Theta_{ki})] = w_{ki}g(\mu_{jri}, \sigma^2_{jri} + \sigma^2_{ki}) + (1 - w_{ki})\frac{1}{b_{ki} - a_{ki}} \int_{a_{ki}}^{b_{ki}} g(o_{ki}; \mu_{jri}, \sigma^2_{jri}) do_{ki}$$

and combines the advantages of (11) and (14) together into one equation. It was shown in Küehne et al. [2008b], that the additional frame-by-frame variance compensation in (16) is superior to the Dirac-Uniform decoding rule, particularly in reverberant conditions.

However, the performance improvements strongly depend on the quality of the variance estimation, which may vary considerably under practical conditions. For example, for low SNR regions the feature means and variances can not be estimated with high precision. In cases, where $\sigma^2_{ki}$ is grossly over-estimated the decoder will no longer be able to discriminate between individual speech models, making the decoding process prone to insertion errors [Deng et al., 2005; Liao and Gales, 2008]. However, Cooke et al. [2001] point out that for spectral features there is additional information available in form of bounds on the spectro-temporal energy surface. In (16), these bounds are utilized only in the Uniform mixture component and, hence, have no effect when the mixture weight
favors the Gaussian PDF component. In order to rectify this problem, we investigate next the potential gains of modeling the bounded support of filterbank energies through a truncated Gaussian component instead.

3 A novel evidence model for spectral feature representations

This section presents the bounded-Gauss-Uniform mixture PDF as a new class of evidence model for spectral feature representations. After introducing the relevant equations, we show how the model’s parameters can be estimated with the aid of a multichannel blind source separation technique. The section concludes with an example illustrating the model parameters in the spectral feature space.

3.1 The bounded-Gauss-Uniform mixture PDF

In probability and statistics, the bounded Gaussian PDF $B$ is the PDF of a Gaussian distributed random variable whose value is either singly truncated from the left or right or from both sides. We define $B$ here as

$$B(o; \mu, \sigma^2, \alpha, \beta) = \frac{G(o; \mu, \sigma^2)}{\Phi(\beta) - \Phi(\alpha)} \mathbb{1}_{[\alpha, \beta]}(o), \quad (17)$$

where $\alpha$ and $\beta$ specify the lower and upper truncation points and $\Phi(x) = \int_{-\infty}^{x} G(o; \mu, \sigma^2) do$ is the cumulative distribution function (CDF) of $G(o; \mu, \sigma^2)$. The denominator in (17) is a normalization factor used to scale up the distribution such that $B$ properly integrates to one. This model is preferred over the simple Gaussian PDF when the tails of the distribution do not reflect the physical reality of the underlying random variable. For example, static filterbank energies are known to have a bounded support that is non-negative. We use this fact here as motivation to replace the unbounded mixture component $G$ in $\varepsilon_4(o_{ki}|\Theta_{ki})$ with its bounded counterpart $B$. The bounded-Gauss-Uniform mixture PDF is then given by

$$\varepsilon_5(o_{ki}|\Theta_{ki}) = w_{ki}B(o_{ki}; \mu_{ki}, \sigma_{ki}^2, \alpha_{ki}, \beta_{ki}) + (1 - w_{ki})U(o_{ki}; a_{ki}, b_{ki}), \quad (18)$$

where $\Theta_{ki} = \{\mu_{ki}, \sigma_{ki}^2, \alpha_{ki}, \beta_{ki}, w_{ki}, a_{ki}, b_{ki}\}$. In practice, we may assume that the truncation points of $B$ are identical with the bounds of the Uniform distribution $U$, e.g., $\alpha \equiv a$ and $\beta \equiv b$, because it is often impossible to guarantee any other bounds on the clean feature value. A graphical illustration of the bounded Gauss-Uniform model for the reduced parameter set $\Theta_{ki} = \{\mu_{ki}, \sigma_{ki}^2, a_{ki}, b_{ki}\}$ is given in Fig. 1.

Next, we derive the decoding rule for the new evidence model $\varepsilon_5(o_{ki}|\Theta_{ki})$ by solving
3. A novel evidence model for spectral feature representations

Figure 1: Evidence model $\varepsilon_{5}(o_{ki}|\Theta_{ki})$ represented in the feature space as a two-component mixture of a bounded Gaussian PDF and a Uniform distribution.

the expectation integral (6). We start by writing (6) as the following weighted sum:

$$
E \left[ p(o_{ki}|\Lambda_{jr_i})|\varepsilon_{5}(o_{ki}|\Theta_{ki}) \right] = w_{ki} I_{ki}^{(B)} + (1-w_{ki}) I_{ki}^{(U)},
$$

where the integrals $I_{ki}^{(B)}$ and $I_{ki}^{(U)}$ for the individual mixture components of $\varepsilon_{5}(o_{ki}|\Theta_{ki})$ are given by

$$
I_{ki}^{(B)} = \int_{-\infty}^{\infty} G(o_{ki}; \mu_{jr_i}, \sigma^2_{jr_i}) B(o_{ki}; \mu_{ki}, \sigma^2_{ki}, \alpha_{ki}, \beta_{ki}) do_{ki},
$$

$$
I_{ki}^{(U)} = \int_{-\infty}^{\infty} G(o_{ki}; \mu_{jr_i}, \sigma^2_{jr_i}) U(o_{ki}; a_{ki}, b_{ki}) do_{ki}.
$$

