Informing Multisource Decoding in Robust Automatic Speech Recognition

BY

Ning Ma

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY AT
THE UNIVERSITY OF SHEFFIELD

DEPARTMENT OF COMPUTER SCIENCE

JULY 2008

Informing Multisource Decoding in Robust Automatic Speech Recognition

Ning Ma

Abstract

Listeners are remarkably adept at recognising speech in natural multisource environments, while most Automatic Speech Recognition (ASR) technology fails in these conditions. It has been proposed that this human ability is governed by Auditory Scene Analysis (ASA) processes, in which a sound mixture is segregated into perceptual packages, called 'streams', by a combination of bottom-up and top-down processing.

This thesis examines a novel ASR framework based on the ASA account, Speech Fragment Decoding (SFD). A 'fragment' is a spectro-temporal region where energy from a single sound source dominates. SFD employs techniques developed from knowledge about the auditory system to identify fragments. A decoding process using statistical speech models is applied to the fragment representation to simultaneously identify speech evidence and recognise speech.

In this study three techniques for improving SFD are investigated. Firstly, explicit duration modelling is exploited to combat the corruption of acoustic data which often causes the decoder to produce word matches with unrealistic durations. Secondly, it is argued that the top-down information in recognition models may be insufficient to mediate the speech identification. Knowledge that can assist the decoder in the choice of speech evidence is investigated. Thirdly, pitch cues derived from structure in the correlogram are used in the fragment generation process.

A range of small-vocabulary speech recognition experiments are conducted for evaluation. The improved SFD system is able to produce word error rates significantly lower than conventional ASR, and is relatively insensitive to a range of noise conditions. In conclusion, the framework provides some progress towards finding a general solution to the robust ASR problem in multisource environments.

Acknowledgements

First special thanks are undoubtedly due to my supervisor, Phil Green, for his continuous support and encouragement throughout my PhD. He has been a main source of my inspiration for this work and a reliable source of my knowledge for beer. I would also like to thank Jon Barker for his tremendous help and advice over the years. His SFD implementation has laid the foundation for many experiments reported in this thesis.

Thanks are also due to my colleagues and friends in the Department of Computer Science at the University of Sheffield for their advice and friendship. I have enjoyed productive discussions with Guy Brown and Martin Cooke, who have provided much inspiration for the work in auditory scene analysis. And it has been a pleasure working with André Coy, Sue Harding, Heidi Christensen, Vincent Wan and Roger Moore, from whom I have learnt a lot. Special thanks to Yasser Hifny and Thomas Poulsen for the typical conversations which have accompanied me over many evenings in the lab. Thanks also to all those past and present members of the Speech and Hearing Research Group for providing a great atmosphere to work in.

I would like to thank Jeff Bilmes for hosting me in the SSLI lab at University of Washington, Seattle in 2007. This visit was funded by the Worldwide Universities Network. I have had the pleasure of working with Jeff, Katrin Kirchhoff, Chris Bartels and Hui Lin on duration modelling using graphical models. Thanks to Lee Damon for his technical support and to Mei Yang, Peng Gang for their lovely spicy dishes.

I would also like to thank Kalle Palomäki from the CIS lab at the Helsinki University of Technology for valuable discussions on modulation filter design.

Thanks to Matthew, Jianbiao, Ruhai, Sue, Thomas, Vicky, Fabien, Vinny and Yan-Chen for their friendship and providing welcome distractions from my PhD. Thanks to Vinny and Sue for proof-reading many chapters.

Thanks to my parents for their support, and to my uncle Xuesheng for his encouragement and friendship.

Finally, thanks to my wife Chunjie, for her continuous support and love.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Speech Recognition in Complex Environments

Imagine sitting in a busy restaurant amongst friends. What can you hear? Perhaps a familiar piece of music playing quietly in the background, the babble of distant voices broken occasionally by unexpected laughter, continuous clinking and clattering of dining-ware ... And yet with all these sounds reaching your ears at once, you do not experience any problem conversing with your friends. In fact, the conversation can be so pleasant that you do not even notice any of the other sounds.

Indeed, in daily listening environments speech is naturally mixed with various other sounds, but human listeners are remarkably adept at recognising speech in such complex environments. This experience is so common that the perceptual ability of listeners is often taken for granted. In the 1950s, Colin Cherry, a British engineer, described this phenomenon as the 'cocktail party problem.'

> *How do we recognise what one person is saying when others are speaking at the same time (the "cocktail party problem")? On what logical basis could one design a machine ("filter") for carrying out such an operation?* – Cherry [28]

So, how could one design a machine that can recognise speech in multisource environments with a performance that matches the robustness of human speech recognition (HSR)? This question asked by Cherry has been the ultimate goal of many scientists and engineers. For many decades, research on automatic speech recognition (ASR) has progressed dramatically

– from recognising isolated words with a limited vocabulary to large vocabulary continuous speech recognition tasks, and from recognising artificially prepared speech to increasingly spontaneous conversations recorded without human intervention. And yet today we are still unable to build a device even for a seemingly simple task (e.g. connected digits recognition) that can work as well as listeners in real (i.e. noisy) acoustic environments. As we will see in Section 2.2, most ASR systems are typically designed to work well in narrowly specified and highly predictable noise conditions. They would generally break down if the operating environment is not carefully controlled [119, 54]. For example, a computer dictation program which performs well in quiet office environments would most likely fail if background music were present.

This characteristic of current 'robust' ASR systems is largely due to, arguably, a historical reason. In the early stage of speech technology research it was a sufficient challenge to just handle the great variability of the speech signal itself. Most research therefore focused on the recognition of 'clean' speech, i.e. an unrealistic listening situation. This is in sharp contrast to studies in computer vision [127], where great attention was focused on scene analysis from the beginning.

The remarkable robustness of human speech recognition has inspired researchers to look for solutions to the robust ASR problem in the underlying processes of human speech perception, just like many successful applications of biological methods and systems found in nature to the study and design of engineering machines. Researchers believe that better understanding of how listeners recognise speech in noise may lead to better strategies for building machines to solve the same problem [32, 176, 86]. This is the question with which the thesis is concerned.

1.2 Auditory Scene Analysis

Understanding the principles underlying the perception of complex acoustic mixtures in humans is a challenging problem. The human auditory system is hugely complicated and involves complex interactions with the brain and other sensory systems. Although many underlying principles still remain unclear, decades of psychoacoustic research has brought us some promising insight into the mysterious process. It is generally believed that listeners are able to segregate individual sound sources from a complex mixture of sounds arriving

at our ears into perceptual packages, allowing whatever package is of interest at the time to be selectively attended to. Bregman [21] described this process as *auditory scene analysis* (ASA).

In one of the first attempts to seek solutions to the ASA problem, Cherry [28] conducted perception experiments in which subjects were asked to listen to two different messages simultaneously mixed on tape and try to repeat one of them word by word. His work revealed that the ability of listeners to separate sound sources is often based on the characteristics of the sounds, such as the gender of the speaker, voices, speaking speed, and the direction from which the sound is coming. Bregman [21] summarised evidence from a large number of perceptual experiments in his well-known book, "Auditory Scene Analysis", which has suggested that listeners solve the ASA problem by interactively exploiting primitive *data-driven* grouping principles, which are innate constraints driven by various properties of the acoustic input, as well as *schema-driven* constraints, which employ prior knowledge of familiar patterns that have been learnt from acoustic environments.

1.2.1 Data-Driven Processing

Primitive data-driven grouping principles describe how elements extracted from the input sound mixture may be grouped across time/frequency (T/F) according to characteristics. For example, T/F components will show a tendency to group together if they share a common fundamental frequency [22, 167, 5], if they have synchronised changes in frequency or amplitude [21, p.250], or if their energy comes from the same direction [44, 47].

Among many, grouping by harmonic relation appears to be the most popular cue in ASA models. Pitch, which represents the perceived fundamental frequency ($F0$) of sound, is the most studied auditory attribute of the quasi-periodic speech signal. Perceptual experiments have shown that pitch plays a significant role in separating simultaneous voices. For example, in the well studied 'double vowel' experiments [167, 5, 43], researchers demonstrated that a very small difference in $F0$ can significantly improve listeners' performance in identifying two simultaneous vowels. There is also evidence that listeners use the harmonicity cue to help them in sequential grouping of sound components. For example, van Noorden [180] described how two rapidly alternating tones may group sequentially based on $F0$ continuity. When

close in frequency the two tones were perceived as one coherent stream with a 'galloping rhythm'. With a larger frequency difference between the two tones the coherence was lost and instead two separate perceptual streams were heard.

Experiments have suggested that the harmonicity-based cue is one of the most robust grouping cues. For example, Darwin and Hukin [49] reported that pitch cue is more resistant to reverberation than spatial cues. Therefore grouping by harmonic relation has been employed in most successful computational models of auditory scene analysis [e.g. 30, 23, 186].

There is plenty evidence that listeners employ more than one cue to help them achieve robust perceptual source separation. For example, listeners have no problems in separating unvoiced sound sources which do not have harmonic structures. For simultaneous vowels with the same $F0$, although with more difficulties than if $F0$ is different, listeners are still able to perform separation with an accuracy significantly greater than the chance level [167, 4] [1]. It is likely that various grouping cues interact in forming an overall segregation [46].

1.2.2 Schema-Driven Processing

Schema-driven processes, by contrast, work 'top-down' by actively finding support for learnt models of commonly occurring sounds in the mixtures – an active 'hearing-out' for a given pattern. The effect of schema-driven processing can be very strong as listeners actively try to associate sound scenes with the patterns they have learnt. With different knowledge of patterns perception can often be different.

A compelling example of such schema-driven processing in audition is the perception of a 'click' sound employed by some African languages in which a change in the position of the clicks within a word can change its meaning. Listeners who are unfamiliar with these languages would hear two separate perceptual streams – with one stream sounding like someone speaking a foreign language, and the other sounding like a rapid sequence of click sound. For those who do not have experience of these languages, the click fits better to their knowledge as a non-speech sound. Native speakers of these languages, however, will hear the clicks and the rest speech sounds as an integrated whole – the clicks sound coherent with other

[1] It should be noted that the schema-driven processing should also have effect in these experiments. See Section 1.2.3 for discussion on interaction between primitive and schema-driven processes.

speech sounds as well as more familiar plosives sound to native English speakers. A detailed discussion of this example can be found in [21, p.686], which also provides more supporting evidence of the schema-based process.

Another example of this schema-based process in everyday environments is that when you hear someone speaking English with a strong accent which you are not familiar with, you often have difficulty in understanding the speaker even the speaker uses perfect grammar. I had personal experience on this example – a friend of mine (a native English speaker) could not initially understand many English words I spoke, but after some time this was less of a problem for her. "It's not because your English has improved," she later told me, which is largely true, "I just know what to listen for." Although she did say this partly because she wanted to disappoint me, this is a common experience for many people when presented an unfamiliar sound. Listeners learn to listen to details of a sound, e.g. a good musician can be trained to identify the instruments in a music mixture even they have the same fundamental frequency.