Using the definition (17), the likelihood contribution of the bounded Gaussian mixture component $\mathcal{B}$ can be computed straightforwardly as

$$
I_{ki}^{(B)} = \int_{-\infty}^{\infty} G(o_{ki}; \mu_{jr_i}, \sigma^2_{jr_i}) \frac{G(o_{ki}; \mu_{ki}, \sigma^2_{ki})}{\Phi(\beta_{ki}) - \Phi(\alpha_{ki})} \mathbb{1}_{[\alpha_{ki}, \beta_{ki}]}(o_{ki}) do_{ki}
$$

$$
= \frac{1}{\Phi(\beta_{ki}) - \Phi(\alpha_{ki})} \int_{\alpha_{ki}}^{\beta_{ki}} G(o_{ki}; \mu_{jr_i}, \sigma^2_{jr_i}) G(o_{ki}; \mu_{ki}, \sigma^2_{ki}) do_{ki}
$$

$$
= \frac{G(\mu_{ki}; \mu_{jr_i}, \sigma^2_{jr_i} + \sigma^2_{ki})}{\Phi(\beta_{ki}) - \Phi(\alpha_{ki})} \int_{\alpha_{ki}}^{\beta_{ki}} G(o_{ki}; \tilde{\mu}, \tilde{\sigma}^2) do_{ki}
$$

(20)
with
\[ \bar{\mu} = \frac{\mu_{jri}\sigma_{ki}^2 + \mu_{ki}\sigma_{jri}^2}{\sigma_{ki}^2 + \sigma_{jri}^2} \quad \text{and} \quad \bar{\sigma}^2 = \frac{\sigma_{ki}^2\sigma_{jri}^2}{\sigma_{ki}^2 + \sigma_{jri}^2}. \]

The last line in (20) was obtained by using the following well known result [Deng et al., 2005; Morris, 1999]:
\[ G(o; \mu_1, \sigma_1^2)G(o; \mu_2, \sigma_2^2) = G(\mu_1; \mu_2, \sigma_1^2 + \sigma_2^2)G(o; \bar{\mu}, \bar{\sigma}^2) \]
where
\[ \bar{\mu} = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad \text{and} \quad \bar{\sigma}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \]

The likelihood contribution of the Uniform mixture component \( U \) is given by
\[
I_{ki}^{(U)} = \int_{-\infty}^{\infty} G(o_{ki}; \mu_{jri}, \sigma_{jri}^2) \frac{1}{b_{ki} - a_{ki}} \mathbb{I}_{[\alpha_{ki}, \beta_{ki}]}(o_{ki}) do_{ki} = \frac{1}{b_{ki} - a_{ki}} \int_{a_{ki}}^{b_{ki}} G(o_{ki}; \mu_{jri}, \sigma_{jri}^2) do_{ki}. \quad (21)
\]

Substituting the solutions for \( I_{ki}^{(B)} \) and \( I_{ki}^{(U)} \) back into (19) yields the decoding rule for the new bounded Gauss-Uniform mixture PDF, shown in Equation (22) on the top of the next page.

Note that (22) is almost identical to the Gauss-Uniform mixture decoding rule in (16). For static filterbank energies the additional integration in the first term of (22) contributes some discriminatory information by penalizing speech models that are inconsistent with the integration limits \( \alpha_{ki} \) and \( \beta_{ki} \). Equation (22) becomes equivalent to (16) when these bounds approach infinity, thus reducing both the integral and normalization factor to unity. Furthermore, several other previously proposed evidence PDFs, such as the Dirac-Uniform mixture PDF or the Gaussian PDF, can also be derived as special cases from (22). On the downside, the new model increases the computational complexity. Extra computations are mainly required for the calculation of \( \bar{\mu} \) and \( \bar{\sigma}^2 \) as well as the additional integration in the first term of (22). However, because the univariate integrals can be evaluated quite efficiently through tabulated Gaussian error functions [Cooke et al., 2001; Morris et al., 1998] the actual increase in computation time remains modest.
3. A novel evidence model for spectral feature representations

\[ E \left[ p(o_{ki}|\Lambda_{jri})|\varepsilon_{5}(o_{ki}|\Theta_{ki}) \right] = w_{ki} \frac{\mathcal{G}(\mu_{ki}; \mu_{jri}, \sigma^2_{jri} + \sigma^2_{ki})}{\Phi(\beta_{ki}) - \Phi(\alpha_{ki})} \int_{\alpha_{ki}}^{\beta_{ki}} \mathcal{G}(o_{ki}; \bar{\mu}, \bar{\sigma}^2) do_{ki} + (1 - w_{ki}) \frac{1}{b_{ki} - a_{ki}} \int_{a_{ki}}^{b_{ki}} \mathcal{G}(o_{ki}; \mu_{jri}, \sigma^2_{jri}) do_{ki} \]

(22)

3.2 Evidence PDF parameter estimation

Decoding with evidence models requires an estimation of the parameter set \( \Theta \) in practice. In this study, we utilize the separation outcomes of a recently developed multi-channel BSS technique [Kühne et al., 2009] for this purpose (see Fig. 2). Our BSS approach is based on a combination of beamforming and time-frequency masking and employs a novel fuzzy cluster algorithm for estimating contextually constrained time-frequency masks in reverberant conditions. A detailed description of the method is beyond the scope of this paper but can be found in Kühne et al. [2009]. We would like to point out that the following procedure for estimating \( \Theta \) is fairly general and, in principle, almost any BSS technique could be used instead.

Consider a reverberant multi-source scenario, where \( M \) speech-plus-noise mixture observations \( X_{mkl} \) are recorded by a small microphone array with \( m = 1, \ldots, M \) sensor elements. Here, \( X_{mkl} \) denotes the magnitude spectrum of the \( m \)-th mixture observation in the Short-time-Fourier-transform (STFT) domain. The subscripts \( k \) and \( l \) specify the time and frequency index on the linear STFT frequency scale. Let us assume that the BSS separation results are available in the form of two magnitude spectra \( T_{kl} \) and \( I_{kl} \), where the former denotes the estimated target signal used for recognition and the latter represents an estimate of the noise intrusion. Furthermore, let \( M_{kl} \) be the estimated time-frequency mask marking the dominant points for the target speaker \( T_{kl} \). If the BSS algorithm in question does not automatically estimate this mask, the BSS output spectra \( T_{kl} \) and \( I_{kl} \) may be utilized to obtain \( M_{kl} \) during a post-processing step [Kolossa and Orglmeister, 2004].

Next, we describe how the evidence PDF parameters of the bounded-Gauss-Uniform mixture model can be estimated for two complementary spectral feature streams. The
first stream consists of logarithmic compressed filterbank energies (FBEs), which measure the absolute energy level in each filterbank channel on the mel-frequency scale. The second stream models the slope of the FBE envelope across frequency, which we implement via a simple linear transformation of the FBEs, called frequency filtering (FF). In particular, we use the FF2 technique first proposed in Nadeu et al. [2001]. Note that both feature streams keep the noise corruption localized within a small frequency band. In our experimental evaluation in Section 4 we then demonstrate how the different parameters in $\Theta$ affect the ASR performance of each of these two spectral feature types.

Now consider the random variable $o_k$ at time frame $k$ consisting of the two feature streams $o_k^{(s)}$, $s \in \{\text{FBE, FF2}\}$. Each stream is modeled as a $n$-dimensional vector $o_k^{(s)}= (o_k^{(s)}(1), o_k^{(s)}(2), \ldots, o_k^{(s)}(n))^T$, where we assume that the two streams as well as the individual vector components are statistically independent. The parameter set for both feature streams is denoted as $\Theta^{(s)}_{ki} = \{\mu_{ki}^{(s)}, \sigma^2_{ki}^{(s)}, \alpha_{ki}^{(s)}, \beta_{ki}^{(s)}, w_{ki}^{(s)}\}$ respectively.

**Feature means - $\mu_{ki}^{(s)}$**

We start with the bounded-Gaussian mixture component $B$ and estimate $\mu_{ki}^{(s)}$ for both feature streams $s \in \{\text{FBE, FF2}\}$.

The mean for the static FBE component is simply the recovered target estimate $T_{kl}$ of the BSS algorithm after the usual mel-frequency conversion. Following Young et al. [2006], we calculate this as

$$\mu_{kb}^{(\text{FBE})} = \log \left( \max \left\{ \sum_{l} \lambda_{bl} T_{kl}, 1 \right\} \right), \quad b = 1, \ldots, B \quad (23)$$

where $\lambda_{bl}$ is the triangular filter of the $b$-th channel in the mel-scale filter bank.

The static means of the FF2 feature stream are derived by the following frequency filtering operation [Nadeu et al., 2001]:

$$\mu_{kb}^{(\text{FF2})} = \begin{cases} 
\mu_{k2}^{(\text{FBE})}, & b = 1 \\
\mu_{kb+1}^{(\text{FBE})} - \mu_{k(b-1)}^{(\text{FBE})}, & b \in \{2, \ldots, B - 1\} \\
-\mu_{k(B-1)}^{(\text{FBE})}, & b = B \end{cases} \quad (24)$$

First order regression coefficients are then appended to both static feature streams using the standard regression formula [Young et al., 2006]

$$\mu_{k(b+\frac{1}{2})}^{(s)} = \frac{\sum_{\theta=1}^{\Theta} \theta \left( \mu_{k(b+\theta)}^{(s)} - \mu_{k(b-\theta)}^{(s)} \right)}{2 \sum_{\theta=1}^{\Theta} \theta^2}, \quad b = 1, \ldots, B; \forall s \quad (25)$$

where $\Theta$ denotes the half-size of the temporal regression window.
3. A novel evidence model for spectral feature representations

**Feature variances - \( \sigma^2_{ki} \)**

Next, we consider a heuristic approach in order to determine the feature variance parameter \( \sigma^2_{ki} \) associated with \( B \). This is accomplished by means of a simple spectral subtraction scheme [Vaseghi, 1996], which utilizes the mixture observations \( X_{mkl} \) as well as the interference estimate \( I_{kl} \) to construct an additional target estimate \( T_{mkl} \) at each microphone. After converting each target estimate to the mel-frequency domain (see previous paragraph), we estimate the feature uncertainty as the (weighted) variance among all \( M + 1 \) feature values.