The schema-driven processing also commonly occurs in vision. For example, the UK clothing company 'French Connection' fully exploited the top-down process to attract consumers' attention with its controversial brand name 'fcuk' (see Fig. 1.1) – an acronym for 'French Connection United Kingdom'.

Figure 1.1: *An example of exploiting schema-driven processing.*

1.2.3 Interaction Between the Two Processes

The boundary between data-driven and schema-driven processes is not clear, and there is often debate over the extent to which auditory organisation is based by the primitive processes [159], or whether it is driven primarily by learnt patterns of sounds [63]. Remez et al. [159] criticised ASA on the grounds that it fails to account for the coherence of speech. They used an artificial speech stimulus, known as sine-wave speech (SWS), to argue that the speech stimulus, despite violating the ASA grouping principles discussed by Bregman, is still perceived as a coherent whole. However, Barker [9] carried out a series of experiments

using SWS and demonstrated that although very few, there are still primitive grouping cues available for SWS (e.g. common onsets). The fact that SWS lacks many auditory grouping cues makes its recognition much harder than normal speech. Barker further pointed out that the argument by Remez et al. [159] was based on an extreme view in which auditory scenes analysis is entirely dependent on primitive processes with no interactive link to knowledge about the sound sources expected to be present.

Other people may take an extreme view that auditory scene analysis can be accomplished entirely by schema-driven processes. For example, Ellis [63] argued that primitive grouping principles can be considered as implicit models of sound sources that have been learnt from acoustic environments. Although such a strong view is valid to some extent, it claims that listeners cannot segregate sources unless they have some kind of prior model for at least one of the sources. Primitive grouping rules are generally difficult and inefficient to be represented using any types of models. They are developed in an early stage of the auditory system [128] and can apply to general sounds. By contrast, schema-driven processing is developed to learn familiar patterns from particular sounds. Without some primitive mechanisms to group the mixed sounds into some initial coherent structures in the first place, there would be nothing for schema-driven processes to work on. As Bregman wrote:

> *It is important to emphasize again that the way that sensory inputs are grouped by our nervous systems determines the patterns that we perceive.* – Bregman [21]

In his book Bregman showed strong evidence that the two processes work interactively together to organise sounds. Primitive grouping alone is not able to segregate sound sources reliably, especially when dealing with highly dynamic sources such as speech. Instead, it appears that the data-driven processing would suggest some initial groupings to form *local* coherent spectro-temporal regions. Schema-driven processing is necessary in order to merge local regions together into common streams based on a logical explanation of the present auditory scenes.

1.3 From Scene Analysis to Speech Recognition

Given the current understanding of auditory scene analysis, there is another tremendously important and practical question: Are there good ways of exploiting knowledge to *engineer* machines that perform ASR in the presence of multiple sound sources? The problems of auditory scene analysis and automatic speech recognition were traditionally addressed separately, often by different research communities. This causes great difficulties in integration of scene analysis and speech recognition systems – primarily their incompatible representations. For example, ASR assumes that the speech input is a sequence of acoustic feature vectors which deliberately excludes detail with great variability such as voicing periodicity. By contrast, periodicity is the most popular cue in ASA systems. ASR systems typically employ decorrelated feature representations (e.g. cepstral features) which can be compactly modelled, while ASA systems work on a spectro-temporal representation of the acoustic input. Despite these difficulties, growing research has focused on building ASR systems that can benefit from the understanding of auditory scene analysis, mainly through the development of *computational auditory scene analysis* (CASA) [187].

1.3.1 Computational Auditory Scene Analysis

Motivated by extensive research on ASA, the field of CASA has evolved with increasing interest, which aims to develop computer programs to perform sound source separation based on perceptual principles. Different from many other sound separation techniques, such as the blind source separation technique [107], CASA is perceptually motivated and provides a general perspective in which speech is regarded as just one of many sound sources in a complex acoustic environment. Therefore, ASA principles can apply equally well to all sound sources. This offers an alternative approach which directly addresses the problems raised by complex acoustic scenes, just like its counterpart in the field of computer vision.

CASA systems usually employ a spectro-temporal representation of sound signals derived from computer models of the peripheral auditory system [30]. Rather than detailed physiological models of auditory mechanics, the computer peripheral models were largely motivated by the known psychoacoustical properties of the human auditory system and provide a frequency analysis which is consistent with the properties of cochlea frequency selectivity. An

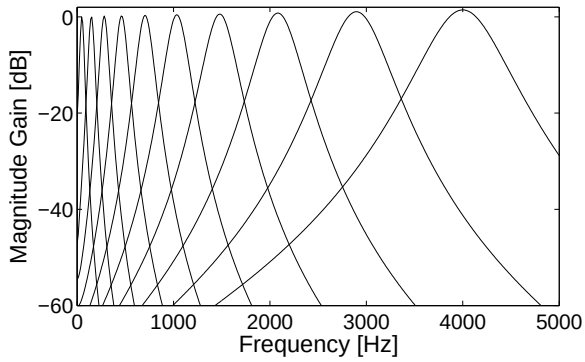

Figure 1.2: *Frequency responses of a gammatone filterbank with ten filters whose centre frequencies are equally spaced between 50 Hz and 4 kHz on the ERB-rate scale.*

auditory filterbank is commonly used to simulate the motion of the basilar membrane within the cochlea as a function of time, in which the output of each filter models the frequency response of the basilar membrane at a single place.

The gammatone filter [101] is widely used in models of the auditory system. Gammatone filter modelling was a physiologically motivated strategy to mimic the structure of peripheral auditory processing stage [30]. The gains of the filters were chosen to reflect the transfer function of the outer and middle ears. More details about the gammatone filter can be found in Appendix A.

A gammatone filterbank is normally defined in such a way that the filter centre frequencies are distributed across frequency in proportion to their bandwidth, commonly known as the equivalent rectangular bandwidth (ERB) scale [75]. The ERB scale is an approximately logarithmic function on which the filter centre frequencies are equally spaced. Fig. 1.2 shows frequency responses of a gammatone filterbank with ten filters whose centre frequencies are equally spaced between 50 Hz and 4 kHz on the ERB scale.

Output of the gammatone filterbank can be further processed with some form of non-linear rectification to derive a spectro-temporal representation: the 'cochleagram'. The cochleagram is commonly employed as a front-end for CASA studies, e.g. for pitch analysis (see Chapter 6). For visualisation the cochleagram contains too detailed information and therefore is normally simplified by smoothing over time, down-sampling and compressing to produce a

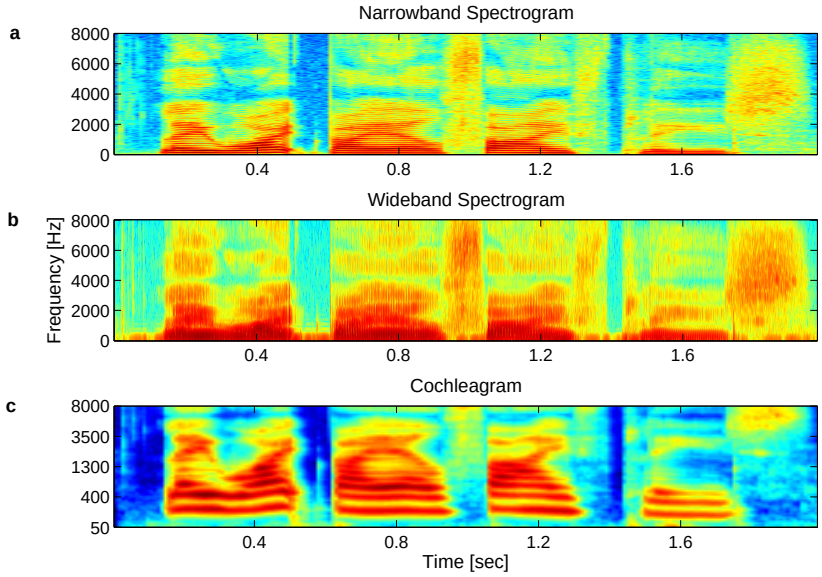

Figure 1.3: *Comparison between (**a**) narrowband spectrogram, (**b**) wideband spectrogram, and (**c**) cochleagram computed for the same utterance 'lay white by L 5 please' spoken by an English female speaker.*

spectrogram-like representation, or an auditory spectrogram [2]. Fig. 1.3 shows a narrowband spectrogram and a cochleagram computed for the same utterance 'lay white by L 5 please' spoken by an English female speaker. The spectrogram was produced using 512 frequency points and a 20 ms Hamming window with 10 ms overlap. The log-compressed cochleagram was produced using a 64-channel gammatone filterbank whose output was smoothed with an 8 ms window and sampled at 10 ms intervals. The cochleagram has a few advantages over conventional spectrograms. First, with the ERB-rate spacing it has better spectral resolution at the low frequency end, which results in individual harmonics of sounds being resolved. Resolved harmonics allow sources to be tracked and detected at unfavourable signal-to-noise ratios (SNRs) as source energy is heavily concentrated at its harmonics. Second, the shorter smoothing window offers better temporal resolution at the high frequency end which causes onsets of acoustic events to be emphasised in the cochleagram.

Using the cochleagram many researchers have proposed automatic sound separation systems based on the known principles of human hearing and achieved some success [e.g. 23, 186, 64]. A good review of CASA development can be found in the recently published CASA

[2] As the cochleagram approximates firing rates of the auditory nerve the auditory spectrogram is also referred to as the 'ratemap' representation [23].

book edited by Wang and Brown [187]. Some successful CASA systems will be discussed in Section 2.4.4.

1.3.2 CASA-Driven Approaches to ASR

Although CASA seems to be a natural solution to the robust ASR problem in achieving human performance, it is not obvious how CASA and ASR can be effectively combined. Weintraub [190] was the first to systematically study this problem. In his work, ASA principles were applied in an attempt to separate monaural voice mixtures of two speakers by finding periodicities in each frequency channel based on an autocorrelation-like neural coincidence function [117]. Each speech signal can be in one of four states (voiced, unvoiced, silent or transitional) and a Markov model was used to find the best state sequence using a dual-pitch tracking algorithm based on dynamic-programming. The pitch estimates were then used to recover the spectra of each voice. Weintraub assessed the performance of his speech separation system using the conventional metrics of recognition accuracy by feeding the separated and reconstructed speech into a speech recogniser. Although the results were disappointing, which is not surprising given the development of ASR at that time, Weintraub himself was quite clear about the limitation of his model and suggested that any complete model of auditory organisation would necessarily involve more than just data-driven processing.