Let \( \xi_{kl} = 20 \log_{10} \left( T_{kl}/I_{kl} \right) \) be an estimate of the SNR on the linear frequency axis and let \( \varsigma_{kb} \) be its equivalent on the mel-frequency scale. Given \( X_{mkl} \) and \( I_{kl} \), we use the following spectral subtraction procedure to compute \( M \) additional estimates of the target signal as

\[
T_{mkl} = \begin{cases} 
X_{mkl}^2 - \xi_{kl}I_{kl}^2 \left( \frac{1}{2} \right), & \text{if } X_{mkl}^2 > \xi_{kl}I_{kl}^2, \forall m \\
\varepsilon X_{mkl} \left( \frac{1}{2} \right), & \text{otherwise}
\end{cases}
\tag{26}
\]

where \( \xi_{kl} = 1 + \left( 1 + \exp(\varsigma_{kb}) \right)^{-1} \) is a SNR dependent subtraction factor with \( 1 \leq \xi_{kl} \leq 2 \) [Vaseghi, 1996] and \( \varepsilon \) is a spectral floor parameter, fixed at 0.01. After converting \( T_{mkl} \) to \( \mu_{mki}^{(s)} \) via (23), (24) and (25), we estimate the feature uncertainty in the mel-frequency domain as

\[
\sigma^2_{ki} = \frac{1}{M} \sum_{m=1}^{M} \rho_{ki} \left( \mu_{mki}^{(s)} - \mu_{ki}^{(s)} \right)^2, \quad i = 1, \ldots, n; \forall s.
\tag{27}
\]

Here, \( \rho_{ki} = \left\{ 1 + \exp[0.25(\varsigma_{kb(i)} - 3)] \right\}^{-2} \) is an empirical SNR weighting factor with the function \( \nu(i) = i - 1_{[1,B]}(i)B \) performing a mapping between static and dynamic feature indices. The purpose of the weighting factor \( \rho_{ki} \) is to bias the variances \( \sigma^2_{ki} \) towards zero in high SNRs whilst retaining higher uncertainty values in mid and low SNRs. Similar ad-hoc weighting schemes have been used in related work on uncertainty decoding [Benitez et al., 2004; Kolossa et al., 2008; Stouten et al., 2004].

**Integration limits - \( \alpha^{(s)}_{ki}, \beta^{(s)}_{ki} \)**

For determining the integration bounds, we make the above mentioned assumption that the truncation points of \( B \) are identical with the bounds of the Uniform distribution \( U \). Similar to previous work [Cooke et al., 2001; Morris et al., 2001], we declare the static clean FBE value to be confined to the interval between

\[
\alpha^{(\text{FBE})}_{kb} = 0, \quad b = 1, \ldots, B
\tag{28}
\]
and

\[ \beta_{kb}^{(\text{FBE})} = \log \left( \max \left\{ \sum_l \lambda_{kl} \tilde{X}_{kl}, 1 \right\} \right), \quad b = 1, \ldots, B \]  

(29)

where \( \tilde{X}_{kl} \) denotes mixture observation with the largest magnitude as recorded by the microphone array.

For the FF2 feature stream, we derive the integration limits based on the bounds (28) and (29) of the static FBE feature stream. This is achieved by considering the largest and smallest possible feature values that can be obtained for \( \mu_{kb}^{(\text{FBE})} \) given the bounds on \( \mu_{kb}^{(\text{FBE})} \) during the frequency filtering operation in (24).

The corresponding integration limits for the dynamic features are obtained in a similar fashion by utilizing the lower and upper bounds of the static features in (25).

We remark that only the relatively tight bounds on static FBE features are expected to contribute significant discriminatory information during decoding. Because the bounds for dynamic and FF2 features are derived from (28) and (29), the integration interval for these features increases as a consequence of the uncertainties associated with the additional feature transformations in (24) and (25).

**Mixture weight - \( w_{ki} \)**

The static mixture weights \( w_{ki} \) are obtained by converting the high resolution mask of the target speaker \( M_{kl} \) to the mel-frequency scale using the same triangular filter weights \( \lambda_{bl} \) as in (23) [Kühne et al., 2007a]. More specifically, the static weights for the FBE feature stream are obtained as

\[ w_{kb}^{(\text{FBE})} = \frac{\sum_l \lambda_{bl} M_{kl}}{\sum_l \lambda_{bl}}, \quad b = 1, \ldots, B \]  

(30)

and are subsequently used to compute the weights of the FF2 stream as

\[
\begin{align*}
    w_{kb}^{(\text{FF2})} &= \begin{cases} 
        \left[ \left( w_{k(b-1)}^{(\text{FBE})} w_{k(b+1)}^{(\text{FBE})} \right)^{1/2}, & b \in \{2, \ldots, B-1\} \\
        0, & b \in \{1, B\} 
    \end{cases}, \\
    & b = 1, \ldots, B 
\end{align*}
\]  

(31)

Similarly, the dynamic mixture weights are determined using the following geometric average

\[ w_{k(b+t/2)}^{(s)} = \left[ \prod_{t=-\tau,t\neq 0}^{\tau} w_{(k+t)b}^{(s)} \right]^{1/\tau}, \quad b = 1, \ldots, B; \forall s. \]  

(32)

This product of static mixture weights will only then indicate a reliable dynamic feature if all individual static weights \( w_{(k+t)b}^{(s)} \) are marked as highly trustworthy.
We conclude this section with Fig. 3 in which an example is shown for the parameter estimation of the static FBE feature stream in the 12-th mel-frequency band of the TIDIGIT utterance "111a". Beside the estimated model parameters we also show the true, clean speech value for comparison.

4 Experimental evaluation

The proposed system was evaluated in terms of speech recognition accuracy on a connected digit task previously used for assessing the performance of binaural segregation models [Harding et al., 2006; Palomäki et al., 2004b; Roman et al., 2006]. A number of experiments were conducted to measure the influence of different room reverberation times, spatial separation angles and various types of noise intrusions on the recognition rate. In order to compare our results with previous work, the room layout and data material closely followed the specifications given in Palomäki et al. [2004b].

4.1 Experimental setup

Room layout and data generation

Sound propagation was simulated at a sampling frequency of 8 kHz for a small rectangular room of dimensions 6 m x 4 m x 3 m (length x width x height). Wall reflections were modeled by a recently proposed version of the widely used image method for simulating small-room acoustics [Lehmann and Johansson, 2008]. The room reverberation time RT$_{60}$\(^1\) was adjusted for five different reverberant scenarios with RT$_{60}$ ∈

---

\(^1\)RT$_{60}$ specifies the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.
A small linear microphone array with six sensor elements was positioned in the middle of the room at a height of 2 m. The speech and noise sources were placed at different horizontal angles facing array broadside and a distance of 1.5 m from the array center. The test set consisted of 240 utterances of four male TIDIGIT [Leonard, 1984] speakers (‘ah’, ‘ar’, ‘at’, ‘be’) mixed with one of three different noise intrusions at SNR levels of 0 dB, 10 dB or 20 dB. All three noise files (male and female TIMIT [Garofolo et al., 1993] speaker, rock music [Cooke, 1993]) were identical to those used in Palomäki et al. [2004b]. Each mixture was pre-emphasized with a pre-emphasis coefficient of 0.97 before splitting the signal into frames using a 25 ms Hamming window and a 10 ms frame shift.

**Speech recognition back-end**

The training set for learning the anechoic, clean speech models consisted of 4235 utterances spoken by 55 male TIDIGIT speakers. The Hidden Markov Model Toolkit (HTK) [Young et al., 2006] was used to train 11 word HMMs (‘1’-‘9’, ‘oh’, ‘zero’) each with eight emitting states and two silence models (‘sil’, ‘sp’) with three and one state. All HMMs followed standard left-to-right models without skips using continuous Gaussian densities with diagonal covariance matrices and 10 mixture components. Two different sets of acoustic models were created. The first set of HMMs was used as a baseline system and employed 13 MFCCs derived from a 32-channel HTK mel-filterbank together with their delta and acceleration coefficients [Young et al., 2006]. To provide robustness against convolutional distortions cepstral mean normalization (CMN) was applied. This kind of baseline has been used in a number of previous missing data studies [Cooke et al., 2001; Harding et al., 2006; Morris et al., 2001; Palomäki et al., 2004b] in order to demonstrate the performance of state-of-the-art features in noise. The second model set was designed for the missing data decoder and employed spectral rather than cepstral features. HTK’s streaming capability was utilized for implementing the 128-dimensional two-feature stream model outlined in Section 3.2. The recognition accuracy on the clean, anechoic test set was 98.3 % for the cepstral baseline while the spectral decoder achieved 97.3 % for the FBE features, 98.3 % for the FF2 stream and 98.9 % for the combined feature set.

**4.2 Results**

**Choice of evidence PDF and feature type**

The first experiment established the performance of various evidence models when applied to the individual as well as the combined feature streams. The reverberation time RT$_{60}$ was fixed at 300 ms. Two male speakers were mixed at a SNR of 0 dB with 40° spa-
4. Experimental evaluation

Table 1: HTK percent accuracy (ACC) and percent correctness (COR) score for several types of evidence models in the presence of an interfering male speaker for a room reverberation time RT<sub>60</sub> of 300 ms. Results are shown for filterbank energy features (FBE), frequency filtered FBEs (FF2) and the combined feature streams (FBE+FF2).

<table>
<thead>
<tr>
<th>Evidence model</th>
<th>Spectral feature type</th>
<th>FBE</th>
<th>FF2</th>
<th>FBE+FF2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ACC  %</td>
<td>COR %</td>
<td>ACC  %</td>
</tr>
<tr>
<td>Dirac</td>
<td></td>
<td>1.0</td>
<td>50.3</td>
<td>29.8</td>
</tr>
<tr>
<td>Gauss</td>
<td></td>
<td>6.0</td>
<td>59.8</td>
<td>29.2</td>
</tr>
<tr>
<td>BGauss</td>
<td></td>
<td>25.1</td>
<td>65.5</td>
<td>29.2</td>
</tr>
<tr>
<td>Dirac-Uniform</td>
<td></td>
<td>50.3</td>
<td>69.0</td>
<td>27.4</td>
</tr>
<tr>
<td>Gauss-Uniform</td>
<td></td>
<td>57.7</td>
<td>72.4</td>
<td>48.7</td>
</tr>
<tr>
<td>BGauss-Uniform</td>
<td></td>
<td>65.1</td>
<td>76.4</td>
<td>48.7</td>
</tr>
</tbody>
</table>

Table 1 shows the ASR performance for six evidence PDFs of increasing complexity. Several observations can be made regarding the use of individual evidence PDFs and feature streams. First, the more complex the evidence PDF the higher the achieved recognition scores for both individual and combined feature streams. Second, the best results were obtained with the FBE+FF2 feature set using a bounded-Gauss-Uniform mixture evidence model. We also note that the use of bounded mixture components was most effective for the FBE stream while for the FF2 stream performance remained unaffected by the rather loose bounds on the FBE slopes. On the other hand, the FF2 stream showed an improved performance for evidence models with a Gaussian mixture component indicating that this type of feature strongly benefits from the dynamic compensation of HMM state variances. With respect to the combined FBE+FF2 feature stream, we observe that a simple stream combination (Dirac model) did not result in an improved recognition performance. Only the two-component evidence mixture PDFs could significantly improve both recognition scores in comparison with the individual stream performances.