Obviously, CASA can be employed as a 'front-end' to produce effectively 'noise-free' speech before passing the output to an ASR 'back-end'. However, this 'separation–resynthesis–recognition' strategy is problematic itself. Although enhanced speech signals may sound more intelligible for human listeners, they are not necessarily suitable for ASR. The artifacts introduced during the enhancement process often cause a dramatic problem for machines, which humans may find a trivial distraction. Partly because of the lack of an advanced statistical framework, after Weintraub's initial efforts researchers often used other metrics to evaluate CASA systems[3]. For example, Cooke [30] examined the similarity between the segregated target and the pre-mixed target signal in the spectro-temporal plane. Brown and Cooke [23] and Wang and Brown [186] measured performance in terms of improvements in signal-to-noise

[3]Another important reason is that ASR is not the only application for CASA and there are many ohter major applications, e.g. hearing aids.

ratio. Evaluation based on subject listening experiments were also reported [177].

In the 1990's, researchers at Sheffield University developed a statistical framework that allows ASA models to be linked with ASR without the need to resynthesise target speech signals – the 'missing-data' approach to robust ASR [33, 81, 119, 35]. Missing-data ASR assumes that when speech is corrupted by noise, some spectro-temporal regions will remain reliable (i.e. the observed energy is close to actual speech energy) and can be identified for automatic speech recognition. Cooke et al. [34] demonstrated that robust speech recognition can be achieved based on only a small portion (10%) of speech evidence. Missing-data ASR was systematically discussed in [35] and will be examined in detail in Section 2.4.

Although missing-data ASR provides a statistical approach to linking the output of CASA with ASR, the problems of segregation and recognition are decoupled. The convenience of this strategy is that the CASA front-end and the ASR back-end can be developed independently. However, as we have seen in Section 1.2, sound organisation requires both the data-driven and schema-driven processes to be interacted when interpreting complex acoustic scenes. Nearly all existing CASA studies have focused on data-driven processes. We will see in Section 2.4.5 that the decoupling of segregation and recognition ultimately limits such a simple 'left-to-right' strategy and more sophisticated solutions are needed.

Recently, a technique termed 'multisource decoding' [10, 15] emerged as a promising framework for CASA-based ASR which follows the auditory scene analysis account of sound organisation. The technique combines segregation and recognition in a tightly coupled process. Primitive grouping techniques are applied to segment the spectro-temporal plane of the acoustic mixture into local regions where energy from a single sound source dominates. These spectro-temporal regions are called 'fragments' in this study. The identities of fragments do not have to be decided at this stage. Statistical schema-driven processes then employ recognition models to simultaneously search for the most likely word sequence and foreground/background segregation. The multisource decoding technique can serve as a potential solution to the problem in linking ASA with ASR. When the target sound source is constrained to be speech, Barker et al. [15] termed the technique 'speech fragment decoding' (SFD) (see Chapter 3). This doctoral work is based on the SFD framework.

1.4 Informing Multisource Decoding

The strength of SFD is that it is designed to operate without strong assumptions about the nature of the interfering noise, in contrast with many conventional approaches to robust ASR. However, there are several additional factors to be considered when multiple sound sources are present:

- SFD bases speech recognition on partial acoustic evidence. The corruption of acoustic features and the weak duration constraints implicitly modelled in HMMs often lead to word matches with unrealistic durations by ASR in noisy conditions. Stronger duration constraints may need to be introduced into the speech decoding process to combat the corruption.
- SFD assumes that each fragment is part of either the speech foreground or the noise background with equal probability. Although accurate noise models are difficult to estimate, knowledge about the noise is often available which can distinguish speech fragments from noise fragments. This information can be exploited to assist the decoder in the choice of fragments.
- The quality of the fragments generated can also affect the recognition performance of SFD. If the fragments are not *coherent*, i.e. contain too much energy that belongs to different sources, the recognition accuracy will be expected to be low.

This thesis will examine the SFD framework and investigate three techniques for improving speech decoding in multisource environments. Firstly, explicit duration modelling is exploited to combat the corruption of acoustic data which often causes the decoder to produce word matches with unrealistic durations. Secondly, it is argued that the top-down information in recognition models may be insufficient to mediate the speech identification. Knowledge that can assist the decoder in the choice of speech evidence is investigated. Thirdly, pitch cues derived from structure in the correlogram are used in the fragment generation process.

This study focuses on single-channel signals. Although there is strong evidence that listeners exploit various (and multi-modal) cues to recognise speech in a noisy environment, e.g. inter-aural information available from the two ears, or lip-reading when speech is barely audible in a noisy bar, they are still able to effectively extract target audio streams from monaural

acoustic mixtures with little effort, for example, listening to speech/music mixtures on a mono radio program. However, separating and recognising speech in single-channel signals, the problem considered in this thesis, still remains a challenging problem for machines.

1.5 Thesis Organisation

In this chapter we have reviewed the fundamentals of auditory scene analysis and its application to ASR. Chapter 2 will briefly explain the standard statistical ASR framework and examine some traditional approaches that have been widely applied to deal with noisy speech data. It will proceed to examine in detail some developments of CASA-inspired approaches to ASR. The traditional ASR approaches will be contrasted with CASA-based approaches. In particular, missing-data ASR [35] will be examined. The CASA-based approaches that can be employed to identify the reliable speech evidence for missing-data ASR will be reviewed at the end of the chapter.

Having examined the ASR approaches which effectively decouple source segregation and recognition, Chapter 3 will begin by arguing that this strategy is ultimately limited and more sophisticated approaches are needed. The speech fragment decoding technique – the framework this doctoral work is based on – is then presented as a potential solution to the problem of CASA and ASR combination. Finally, Chapter 3 will show that there are several possibilities to improve speech recognition for the SFD model in multisource environments, which will be investigated in following chapters.

It is well known that hidden Markov models (HMMs) do not directly characterise some important temporal information such as duration constraints [153]. The weak duration constraints may cause speech decoders to produce word matches with unrealistic durations. This is of particular importance for missing-data ASR and SFD which uses missing-data techniques at its core as speech recognition is based on only partial acoustic evidence. Chapter 4 will investigate the effect of explicit duration modelling in the context of missing-data ASR. Duration constraints are modelled at both state-level in Section 4.2 and word-level in Section 4.3. In particular, the 'prepausal lengthening effect' [41] – the property that before a speech pause, the preceding speech unit (particularly vowels) tends to lengthen – is investigated.

Chapter 5 will demonstrate that the top-down information provided by speech models is often insufficient to recruit enough speech fragments, which will lead to poor recognition performance. To assist the decoder in the choice of fragments, Chapter 5 will introduce a 'speechiness' measure for fragments – a degree of confidence that the fragment is part of the speech foreground. A technique based on the modulation spectrogram [82] is proposed as a speechiness measure, which emphasises the characteristic low-frequency modulation energy of speech.

Chapter 6 is concerned with the use of primitive CASA models to address the problem of fragment generation for SFD. The quality of fragments can affect the recognition performance of SFD. Chapter 6 will commence by reviewing correlogram-based CASA models. Pitch cues derived from a particular tree-like correlogram structure are employed in fragment generation. The coherence of the fragments is compared with that of fragments generated using a technique based on the summary correlogram. Chapter 6 will also present evaluation using ASR experiments on a simultaneous speech recognition task.

Chapter 7 summarises the thesis and presents future development.

Chapter 2
Robust Automatic Speech Recognition

This study is concerned with single-channel approaches. Before examining strategies for coupling CASA and ASR, this chapter will first review the standard statistical ASR framework and some traditional approaches that have been widely employed to deal with noise. After all, our goal is to *engineer* a machine that can perform robust speech recognition and it is important to understand the limitations of many existing robust ASR techniques. Readers are referred to [154, 100, 92, 72] for a detailed account of the statistical ASR framework. Comprehensive reviews of robust ASR can be found in [104, 77, 72]. This chapter will also review, in detail, some of the earlier attempts that have been made to tackle the problem of robust ASR using perceptually inspired approaches. These approaches will be contrasted with traditional ASR approaches at the end of the chapter where their limitations will also be discussed.

2.1 The Statistical ASR Framework

Automatic speech recognition (ASR) is the process of automatically converting spoken words to machine-readable input. Typically the audio waveform is first converted into a temporal sequence of acoustic feature vectors as the acoustic input to a speech recogniser. Popular choices of ASR features are those represented in an orthogonal domain, such as mel-frequency cepstral coefficients (MFCC), which are decorrelated and can be modelled more efficiently. The goal of the recogniser in continuous speech recognition is to find the most probable sequence of words W given the acoustic input X, an acoustic model and a language model.

We can treat the acoustic input X as a sequence of individual observations:

$$X = x_1, x_2, x_3, \ldots, x_T \tag{2.1}$$

Similarly, we can treat the sequence of words as:

$$W = w_1, w_2, w_3, \ldots, w_L \tag{2.2}$$

When formulated in a statistical manner, the goal is then to find the word sequence $\hat{W}$ that has the maximum *a posteriori* (MAP) probability given the sequence of acoustic observations X:

$$\hat{W} = \arg\max_{W} P(W|X) \tag{2.3}$$

Eq. 2.3 is guaranteed to give us the optimal sentence $\hat{W}$. However, it is much simpler and more practical to calculate the likelihood, $P(X|W)$, and this can be interpreted as the probability that a sequence of feature vectors generated by a particular word sequence. We can use Bayes' rule to break Eq. 2.3 down as follows:

$$\hat{W} = \arg\max_{W} \frac{P(X|W)P(W)}{P(X)} \tag{2.4}$$

Since the probability of the acoustic observation sequence, $P(X)$, remains constant for each word sequence, Eq. 2.4 becomes:

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \tag{2.5}$$

This formulation was developed by Baker [8], Jelinek [99], Bahl et al. [6]. The first term, $P(X|W)$, the *observation likelihood*, is termed the *acoustic model*. $P(W)$, the *prior probability*, is termed the *language model*. For the sake of simplicity it is assumed that the parameters in the likelihood and priors are independent and can be estimated separately.

In most successful ASR systems the acoustic model is represented using hidden Markov models (HMMs). HMMs are a general stochastic representation that can be applied to various problems. Their applications have been developed in the context of a long history of pattern recognition technology. Although specific methods are changing, the pattern recognition perspective continues to be useful for the description of many problems and their proposed solutions. An HMM is defined with a sequence of states $Q = q_1, q_2, \ldots, q_N$, a set of transition probabilities a_{ij} representing the probability of moving from state q_i to state q_j, and a set of

output observation distributions $b_j(x_t)$ expressing the probability of an observation x_t being generated from a state q_j. Two conditional independence assumptions are made for HMMs:

- states q_j are conditionally independent of all other states given the previous state q_{j-1} (i.e. the first-order Markov assumption);
- observations x_t are conditionally independent of all other observations given the state q_t that generates x_t

The topology of an HMM is defined by state transitions. There are three commonly used topologies of HMM: the ergodic topology where each state is connected to each other, the left-to-right no-skip topology where each state must be traversed from left to right as time passes, and the skipped state topology in which some states may be skipped. In speech recognition the left-to-right no-skip topology (see Fig. 2.1) is mostly employed, because it is useful for modelling signals with characteristics that change with time, such as speech signals. Each state has a probability density function (*p.d.f.*) for the feature vectors that is used to determine the probability that a particular feature vector could be generated by the state. The fact that the exact state sequence that determined the output is unobserved makes the model a *hidden* Markov model.