We conclude that despite the complementary acoustic information provided by each feature set, an appropriate type of evidence model seems to be required for realizing the full potential of this spectral feature combination.

Influence of reverberation time

The second experiment investigated the effect of reverberation on ASR performance. The reverberation time RT<sub>60</sub> was varied between 0 ms (anechoic) and 600 ms (‘live’ office). Two male speech sources were mixed at SNRs of 0, 10 and 20 dB for each reverberant condition. The spatial separation between sources was 40°, with the TIDIGIT target speaker...
Table 2: HTK percent accuracy (%) in the presence of an interfering male speaker for six room reverberation times. Results are shown for the missing data recognizer using three types of evidence PDFs and an uncompensated MFCC-CMN baseline scoring on the sound mixture.

<table>
<thead>
<tr>
<th>Evidence model</th>
<th>Reverberation time RT60 (ms)</th>
<th>0</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>450</th>
<th>600</th>
</tr>
</thead>
<tbody>
<tr>
<td>No processings</td>
<td>Baseline</td>
<td>2.6</td>
<td>5.5</td>
<td>10.3</td>
<td>14.4</td>
<td>14.9</td>
<td>13.0</td>
</tr>
<tr>
<td>A posteriori</td>
<td>Dirac-Uniform</td>
<td>98.7</td>
<td>95.9</td>
<td>84.7</td>
<td>63.6</td>
<td>43.3</td>
<td>29.7</td>
</tr>
<tr>
<td></td>
<td>Gauss-Uniform</td>
<td>98.2</td>
<td>97.1</td>
<td>90.8</td>
<td>72.7</td>
<td>51.9</td>
<td>33.6</td>
</tr>
<tr>
<td></td>
<td>BGauss-Uniform</td>
<td>98.6</td>
<td>97.5</td>
<td>91.9</td>
<td>78.5</td>
<td>55.9</td>
<td>41.0</td>
</tr>
<tr>
<td>A priori</td>
<td>Dirac-Uniform</td>
<td>98.6</td>
<td>97.3</td>
<td>90.2</td>
<td>82.4</td>
<td>80.4</td>
<td>72.4</td>
</tr>
<tr>
<td></td>
<td>Gauss-Uniform</td>
<td>98.3</td>
<td>98.5</td>
<td>97.7</td>
<td>97.6</td>
<td>96.7</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>BGauss-Uniform</td>
<td>98.3</td>
<td>98.5</td>
<td>97.8</td>
<td>97.6</td>
<td>96.7</td>
<td>95.2</td>
</tr>
</tbody>
</table>

located at 20° azimuth and the interfering TIMIT speaker at −20° azimuth.

Table 2 shows the ASR performance for each room condition when the a posteriori evidence model parameters were estimated as described in Section 3.2. Additionally, we show the performance of a priori evidence models, where knowledge about the clean speech and noise signal was utilized for deriving the PDF parameters. In particular, ideal binary mixture weights were constructed as in Kühne et al. [2008a] by comparing the local SNR in each time-frequency slot. The a priori variances were estimated as in Deng et al. [2005] by squaring the distance between the estimated feature mean and the true clean feature value.

Looking at Table 2, we note that the missing data decoder with a posteriori evidence models achieved substantial improvements in recognition accuracy over the cepstral baseline for all room reverberation times. However, caused by the limitations of our source separation technique, the performance of these models degraded significantly as the reverberation time increased. While the best results were obtained with the new bounded-Gauss-Uniform mixture PDF, the worst performance was achieved by the Dirac-Uniform mixture model. A similar trend was observed among the a priori evidence models. The bounded-Gauss-Uniform as well as the Gauss-Uniform mixture PDFs remained nearly at ceiling performance revealing an impressive robustness against both additive and reverberative distortions. As has been criticized in Palomäki et al. [2004b] and Roman et al. [2006], the Dirac-Uniform mixture model with its simple data clean-or-corrupted assumption was unable to deal with the increasing amount of spectral distortions and as a consequence performed less robust for higher reverberation times.
4. Experimental evaluation

Influence of spatial separation

The third experiment evaluated the effect of the spatial separation between speech and noise sources on ASR performance. Two speech sources were positioned symmetrically about the median plane at azimuth angles of \((-5^\circ, 5^\circ), (-10^\circ, 10^\circ)\) and \((-20^\circ, 20^\circ)\).

Table 3 shows the obtained recognition performance of our system in comparison with the binaural missing data system of Palomäki et al. [2004b] for the three spatial separation angles. As is evident from the table, increasing the spatial separation between target source and interfering speech improved the recognition performance for all evidence models, most notably in the 0 and 10 dB SNR conditions. Apart from the Dirac-Uniform mixture in 10\(^\circ\) angular separation, the performance of the remaining two evidence models exceeded that of the cepstral baseline in all other conditions. The bounded-Gauss-Uniform mixture PDF performed best while the runner up was again the Gauss-Uniform mixture model followed by the Dirac-Uniform mixture. The performance differential between these models became more prominent the lower the separation angle and the lower the SNR level. This clearly indicates the potential of more advanced evidence PDFs to provide higher noise robustness in challenging conditions. In comparison with Palomäki et al. [2004b], our proposed system achieved better recognition scores in all conditions even when starting from a lower baseline.