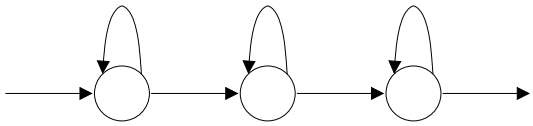

Figure 2.1: *Outline of a left-to-right no-skip HMM.*

HMMs can have either discrete or continuous probability distributions. Discrete density HMMs use vector quantisation to assign probabilities to a discrete set of code words (or symbols) representing acoustic data being generated. In continuous density HMMs (CDHMMs) each state is associated with a *p.d.f.* that models the distribution of the acoustic data. Before the HMMs can be used in recognition they have to be trained, most commonly using the Baum-Welch algorithm [18], which is a generalised example of the expectation-maximization (EM) algorithm [53]. The Baum-Welch algorithm gives a local optimum which is known to produce an estimate of the parameters that is *more likely* than the initial estimate [153]. The re-estimation is terminated when the increase in likelihood falls beneath some pre-determined

threshold.

After the estimation phase, the HMMs can be used for recognition. In connected word recognition, the Viterbi algorithm [185] is widely used to find the most probable path through a probabilistically scored time/state lattice, given the observation features [140]. This approach is time-synchronous, and computes the most likely path to every state in every model at time t, given the corresponding paths at time $t-1$. An exhaustive search over all state sequences is effectively performed without the need to calculate all possible paths. The Viterbi algorithm is a special case of dynamic programming. It makes use of the fact that the probability of generating the first t observations and being in state i at time t, depends only on the state occupied at time $t-1$, i.e. it exploits the first order Markov property on which the model is based. In this way the search effectively finds the best path through the utterance without evaluating all possible paths independently. A detailed treatment of HMMs and their operations can be found in [153, 72].

2.2 Conventional Approaches to Robust ASR

HMM based automatic speech recognition has achieved great success in controlled environments. The scale and complexity of viable speech recognition tasks has significantly increased in recent years. However, while most ASR systems produce acceptable recognition accuracy for speech collected in quiet situations, their performance degrades dramatically in noisy environments [77]. Traditional approaches to achieving noise robustness exploit the differences that are assumed to exist between the training and operating environments and their goal is to minimise the mismatch, typically using engineering methods. Many well established techniques have been widely applied with some success, either alone or in combination. These techniques can be split roughly into three categories by their initial objectives: exploitation of noise-robust features, feature compensation and model compensation.

2.2.1 Exploitation of Noise-Robust Features

Acoustic features that are inherently less sensitive to noise can be employed to improve noise robustness. An example of this approach is the use of a speech feature representation known as RASTA, an acronym for 'RelAtive Spectral TrAnsform' [88, 87]. Conventional acoustic

features typically represent the short-term speech spectrum, which is vulnerable to spectral distortions. Human auditory perception tends to operate with long constants, especially in adverse environments [87]. The temporal properties of speech are often quite different from those of environmental effects. Therefore, the RASTA processing technique can be employed to filter out noises with modulation frequencies outside the narrow frequency range (4 to 50 Hz) that characterises speech [82]. RASTA applies band-pass filtering of time trajectories to the energy of each spectral component in order to smooth over short-term noise variations and to provide a cancellation of slowly varying additive noise resulting from static spectral correlation in the speech channel, e.g. from a telephone line. The technique is generalised in JRASTA to accommodate non-stationary convolutive noise. RASTA features are often combined with other signal processing algorithms such as perceptual linear predictive (PLP) analysis [85], which was originally proposed by Hermansky as a way of warping spectra to minimise the differences between speakers while preserving the important speech information. The combination has become a popular speech feature representation known as RASTA-PLP [88].

Speech recognition experiments [87] employing RASTA-PLP showed better noise robustness over standard PLP or MFCC parameters. Although these techniques can be very effective in some situations, they are clearly limited by the dependence on an easily characterised difference between the target speech and the interfering noise. If the noise source has a similar acoustic property to that of speech, e.g. a second speaker, then such techniques will be unlikely to bring robustness to an ASR system.

2.2.2 Feature Compensation

This category of methods attempt to pre-process the noisy speech in such a way that the resulting features better fit the models trained using clean speech. This scheme is often termed 'speech enhancement'. The simplest approach in this category is to 'clean-up' noisy speech using spectral subtraction [e.g. 118, 121, 111]. The approach assumes that the noise is statistically stationary. It makes use of the fact that power spectra of additive independent signals are also additive. Hence, additive noise signals can be effectively removed if they have a relatively stationary spectrum compared to that of the speech signal. As a result of the fluctuations of noise spectrum around its mean value, negative estimates of the

speech spectrum may occur. Usually some ad hoc flooring is required to make the estimates consistent. This non-linear operation puts residual noise in the output signal commonly referred to as 'musical noise' [77].

The popularity of spectral subtraction is largely due to its relative simplicity and efficiency. However, it would fail with non-stationary noises. A common approach to this problem is to estimate the noise spectrum in periods where speech is known to be absent [e.g. 12]. This approach, however, requires prior knowledge of speech segmentation, and is ineffective when the noise spectrum is difficult to estimate.

Cepstral mean normalisation (CMN) [120] is also a common technique employed to remove the global shift of the mean affecting the cepstral vectors. This normalization compensates for the main effect of channel distortion and some of the side effects of additive noise. However, the nonlinear effects of additive noise on cepstral features cannot be treated by CMN and this limits its effectiveness to only moderate levels of additive noise.

Recently, Cui and Alwan [42] proposed a feature compensation technique based on polynomial regression of utterance signal-to-noise ratio (SNR) for noise robust ASR. In this method the bias between clean and noisy speech features is approximated by a set of polynomials which are estimated from adaptation data for the new environment using an EM algorithm under the maximum likelihood criterion. During decoding the utterance SNR is first estimated and noisy speech features are then compensated for by corresponding regression polynomials.

de la Torre et al. [52] proposed a method to compensate for nonlinear distortions in acoustic features based on the histogram equalization (HEQ) technique commonly used in digital image processing [79]. This method assumes that the effect of the noise distortion is a monotonic transformation in the feature space. It provides a transformation mapping the histogram of each component of the feature vector onto a reference histogram.

The noise can also be successfully separated from speech using cues from multiple sensors, e.g. when a microphone array is present. If there are at least as many sensors present as sound sources, techniques such as blind source separation (BSS) [107] based on independent component analysis (ICA) can be employed to recover independent sources given only sensor observations, which are are treated linear mixtures of the source signals [146]. However, these techniques cannot be applied for single-channel mixtures.

It should be noted that many techniques in this category were originally developed to improve speech quality. Therefore they may increase speech intelligibility for human listeners, but not necessarily the performance of automatic speech recognisers.

2.2.3 Model Compensation

Instead of estimating the clean speech from the noisy observations, model-based techniques attempt to modify the speech models in order to account for the interfering noise. Such schemes are potentially able to deal with time varying noises, although normally require a statistical model of the corrupting noise. For example, HMM decomposition [182] involves creating a noise model that captures the variability expressed in the noise. This noise model is decoded in parallel with speech models which can jointly explain the noisy observations. In order to be mathematically feasible, it requires that the noise and speech be modelled in the log-spectral domain and treated as independent. Gales and Young [71] extended the idea in the parallel model combination (PMC) technique to perform model compensation in the cepstral domain. It combines the models for speech and noise to derive a 'noisy speech model' using a mismatch function that approximates the effect of the noise on speech. Therefore the model compensation can be done before recognition, unlike HMM decomposition in which model combination is performed during recognition time. PMC also has the advantage over HMM decomposition that it can be applied in the cepstral domain and therefore orthogonal features can be employed.

Both HMM decomposition and PMC techniques have been shown to produce very low word error rates when the noise can be adequately modelled. However, the requirement that detailed models are available for all the noise sources is often difficult to meet, especially in an unpredictable noisy environment. One solution would be to keep a library of many different noise models, if feasible, and combine the speech model with each of the noise models during recognition. This is, however, extremely computationally heavy due to the factorial nature of HMM combination. When more than a few models are involved the model combination can have an explosion in the size of the state space.

Although originally developed for speaker adaptation, maximum likelihood linear regression (MLLR) [112] is also an effective way to adapt the clean acoustic models to a different op-

erating environment [e.g. 70, 191]. MLLR obtains the environmentally matched models by rotating and shifting the means of the Gaussian mixtures in clean HMMs using linear regression. In comparison with PMC, MLLR is more computationally cheap but the performance of linear transformation is limited. Zhang and Furui [191] proposed piecewise-linear transformation (PLT) technique based on MLLR, in which various types of noise are clustered according to their spectral property. A noisy speech HMM set corresponding to each clustered noise and SNR condition is made and the best matching HMM set is selected and further adapted using MLLR.

If the noise conditions are known in advance, another simple strategy is to train recognition systems on a range of noisy examples of the speech – 'multicondition training' [150]. Multicondition training can bring more robust recognition performance against noise than training using clean speech provided that the operating noise condition is similar to that during training. However, without prior knowledge of the noise it is often difficult to design a proper training set of noisy speech. Training systems on a greater range of noisy examples might sound advantageous but will significantly decrease the discriminating ability of recognition models.

2.2.4 An Alternative Hypothesis for Robust ASR

The problem caused by the interfering noise is conventionally viewed as a mismatch between the conditions where a recogniser is trained and operates. Therefore most of the solutions attempt to minimise the mismatch, either by compensating features to match the pre-trained speech models or by adapting the speech models to accommodate the interfering noise. However, most of them are designed to operate in narrowly specified and highly predictable noise conditions [54]. The narrowness of conventional robust ASR solutions is such that systems designed for one environment cannot be expected to work in another (e.g. multicondition training). If ASR systems were to operate in the wide range of noisy environments with which listeners regularly cope, more general techniques have to be developed.

Common aspects of many distortions include the absence of spectro-temporal regions or the presence of additive noise, resulting in incomplete (both spectral and temporal) acoustic evidence. There is much evidence that listeners routinely handle the incomplete data sit-

uation [67, 2, 188]. In the last two decades researchers started to consider an alternative hypothesis – that sound being distorted by noise is a valid characterisation of the normal listening situation [86, 32]. The rest of this chapter will examine several approaches to robust ASR motivated by this hypothesis.

2.3 Multistream Speech Recognition

Fletcher [66] found that the error rate (represented as a fraction of one) for human phoneme perception using the full frequency range was approximately equal to the product of the error rate using high-pass filtered speech and the error rate using low-pass filtered speech at the same cut-off frequency [2]. Furthermore, the overall error rate is independent of the cut-off frequency used. Allen [2] interprets the work of Fletcher as suggesting that the acoustic information in the speech signal is decoded independently in narrow frequency sub-bands and the final decision is based on recombining the sub-band decisions. An alternative interpretation of Fletcher's work is that as long as any sub-band combination contains sufficient information to decode the linguistic message, the information from the remaining sub-bands (possibly corrupted) can be ignored [89]. This research inspired the use of multistream recognition methods to deal with band-limited noise corruption [20, 89, 83, 139]. If noise corrupts the data in some spectro-temporal regions, can results for uncorrupted regions be made to prevail? This is the motivation behind multistream recognition methods as an alternative to traditional 'full-band' based methods.

2.3.1 Full-band ASR vs. Multistream ASR

Traditional full-band ASR methods are sensitive to local feature corruption. Assuming features are independent, the joint observation likelihood is a product of local observation probabilities of each feature. As a result, the product is typically dominated by small probabilities. This characteristic provides the model with a better discrimination between speech units based on local feature differences, but also an increased sensitivity to local feature corruption.

In the multistream recognition approach, the whole speech spectrum is split into a number of narrow frequency sub-bands, with features extracted independently from each sub-band

being allocated a separate acoustic model. Recognition is done independently and the sub-band recognition results are then recombined to give a final recognition decision. When speech is corrupted by band-limited noise, a local sub-band corruption only degrades its corresponding sub-band features and hence will not spread over the entire feature space. Hermansky et al. [89] showed that the partial information from individual sub-bands can be successfully merged, producing recognition error rates very close to that of a full-band recogniser when tested on clean (i.e. uncorrupted) speech. The study by Hermansky et al. is essentially the proof of concept of multistream ASR. Studies [20, 89, 83, 135] have also shown that the sub-band based approach is able to improve ASR robustness against band limited distortions over a conventional full-band based method.

2.3.2 The Multistream Recombination Problem

For multistream speech recognition, a critical issue is to find a way of recombining sub-band information. Ideally, those bands that are less corrupted by noise should be selected and those unreliable bands should be suppressed (wholly or partially) during the recombining process. Identifying the reliable bands, however, is difficult without prior knowledge of the distortions. Many researchers have proposed various methods for sub-band recombination, such as linear combination [20, 89], neural networks [20, 89], the full-combination approach [19, 83, 139], or the probabilistic union model [134, 135].

Bourlard and Dupont [20] examined a weighting scheme in which the contribution of each sub-band was assigned a recombination weighting factor. Two different recombination functions were examined: i) a linear weighting function and ii) a non-linear multilayer perceptron (MLP). With the linear function the weighted sum of the log-likelihoods of all sub-band observations were calculated. The weighting factors can be estimated from normalised phoneme-level recognition rates in each sub-band, or local signal-to-noise ratios (SNRs) estimated using spectral subtraction. The weighting factors sum to 1. With the non-linear function, an MLP parametrised in terms of the weighting factors was trained to estimate posterior probabilities of each speech unit given the log-likelihoods of all sub-band observations. Both weighting schemes allow different recombination levels such as HMM state level, word level or other sub-unit levels. The schemes were also investigated by Hermansky et al. [89]. Their experiments showed great potential of the multistream ASR technique for speech recognition in

noise.

The multistream approach reported in [20] has a few limitations. For example, the independent sub-band processing loses important information about correlation between sub-bands [19], such as spectral envelope shape. To better exploit the uncorrupted partial information, Hermansky et al. [89] trained an independent neural network for each possible sub-band combination, which is feasible provided the number of combinations is not too large. They employed a seven sub-band system and therefore 127 networks (MLPs) were needed. Their study showed a strong robustness to band limited distortions with manual selection of the correct network given the prior knowledge about the distortions. Hermansky et al. [89] also examined a few primitive methods to select the right combination based on some heuristics. The problem was further addressed in the 'full combination' approach [19, 83], in which the probabilities of different sub-band combinations are merged using a weighted sum method. The full combination approach has the advantage of avoiding the independence assumption between sub-bands and also allows orthogonalisation of the combinations. However, the utility of this approach depends directly on how accurately the combination weights can be estimated. While this is possible with stationary noise, adapting to changing noise conditions is still a challenging problem.

Ming and Smith [134] proposed a probabilistic union model to formulate the recombination of the sub-band features. The union model approach also considers all possible noise positions in order to find the best sub-band combination. It deals with the unknown partial corruption by calculating the union likelihoods of different feature combinations (i.e. using the 'or' operator to combine subsets of features). The order of the model can be chosen given the knowledge that we know how many sub-bands are corrupted, but no information about the corruption location is needed. In practice, a low model order can yield better phonetic discrimination and a high order can accommodate more corrupted features [135]. Therefore, the union model needs a balance between the feature corruption uncertainty and noise robustness.

2.3.3 Difficulties with Multistream ASR

In multistream recognition it is assumed that nothing is known *a priori* about which streams of the speech evidence are clean and which are corrupted. This problem is tackled by con-

sidering all possible noise positions in order to find the best match. As a result, although multistream ASR has shown some robustness against partial frequency band corruption, its ability to handle non-stationary noise is ultimately limited. The sub-band boundaries are determined and fixed at the training stage, therefore unable to adapt to varying noise conditions, especially at low SNRs. The low band resolution usually employed in multistream ASR also limits its ability to localise noise. Employing more and narrower sub-bands may be considered, but narrower bandwidth of each sub-band could also produce a poor phonetic discrimination [89]. More sub-bands will also significantly increase the number of possible sub-band combinations, which will bring significantly more computation load to multistream ASR. Hence typical multistream ASR systems so far have employed fewer than eight sub-bands.

In contrast, as we will see in the next section, missing-data techniques employ prior knowledge of the location of the reliable regions. Primitive CASA processes act to identify spectro-temporal regions of energy dominated by the speech source in the first step and recognisers are adapted to handle the incomplete data. Therefore the missing-data approach is able to handle varying noises, provided that the reliable speech regions can be identified. Furthermore, a much higher frequency resolution can be employed in the missing-data approach (usually from 30 up to 64 frequency channels), without significantly increasing the computation cost.

2.4 The Missing-Data Approach to Robust ASR

In everyday listening environments speech signals are naturally mixed with noise. Information is often lost in spectro-temporal regions that are energetically dominated by the noise – in these regions the speech source is effectively 'masked'. We refer to the masked regions as 'missing data'. Since the segregation of the corrupted signal will never completely recover all the speech evidence [33], a robust ASR system needs to handle the missing data condition. In this section we review a statistical approach to handling the missing data which has demonstrated some success – 'missing-data' ASR [35].

2.4.1 Is Missing Data a Problem to ASR?

If missing data occurs in a normal listening environment, is it possible to recognise speech without complete speech evidence? Before examining the missing-data approach to robust ASR, we will first discuss some properties of the speech signal which essentially enable recognition based on only partial data.

Time-Frequency Masking

The ear has an ability to analyse incoming sound in the spectro-temporal domain (similar to the Fourier transform). The fact that human listeners are able to perform robust speech perception in a multisource environment is largely attributed to two properties of speech energy distribution in the spectro-temporal plane. First, speech energy is concentrated in local regions and these high energy regions are typically *sparsely distributed* in time and/or frequency [31]. For example, vowels have concentrated energy close to formant peaks, and energy of fricatives is typically concentrated in high frequency bands. This means that when speech is corrupted by noise, there will be some regions that are totally dominated by the energy from the noise sources, and other regions where the amount of energy from the noise sources can be ignored compared to that from the speech source. If the noise is non-stationary, then more regions of speech are likely to be uncorrupted. Therefore simultaneous sound sources overlapping in time are only partially overlapping in the spectro-temporal plane.

The sparsity of speech energy is also a valid characteristic of compressive computer representations of sound, such as the cochleagram we discussed in Section 1.3.1. Fig. 2.2 demonstrates the masking effect of two simultaneous sound sources with cochleagrams. Fig. 2.2a shows a log-compressed cochleagram computed for a clean speech utterance. In Fig. 2.2b the same utterance is mixed with babble noise at a global SNR of 0 dB. Although both sound sources have equal global energy in the mixture, many important features of the speech, such as harmonics and spectral shape, are clearly visible in speech-dominated frequency regions above 700 Hz. These speech features are less clear in regions below 700 Hz, where the energy of the babble noise dominates. However, in these regions that are seemingly dominated by the babble noise, speech energy still exceeds that of the noise frequently. Fig. 2.2c displays the spectro-temporal regions in the mixture where the energy from the speech source is greater by

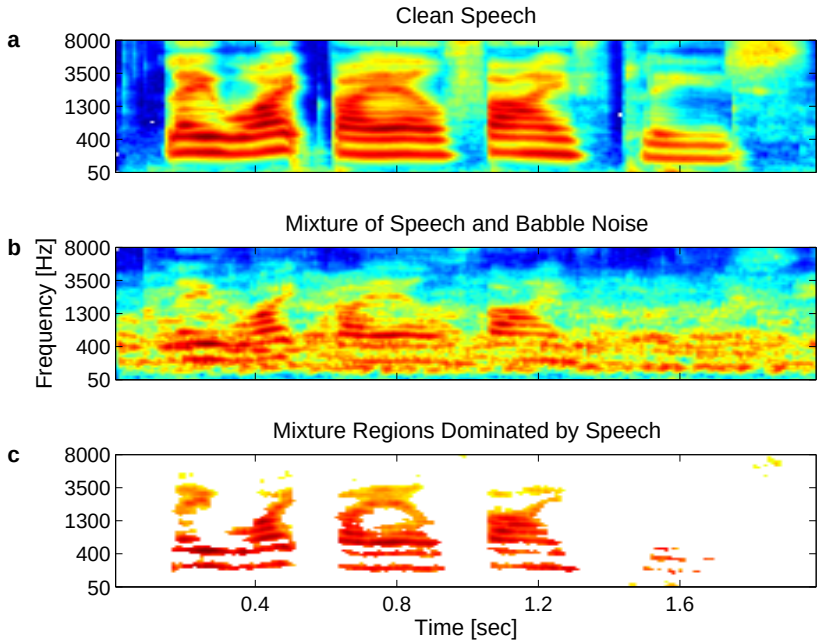

Figure 2.2: *Demonstration of the masking effect of two simultaneous sound sources. (**a**) A log-compressed cochleagram of the clean speech utterance 'lay white by L 5 please'. (**b**) The same utterance has been mixed with babble noise at a global SNR of 0 dB. (**c**) The spectro-temporal regions in the mixture (panel b) where the energy from the speech source is greater by at least 1 dB than that from the babble noise source (i.e. regions dominated by the speech energy).*

at least 1 dB than that from the noise source (i.e. regions dominated by the speech energy). This information is obtained using prior knowledge of pre-mixed signals. It is clear that most of the significant speech energy remains unmasked.