Influence of noise type

In our last experiment, we studied the impact of different kinds of noise intrusion on ASR performance. The angular separation between speech and noise source was 40\(^\circ\) with speech and noise source placed at azimuths of 20\(^\circ\) and -20\(^\circ\), respectively.

Table 4 shows the obtained recognition performance of our system in comparison with the system of Palomäki et al. [2004b] for three different types of noise intrusions. When considering the most difficult SNR scenario of 0 dB, the results suggest that our system performs best with the female speech intrusion and worst with rock music. This is in line with Palomäki et al. [2004b], who reached similar conclusions. For the 10 and 20 dB cases the performance differences in our system were less obvious. With respect to the type of evidence model, the best results were generally obtained by the bounded-Gauss-Uniform mixture, followed by the Gauss-Uniform and the Dirac-Uniform mixture PDFs. Only for the 0 dB rock music case the Gauss-Uniform model performed best. In comparison with Palomäki et al. [2004b], our missing data system achieved considerably higher recognition rates, particularly for the lower SNR conditions.
Table 3: HTK percent accuracy (%) for three angular separations between target speech and an interfering male speaker. Results are shown for the missing data recognizer using three types of evidence PDFs and an uncompensated MFCC-CMN baseline scoring on the reverberant sound mixture ($R_{60} = 300$ ms).

<table>
<thead>
<tr>
<th>Separation ($^\circ$)</th>
<th>Ref.</th>
<th>Evidence model</th>
<th>SNR level (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Palomäki et al. [2004b]</td>
<td>Baseline*</td>
<td>17.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform*</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>Palomäki et al. [2004b]</td>
<td>Baseline</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gauss-Uniform</td>
<td>14.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BGauss-Uniform</td>
<td>26.4</td>
</tr>
<tr>
<td>20</td>
<td>Palomäki et al. [2004b]</td>
<td>Baseline*</td>
<td>14.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform*</td>
<td>38.7</td>
</tr>
<tr>
<td></td>
<td>Palomäki et al. [2004b]</td>
<td>Baseline</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform</td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gauss-Uniform</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BGauss-Uniform</td>
<td>40.8</td>
</tr>
<tr>
<td>40</td>
<td>Palomäki et al. [2004b]</td>
<td>Baseline*</td>
<td>14.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform*</td>
<td>54.9</td>
</tr>
<tr>
<td></td>
<td>Palomäki et al. [2004b]</td>
<td>Baseline</td>
<td>14.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform</td>
<td>63.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gauss-Uniform</td>
<td>72.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BGauss-Uniform</td>
<td>78.5</td>
</tr>
</tbody>
</table>

* as reported in Palomäki et al. [2004b]

5 General discussion

Our discussion starts with a brief summary of the main findings before establishing their significance in the context of previous work. We then conclude the paper by commenting on several limitations and pointing out further research directions.

As the main contribution of this paper, we proposed with the bounded-Gauss-Uniform mixture PDF a novel type of evidence model for missing data ASR with uncertain data. We described a simple method to estimate the mixture PDF parameters using information provided by a multi-channel BSS front-end. The model’s performance was assessed in a variety of test conditions to verify whether it can deal with mixtures of speech and various noise types at different SNRs, room reverberation times and angular separation angles.

Among the tested evidence PDFs, the proposed bounded-Gauss-Uniform mixture model consistently achieved the highest recognition results. Performance gains were most evident for the more challenging setups with higher reverberation times and lower spatial separation of the sources. This suggests a great potential for more complex types of evidence models to perform well under additive and reverberative noise distortions.
5. General discussion

Table 4: HTK percent accuracy (%) for a connected digit recognition task with various types of noise intrusions. Results are shown for the missing data recognizer using three types of evidence PDFs and an uncompensated MFCC-CMN baseline scoring on the reverberant sound mixture (RT$_{60}$ = 300 ms).

<table>
<thead>
<tr>
<th>Noise type</th>
<th>Ref.</th>
<th>Evidence model</th>
<th>SNR level (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Male speech</td>
<td>Palomäki et al. [2004b] Baseline$^a$</td>
<td>14.3</td>
<td>47.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform$^a$</td>
<td>54.9</td>
</tr>
<tr>
<td></td>
<td>Male speech Baseline</td>
<td>14.4</td>
<td>47.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform</td>
<td>63.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gauss-Uniform</td>
<td>72.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BGauss-Uniform</td>
<td>78.5</td>
</tr>
<tr>
<td></td>
<td>Female speech Baseline</td>
<td>16.5</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform$^a$</td>
<td>53.9</td>
</tr>
<tr>
<td></td>
<td>Female speech Baseline</td>
<td>18.2</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform</td>
<td>73.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gauss-Uniform</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BGauss-Uniform</td>
<td>81.4</td>
</tr>
<tr>
<td></td>
<td>Palomäki et al. [2004b] Baseline$^a$</td>
<td>9.5</td>
<td>50.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform$^a$</td>
<td>32.7</td>
</tr>
<tr>
<td></td>
<td>Rock music Baseline</td>
<td>13.0</td>
<td>50.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirac-Uniform</td>
<td>63.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gauss-Uniform</td>
<td>70.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BGauss-Uniform</td>
<td>68.5</td>
</tr>
</tbody>
</table>

$^a$ as reported in Palomäki et al. [2004b]

Furthermore, we found that the more complex evidence models also performed favorably when compared with the related binaural missing data system of Palomäki et al. [2004b]. The observed gains in recognition accuracy may result from several key differences in which our system improves over that of Palomäki et al. [2004b]. Firstly, while Palomäki et al. did not attempt any form of speech enhancement in their processing and instead concentrated on localizing the noise corruption in the time-frequency plane, our system enhances the speech signal prior to recognition. Secondly, Palomäki et al. performed decoding through hard bounded marginalization using a Dirac-Uniform mixture PDF with binary mixture weights. In contrast, our system employs a more complex evidence model and relies on soft bounded marginalization, which has been shown to significantly improve the performance of missing data decoders [Barker et al., 2000; Morris et al., 2001].