The second property of the speech signal is that it has a *redundant* encoding such that speech remains intelligible even when a large part of the speech spectrum is removed [67, 2, 188]. For example, intelligibility tests reported by Fletcher [67] demonstrated that speech syllables remains highly intelligible after all frequencies above 1800 Hz have been artificially removed. What is striking is that if the frequencies above 1800 Hz are retained and low frequencies are removed, the speech is equally intelligible. In fact, experiments using mutually exclusive frequency bands demonstrate that there is no one frequency band that is essential for human speech perception. Subjective listening experiments [188, 189] also suggested that speech intelligibility remains high even with an extremely narrow band (1/3 of an octave). In the

time domain similar masking effects are observed. When portions of an speech utterance are replaced by broadband noise less than 10 times per second, the utterance sounds natural and continuous [133]. Hence, a certain degree of information lost due to masking will not necessarily decrease overall speech intelligibility – the redundancy allows listeners to perceive speech in noise based on relatively sparse information.

The listeners' ability to exploit these properties to achieve robust speech recognition has motivated many researchers to take an alternative perspective to the noise robustness issue. Inspired by ASA studies, the solution given by Cooke et al. [35] is the 'missing-data' speech recognition technique, which is based on the exploitation of inherent redundancy in the speech signal rather than explicit characterisation of the noise. Researchers have demonstrated that ASR can be based on a very small amount (e.g. 10%) of the original time/frequency (T/F) components without serious deterioration in the recognition accuracy [34].

The Log-Max Approximation

The motivation behind missing-data ASR has a very good visual analogy. When part of an object is blocked (i.e. masked) by other objects, its identity may still be revealed based only on the visible parts of the object. One of the most important reasons that the partially blocked object can be identified is the principle of 'exclusive allocation' described by [21], i.e. each visual element can be exclusively assigned to an individual source. Otherwise all the present objects will be blurred together and no perceptual segregation is possible. In fact, the missing-data problem has been a subject of many studies in computer vision [e.g. 1, 73]. However, unlike predominantly *opaque* visual objects, sound signals combine additively and all present sources will contribute to the energy observed at each T/F point. With this energy combining attribute, is it reasonable to apply the same principle to audio signals?

Fortunately, the principle of exclusive allocation appears to be a good assumption for the speech signal justified by the sparsity of speech encoding in the spectro-temporal plane. When two sources combine, even if they have similar energy at a coarse scale, at a fine scale the magnitude ratio between the two sources in most T/F components is so great that the weaker one can be safely ignored. Especially when using a representation that compresses the energy, for example, via a log function, the observed log energy of the combination is very

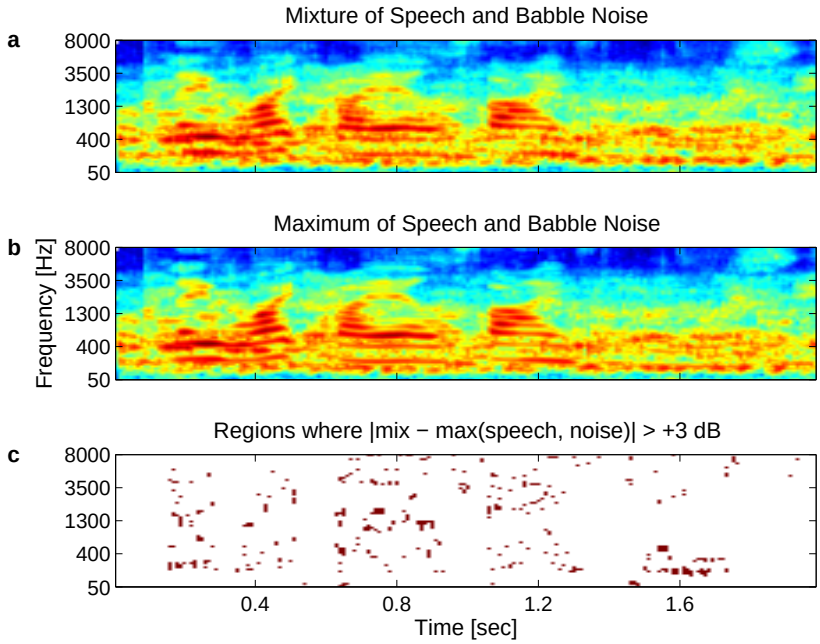

Figure 2.3: *Illustration of the 'log-max approximation'. (**a**) A cochleagram (log-compressed) of the mixture of speech and babble noise at a SNR of 0 dB. (**b**) The energy maximum of the individual sources prior to mixing. (**c**) Time/frequency components in which the log energy of the mixture is more than 3 dB different from the maximum of the log energy of the individual sources (4.9% of elements in this example).*

close to the log energy of the larger component, i.e. $log(x_1 + x_2) \approx \max(log(x_1), log(x_2))$. This observation is commonly known as the 'log-max approximation' [182, 164].

Fig. 2.3 illustrates the log-max approximation. Fig. 2.3a shows a log-compressed cochleagram of the mixture of speech and babble noise at a global SNR of 0 dB. The energy is log-compressed. Fig. 2.3b shows the maximum of the cochleagram energy of the two sources in isolation. This can be generated by first computing cochleagrams of the two sources prior to mixing, and then comparing the energy between them 'pixel-by-pixel'. The two cochleagrams are combined by selecting each T/F component from the one with the maximum energy value. It is clear that that the product is almost identical to the mixture cochleagram shown in panel **a**. This can be further confirmed by Fig. 2.3c, which shows T/F components in which the log energy of the mixture is more than 3 dB different from the maximum of the log energy of the individual sources. These elements are rare and sparsely distributed. In fact, in this example there are only 4.9% of the total T/F components that have an energy difference more than

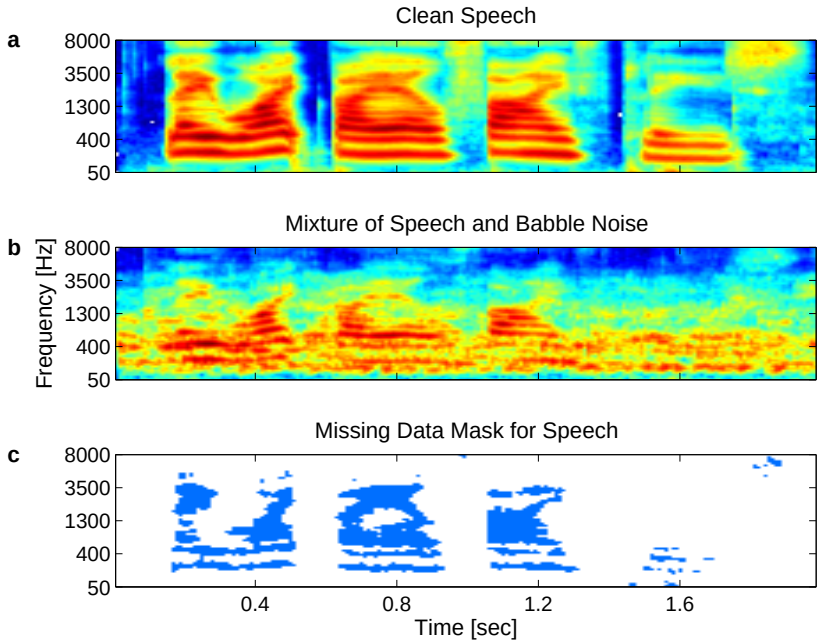

Figure 2.4: *(**a**) An auditory spectrogram of the clean speech utterance 'lay white by L 5 please'. (**b**) The same utterance has been mixed with babble noise at a global SNR of 0 dB. (**c**) The 'oracle' missing-data mask for speech representing spectro-temporal regions that are dominated by the target speech source.*

3 dB, which can be safely ignored compared to the dynamic range of the speech signal.

2.4.2 The Missing-Data Mask

The missing-data approach assumes that when speech is corrupted by noise, some spectro-temporal regions will remain unmasked (the sparsity) and can be identified as reliable evidence for recognition (the redundancy) using models trained on clean (i.e. noise-free) speech. Extensive research on CASA has shown that the sparsity of speech energy distribution allows primitive grouping principles to identify reliable regions that are not masked by noise sources in the spectro-temporal domain. This information is usually represented as a binary spectro-temporal map, referred to as the discrete 'missing-data mask', in which each T/F component is labelled as being either 'reliable' or 'unreliable'. The missing-data mask is typical output of many CASA systems, which essentially motivated the development of the missing-data ASR techniques.

Fig. 2.4c shows an 'oracle' missing-data mask [1] for the mixture example shown in Fig. 2.2, which represents spectro-temporal regions that are dominated by the speech source. The oracle mask is usually obtained by making use of prior knowledge of pre-mixed speech and noise signals. Cochleagrams of the pre-mixed signals are compared 'pixel-by-pixel' and those time/frequency components where the speech energy is higher than the noise energy are labelled as being reliable.

With the discrete missing-data mask wrong decisions made in mask estimation are irreversible. Therefore poor mask estimation has significantly impact on recognition performance. One method to limit the effect is to 'soften' the binary decision. The missing-data mask was extended by Barker et al. [11] to a 'soft mask', in which each T/F component is associated with a probability value in the range $[0, 1]$ expressing a degree of confidence in the reliability of the data. As a result, feature components are no longer exclusively labelled as either reliable or unreliable, which allows recruitment of features based on how well they match recognition models.

2.4.3 Application to ASR

Application of missing-data techniques to robust ASR requires the solutions to two problems. First, regions of reliable acoustic evidence (i.e. a missing-data mask) need to be identified. Estimating the missing-data mask remains a challenging problem. Solutions range from simple signal processing techniques to complex CASA models. Section 2.4.4 will review some of the solutions. Second, the recognition system needs to be modified to handle the missing data. Two commonly employed strategies are reviewed here.

In conventional HMM-based speech recognition each speech unit is typically represented by an HMM with a number of states. Each state, q, is characterised by a multivariate mixture Gaussian distribution over the components of x from an observation sequence, X. Speech recognition in general assigns an observation vector x to a state q. Essentially the state likelihood $p(x|q)$ needs to be computed. In the missing-data approach HMMs are trained using clean speech and there is no re-training for noise conditions. However, each observed acoustic feature vector, x, may be corrupted by noise and therefore $p(x|q)$ cannot be computed

[1] The oracle missing-data mask in some literature is also referred to as the *a priori* mask or the ideal binary mask.

directly. Let us assume that some preceding segregation process (e.g. CASA systems) has partitioned x into reliable components, x_r, and unreliable components, x_u. The order of the components in each feature vector x can be rearranged without loss of generality so that we can write $x = (x_r, x_u)$. There are essentially two approaches to classification with unreliable components x_u: *marginalisation* which evaluates $p(x|q)$ by considering all possible values of x_u, and *imputation* which first estimates values for x_u and then computes $p(x|q)$ based on the reconstructed features. Each approach has its own advantages and disadvantages.

Marginalisation-Based Approach

Many techniques in the marginalisation-based approach were originally developed by researchers at Sheffield University. Detailed analysis can be found in the frequently cited missing-data ASR paper [35]. Marginalisation bases classification on the marginal distribution of the reliable features by integrating over the unreliable components x_u in the state output distributions:

$$p(x|q) = \int p(x|q)dx_u = \int p(x_r, x_u|q)dx_u \tag{2.6}$$

Missing-data masks are naturally defined in the spectral domain, therefore most missing-data work is based on spectral features (e.g. the cochleagram representation we have seen in Section 1.3.1). However, traditional speech recognition systems often employ coefficients in an orthogonal domain derived from log spectra such as MFCC. Cepstral features are decorrelated such that they allow fewer dimensions and diagonal covariance Gaussians to be used. However, the cepstral transform also smears corruptions localised in the spectral domain over the entire feature vector, which brings difficulties for the detection of corrupted cepstral components. Unlike the cepstral representation, spectral features have a high degree of correlation across feature dimensions. In order to produce effective acoustic models using spectral features early missing-data work [81, 34] employed multivariate Gaussian distributions with full covariance matrices. However, with the incomplete features employing the full covariance model is not just computationally heavy but also impractical as the knowledge about which features are missing is not available at the training time. A more flexible and efficient way for modelling the correlation between spectral features is to employ a Gaussian mixture model (GMM), in which the distribution $p(x|q)$ is modelled by a number of Gaussian distributions with diagonal covariance matrices. Exploiting the independence within each mixture

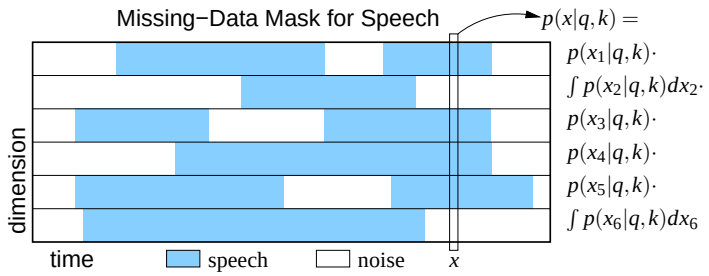

Figure 2.5: *Illustration of evaluating state likelihood with missing data based on marginalisation.*

component, we get:

$$p(x|q) = \sum_{k=1}^{M} P(k|q)p(x|q,k) \tag{2.7}$$

where $P(k|q)$ is the weight for the mixture component k. Assuming the components of x are independent, Eq. 2.6 becomes:

$$p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i \in r} p(x_i|q,k) \prod_{i \in u} \int p(\hat{x}|q,k)d\hat{x} \tag{2.8}$$

where $p(x_i|q,k)$ is the univariate Gaussian distribution.

The marginalisation approach to evaluation of the state likelihood with missing data is illustrated in Fig. 2.5. The integral term introduces constraints on the true values of the unreliable components. In the case where the unreliable components are completely ignored it reduces to unity (i.e. integral from $-\infty$ to $+\infty$). If x is a spectral energy vector in which the unreliable channels are contaminated by additive noise, the true speech energy in these channels must lie between zero [2] and the observed energy x_u. This forms an additional constraint that can be applied by bounding (or limiting) the range over which the unreliable features are integrated. Applying the 'bounds constraint' in Eq. 2.8 we obtain the bounded marginal estimation of $p(x|q)$:

$$p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i \in r} p(x_i|q,k) \prod_{i \in u} \frac{1}{x_i} \int_0^{x_i} p(\hat{x}|q,k)d\hat{x} \tag{2.9}$$

In the bounded marginals the integral term gets bigger as more of the probability mass associated with a particular state lies in the bounded range defined by the observed energy

[2] For cube-root compressive feature representations. For log-compressive representations the lower bound can become negative and a negative lower bound should be used.

x_u, i.e. given a low x_u quieter states (with lower means) will score better than more energetic ones. It effectively represents the 'counter-evidence' [45] against a particular state. The use of bounded marginals has been shown to produce consistent performance improvements over using unbounded marginals [34]. For multivariate Gaussians, the integral required to evaluate the bounded marginals can be approximated by a difference of error functions [35].

Eq. 2.9 assumes the missing-data mask to be discrete. With a soft mask as each feature component is no longer exclusively labelled, it should make a weighted contribution to both the reliable term and the unreliable term. This is reflected in the way of evaluating the likelihood $p(x|q)$:

$$p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i=1}^{N} \left(w_i p(x_i|q,k) + (1-w_i) \frac{1}{x_i} \int_0^{x_i} p(x'|q,k) dx' \right) \tag{2.10}$$

where N is the dimension of the feature vector x, and w_i is the probability that the i^{th} feature is reliable, which is defined in the soft mask. The newly introduced factor $1/x_i$ is a normalising constant. Note with Eq. 2.10 when the probabilities in the soft mask become binary, the distribution is equivalent to that defined by Eq. 2.9. The idea of soft masks is similar to the use of a reliability measure in the missing-data imputation work by Renevey and Drygajlo [162]. With the use of soft masks Barker et al. [11] have shown significant recognition accuracy improvement over the standard bounded marginal approach with discrete masks.

Marginalisation-based techniques are attractive because they follow the paradigm of using existing knowledge about the signal, as listeners are known to do. However, the fact that they require unorthogonal spectral features constrains their application to large vocabulary tasks and many existing 'optimal' ASR systems. In the next section we will review an alternative strategy which allows missing-data techniques to be utilised in orthogonal feature domains.

Imputation-Based Approach

Data-imputation approaches to the missing data problem involve estimating values for the unreliable features. In a state-based imputation scheme [103, 35] a separate reconstruction within each HMM state is formed for an HMM-based recogniser. For any state, the distribution $p(x|q)$ can be computed by replacing the unreliable components x_u by their maximum *a posteriori* estimates, $p(x_u|x_r, q)$, obtained given the knowledge of the reliable components

x_r and the prior distribution of that particular state q. Although it is possible, state-based imputation does not directly reconstruct the complete spectra for transformation into the cepstral domain. Experiments using spectral features demonstrate that the marginalisation-based approach consistently gives better recognition results than the state-based imputation scheme [137, 35].

If imputation is able to reconstruct complete spectral feature vectors then speech recognition can be performed in a standard manner. This approach is obviously attractive as with the reconstructed observation vectors a vast number of techniques available in traditional ASR can be employed. For example, many researchers used imputation to estimate the complete spectral feature vectors which are then transformed into the cepstral domain in order to make use of many ASR systems already well built using cepstral coefficients [e.g. 61, 156, 161, 158]. This scheme is called 'feature compensation', in which the unreliable spectral components are estimated based on the reliable components and the statistical properties of spectral vectors of clean speech. Most techniques in this category model the distribution of clean speech spectra using a Gaussian mixture model. Early work includes [61, 156], which compute a minimum mean square error (MMSE) estimation of the unreliable spectral components based on the reliable components, but with complete ignorance of the observed energy of the unreliable components (the bounds constraint). Renevey and Drygajlo [161] estimate the missing features in a similar way, but adapt the Gaussian mixture model using parameters from an explicit statistic model of additive background noise.

Raj et al. [158] present two reconstruction techniques for feature compensation which exploit information represented in the bounds. The first one is the 'cluster-based reconstruction' which models (or clusters) the spectral vectors of clean training data with a Gaussian mixture model. This technique assumes that the clusters have Gaussian distributions. Clustering is accomplished via conventional EM re-estimation [53]. To estimate the unreliable components the cluster to which an incomplete vector belongs is first identified based on the reliable components, i.e. the marginals. The distribution of that cluster is used to impute the missing values given the present values. The second technique is the 'correlation-based reconstruction' in which log-spectral vectors of the clean speech are considered to be samples of a stationary Gaussian random process. The correlations (both spectral and temporal) between any elements in pairs of spectral vectors in clean speech are learnt. The MAP estimates of unreliable

features are obtained based on all the reliable neighbouring features whose correlation with the unreliable feature is above some fixed threshold according to the learnt knowledge. By modelling temporal correlations between features, the correlation-based technique is able to reconstruct features even when entire frames are missing. It is also computationally cheaper than the cluster-based technique. However, speech recognition experiments [157, 158] show that the cluster-based reconstruction, evaluated on the DARPA 1000-word Resource Management task [152], performs generally better than correlation-based reconstruction.

Applying missing-data techniques in the cepstral domain generally requires significantly more computation. More recently, van Hamme [179] reported an imputation technique which maximises cepstral likelihoods subject to the bounds constraint thus imputing missing components directly in the cepstral domain. The 'PROSPECT' representation (PROjected SPECTra) of speech was proposed as an efficient alternative to the cepstral features. In their experiments on the Aurora 2 task PROSPECT shows comparable accuracy to cepstral features at high SNRs but is computationally cheaper [179, 181].

When measured on the Resource Management task, Raj et al. [158] showed that employing the cepstral features derived using these imputation techniques gives better recognition performance than that of marginalisation. However, the practical benefit is largely from the decorrelation of features which are well matched to HMM/GMM recognition systems. When recognition is performed in the spectral domain, marginalisation shows superior performance [35, 158]. In principle, bounded marginalisation is preferable to imputation because the latter demands an estimation of an individual value which depends on adequacy of reliable features and may be incorrect. In contrast, the marginalisation-based approach can take into account all the possible values within a probabilistic framework. Furthermore, with oracle masks obtained using *a priori* information the marginalisation shows striking performance [35], which suggests a great potential of this technique.

2.4.4 Missing-Data Mask Estimation

As we have seen in the previous section, application of missing-data techniques requires some preceding segregation process to identify reliable time/frequency components. This information needs to be represented by either a discrete mask with a binary reliable/unreliable

decision for each element, or a soft mask where each element is assigned a probability value to express a confidence of its reliability. The missing-data approach is attractive mainly because it does not make strong assumptions about the background noise. Provided that reliable speech evidence can be identified in the first stage, the decoder requires only models trained using clean speech and there is no need for explicitly modelling the noise. Therefore, the missing-data mask should be ideally produced using knowledge of the speech source with minimum assumption about the noise sources present.

Primitive CASA grouping principles are based on innate constraints of the incoming acoustic data and the physics of sound. Partial descriptions of present sound sources can be recovered by grouping spectro-temporal elements that have sufficient common features (e.g. harmonicity and spatial location). Rather than tailored to a particular type of sound, the same grouping principles can be applied to general sound sources. Therefore employing CASA-based techniques for missing-data mask estimation is very appealing. Auditory segmentation by CASA includes simultaneous (or spectral) grouping which organises sound components across frequency, and sequential (or temporal) grouping which links segments to form continuous temporal streams. A detailed review of CASA models can be found in [32] and for more recent CASA development readers are referred to [24].

It should be noted that many of these CASA-based algorithms were not originally designed for ASR and were often evaluated using different criteria (see Section 1.3), partially because of the lack of an advanced statistical framework. The missing-data mask can be the natural output of many primitive CASA systems, and therefore acts like a bridge to link CASA and ASR within the missing-data framework. In this section we will review some recent (post–2000) CASA-inspired systems that have been successfully applied to missing-data ASR.