In terms of other related work, similar HMM model compensation techniques have been proposed for dealing with uncertain observations [Arrowood, 2003; Deng et al., 2005; Kolossa et al., 2008]. For example, in Deng et al. [2005] a method, called uncertainty decoding, was presented that uses a univariate Gaussian PDF for modeling the distribution of the features after speech enhancement. Their method is based on a statistical single-channel enhancement algorithm and relies on a probabilistic and parametric
model of speech distortion for feature mean and variance estimation.

Similar HMM variance compensation schemes have also been proposed for the cepstral feature space [Kolossa et al., 2008; Srinivasan and Wang, 2006]. While such an approach offers important advantages (e.g., a nearly decorrelated feature space) it also suffers from some drawbacks. For example, working in the cepstral feature space requires the feature uncertainties, which are often estimated in the spectral domain, to be propagated to the cepstral domain. During this process the linear mixing of several spectral bands leads to an increase in the uncertainty of the cepstral feature coefficients. Also, no effective bounds for the clean feature value are known in the cepstral feature space, thus restricting the choice of possible evidence models to the class of unbounded PDFs.

At present, the question whether model compensation should be performed in the spectral or cepstral feature domain is still subject to research. In a recently published study by Srinivasan and Wang [2007], the performance of missing data recognition and cepstral uncertainty decoding was compared for a low vocabulary connected digit task. The study reported that the uncertainty decoder outperformed the missing data approach for SNR levels of 10 and 5 dB whilst achieving comparable recognition performance at 0 dB. However, the missing data decoder in their study used a Dirac-Uniform mixture model with binary mixture weights. With respect to the findings of this paper, it would be interesting to repeat these experiments using an extended spectral feature set and a different type of evidence model. For example, the use of the bounded-Gauss-Uniform mixture PDF would equip the missing data recognizer with the same ability to perform a frame-by-frame HMM variance compensation as used by the uncertainty decoder.

With respect to the study’s limitations, we would like to point out some issues that may restrict the generalizability of our findings.

First, like a number of previous studies, our ASR evaluation was conducted with a low vocabulary connected digit task. For more complex recognition tasks, such as large vocabulary ASR, the HMM model space is much denser, requiring a higher accuracy in terms of acoustic modeling. Especially for the highly correlated FBE feature stream, we expect the use of diagonal covariance GMMs to become increasingly problematic. While mixture models with diagonal covariances can handle feature correlations to some extent, past research has shown that missing data marginalization schemes perform better with full covariance structures [Cooke et al., 1997; Kühne et al., 2008a]. Unfortunately, the processing costs involved with full covariances are significantly higher, making strategies to reduce resource requirements and speed up computations essential [Cooke et al., 1997; Pullella et al., 2009].

Second, we acknowledge that the data in this study was artificially mixed using a room image model for simulating sound reflections. Although quite a challenging setup
the simple ‘shoe box’ model does not fully represent real world conditions. Our simulations were limited to small-room environments with low to moderate levels of reverberation. For stronger levels of reverberation, it may be worthwhile to incorporate a spectral normalization scheme tailored to the missing data framework [Palomäki et al., 2004a].

Several other research directions may also be pursued to further improve the robustness of the presented system.

Future work needs to focus on the development of a more theoretically consistent approach for the rather ad-hoc PDF parameter estimation technique presented in Section 3.2. Possible extensions in this regard could include statistical cocktail party processing [Nix, 2005] as well as tailor-made feature extraction strategies [Kolossa et al., 2005] for propagation of the uncertainty information from the BSS front-end to the ASR back-end.

Another point of interest could be the search for tighter integration limits in bounded marginalization techniques in order to fully exploit the true potential of our proposed evidence model. A first successful attempt in this direction has been reported in Srinivasan et al. [2007].

The findings of this study also warrant further investigation to identify optimal strategies for spectral feature combinations. Of particular interest would be studies that investigate which type of evidence model and frequency filtering technique is best suited for combination in a missing data based HMM speech recognition system.

Lastly, another promising avenue for future research is to consider the use of top-down processing. Such information could not only assist the decoder in detecting inconsistencies between learned expectations and incoming bottom-up “evidence”, but could also help in the automatic identification of the target source prior to recognition. Like in Palomäki et al. [2004b], the recognizer was informed here which was the desired source for recognition. Multi-source decoding [Barker et al., 2005] or the integration of an attention model [Wrigley, 2002] are possible extensions to make the system more autonomous.

As a final remark, we point out that there remains a considerable performance gap between a priori and a posteriori evidence models. Closing this gap and moving towards more realistic conditions represent challenging yet exciting topics for future research.

Acknowledgment

The authors would like to thank Dr. Eric Lehmann for providing the MATLAB™ code for the room image model.
References


D. Kolossa, H. Sawada, R. Astudillo, R. Orglmeister, and S. Makino. Recognition of con-


S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, Odell. J., D. Ol-