Local SNR Estimation

Although CASA is an appealing choice, missing-data mask estimation is not limited to CASA models. Many systems have achieved good recognition results using masks derived from simple signal processing techniques. For example, a simple technique can be employed for stationary noise on the basis of local SNR estimates. The stationary noise spectrum is obtained by averaging the noise spectrum over short periods where it appears that no speech

is present. T/F components are labelled as reliable if the local SNR is above 0 dB, i.e. the observed energy is greater than the noise estimate. This process effectively produces a 'SNR mask' [35]. This discrete decision can be softened by mapping the local SNR estimates onto the range $[0, 1]$ using a sigmoid function [11] centred at 0 dB. To allow a margin that can guarantee the reliability of present data, the centre can be shifted to a higher SNR, e.g. at 3 dB. With such a margin only T/F components whose SNR is higher than 3 dB will have a probability value greater than 0.5. If a Gaussian noise distribution can be estimated, then a true probability can be computed to generate a soft mask [162]. Although simple, these techniques are surprisingly effective for small vocabulary tasks, as demonstrated in [12]. The soft values in the mask give missing-data systems an ability to recover mistakes made in mask estimation to some extent. However, for speech corrupted by highly non-stationary noise more general solutions are needed.

Harmonicity

An implementation of primitive CASA was also attempted in [13] which based mask estimation on harmonicity cues. The harmonic structure of voiced speech is known to provide cues for auditory grouping [43, 46]. The technique assumes that the speech signal is the only (or dominant) harmonic source present in the sound mixture. Therefore any T/F components with a certain degree of harmonicity can be grouped together as the missing-data mask. The degree of harmonicity of each element is determined using the 'autocorrelogram model' [117], a popular computational model of auditory pitch analysis. Chapter 6 gives a detailed review of this model. In brief, it computes autocorrelation on the output of each frequency channel of an auditory periphery model to reveal sub-band periodicity. The signal periodicity at each time frame can be emphasised by summing the autocorrelogram across frequencies. For a periodic signal the autocorrelation delay (apart from the zero delay) which gives the largest peak corresponds to its fundamental frequency ($F0$). This process is illustrated in Fig. 2.6. Once the location of the $F0$ delay is identified, a slice through the autocorrelogram is taken at this delay. The degree of harmonicity of each frequency component is computed as the ratio of the energy at the delay to the zero delay energy. If a time/frequency component shows a high degree of harmonicity the energy ratio will be expected close to 1. These energy ratios are rescaled using a sigmoid function to form a soft 'harmonicity mask'.

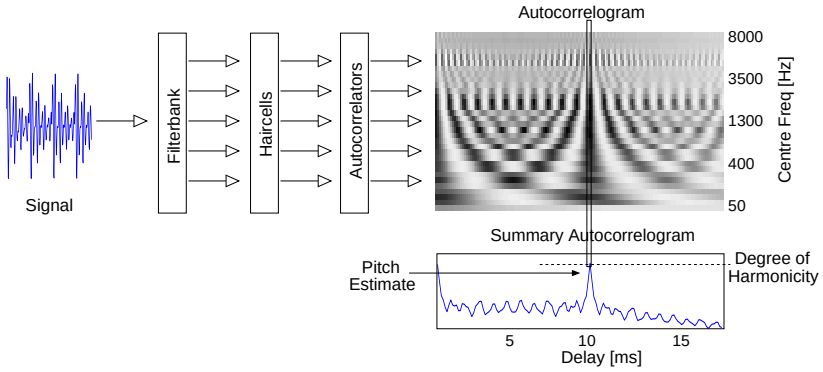

Figure 2.6: *Estimating the missing-data mask based on harmonicity (after [13]).*

Estimating masks based on harmonicity will inevitably fail for inharmonic speech regions (e.g. unvoiced speech). Barker et al. [13] therefore combined the harmonicity masks with SNR masks, which produce improvements over SNR masks alone. However, this technique is unable to deal with the situation where the noise is also a harmonic source (e.g. speech with music in background). The presence of non-speech harmonic source will inevitably result in most T/F components being labelled as reliable which hampers speech recognition.

The problem of how to combine harmonicity and SNR-based cues for missing-data speech recognition was also addressed by Brown et al. [25] using a 'neural oscillator' mechanism. Neural oscillator models have been successful at providing accounts of the interaction of cue combinations [186]. Brown et al. [25] associated each T/F component with a node in an oscillator network. The oscillators responding to related events will automatically synchronise in time and the associated T/F components can be grouped together.

Grouping by harmonicity has been an important cue for sound segregation in many CASA models [e.g. 23, 186] in which the autocorrelogram model is often employed as a front-end to reveal the harmonicity. Based on a similar idea to that of Wang and Brown [186], Hu and Wang [96] proposed a mask estimation technique which treats low-frequency and high-frequency regions differently. The motivation is from the psychophysical evidence that the human auditory system uses different mechanisms to deal with resolved and unresolved harmonics [27]. The system uses both $F0$ and amplitude modulation (AM) cues to group frequency components locally which are then linked across time based on temporal continuity. Evaluated using a conventional SNR metric the system performs better than the Wang-Brown

model [186]. However, it again assumes that the target speech is the dominant harmonic source in the mixture. If the intrusion also presents a harmonic structure, e.g. simultaneous speech, its performance is relatively limited. Although the system is able to output a binary mask, it has not been evaluated as an ASR front-end. Shao and Wang [171] employed this model in a missing-data based speaker identification system, which shows superior results compared to conventional systems.

Common Onsets/Offsets

Unvoiced speech, such as unvoiced consonants, does not have a harmonic structure. Therefore the models exploiting harmonicity cues, reviewed so far, are unable to handle unvoiced speech. Since the energy of unvoiced speech tends to concentrate on local spectro-temporal regions, Hu and Wang [97] presented a technique for unvoiced speech segmentation on the basis of common onsets/offsets analysis [23]. The technique works by grouping T/F components with synchronised energy onsets/offsets into local source regions. Regions with voiced energy are then removed. A binary Bayesian classifier is applied on the remaining regions in order to determine which fragments belong to unvoiced speech. The classification is performed with Gaussian mixture models which are trained using features including the spectrum and duration. In order to deploy the Bayesian classifier, the system requires a background model which is trained using various types of non-speech interference. Hu and Wang [97] assumed the background model to be generic which can accommodate most non-speech sounds. The quality of the background model is critical to the performance of the technique. Therefore their system lacks a general solution to unpredictable intrusions. It is also unable to tackle the situation where the intrusion is speech, e.g. cross-talk.

Hu and Wang [98] extended the model to segment both voiced and unvoiced speech by using a multiscale integration in which the onsets/offsets were examined at various scales. However, their system did not address the issues how the segmented regions that belong to the same source can be identified. A similar technique was reported in Coy and Barker [39], which employed a common image processing approach, known as the 'watershed algorithm' [78], to process the speech mixture after its harmonic regions are removed. The system produces 'inharmonic fragments' (see Section 6.5), which are then combined together with harmonic regions as input to an ASR system.

Spatial Location

Another popular way to construct missing-data masks is to use cues from sound source direction if more than one sensor is available [e.g. 130, 163, 144, 84]. However, as this study is concerned with single-channle signals, work on this topic will not be reviewed in this thesis. Readers referred to [187] for a detailed review.

Bayesian Classification

Rather than relying on the quality of noise models, researchers at Carnegie Mellon University addressed the problem of mask estimation as a Bayesian classification task based on the idea of extracting features which distinguish speech from noise [169]. Although a noise model is used, the features they select ensure a narrow distribution for the speech class so that the decision boundary of the classifier is insensitive to the broad distribution of the noise class. For voiced speech they employ harmonicity-based features extracted from harmonic comb-filter output and autocorrelation function, and additional features based on sub-band energies, spectral shape and kurtosis. For unvoiced segments only harmonic-independent features are used. Each class (reliable/unreliable) is modelled by Gaussian mixture models with a single full-covariance matrix and a separate classifier is constructed for unvoiced segments. Evaluated on the Resource Management corpus [152] the system produced superior recognition results across various SNR levels and noise types to those of conventional mask estimation techniques based on noise models and local SNR estimation.

Reverberation

The missing-data approach has also been adapted to deal with reverberation. In reverberant conditions, some spectro-temporal regions will be dominated by reflected sound, which can be potentially identified because it does not come from the same direction as the direct speech. In the system described in Palomäki et al. [144] the reverberant problem was also addressed using similar binaural processing. Spectro-temporal regions contaminated by reverberant energy usually have clean beginnings with more noisy reverberation tail to follow, which demonstrate modulation characteristics different from direct speech. Palomäki et al. [143] therefore apply a band-pass modulation filter centred around 4 Hz to each frequency band and

the resulting filtered signal captures the onsets of strong speech modulations. The process is similar to the modulation spectrogram [109]. After filtering, spectro-temporal regions with sufficient energy are considered to be dominated by direct speech and reliable. The remaining regions dominated by reverberant energy are ignored during recognition. Since not all the T/F components will be equally contaminated by reverberation, the technique can produce a 'reverberation mask' with soft values. Their experiments show that the system performs well over a variety of reverberant conditions.

2.4.5 Limitations

The missing-data approach to ASR provides a statistical way to make use of CASA output without the need to resynthesise the speech signal. It has demonstrated striking performance in adverse acoustic environments when the reliable speech evidence can be accurately identified. However, most reported experiments were performed on speech with artificially added noise. This is an unrealistic situation which assumes speech production remain unchanged under clean and noise conditions. It is known that in response to background noise talkers increase their voice levels to maintain adequate conditions for speech communication – the Lombard effect [106]. The increase of voice levels will cause variations in speech properties and therefore a mismatch between observed speech and recognition models trained using clean speech. Another significant difference in simulated and real situations is the effect of reverberation [142]. For missing-data ASR to be applied to realistic situations, techniques which can help accommodate these situations are necessary.

Another fundamental concern with all the approaches discussed so far is that the problems of segregation and recognition are decoupled. This strategy is probably due to the fact that these two fields emerged separately and were historically addressed by different research communities. Speech recognition can be seen as a pattern matching problem which aims to find a sequence of words that best match the observed acoustics. These recognition models (both acoustic and language models) employed by ASR are essentially learnt patterns of speech. Although these models are arguably imperfect representations of speech, they still provide some schema-driven information.

There is much evidence that listeners with better language-specific knowledge will perceive

the language more easily – a common experience when learning a new language. For example, Cooke et al. [37] recently examined the non-native speech perception problem. They asked a group of English and Spanish (who learnt English as the second language) listeners to identify keywords in English sentences in three listening conditions: quiet, corrupted by stationary noise and corrupted by a competing speaker. The result across the two language groups shows that in the simultaneous-speaker condition both groups benefited equally from differences in fundamental frequency between the two speakers. This suggests that processes which make use of the pitch cue (presumably mainly the primitive grouping process), are not affected by language-specific knowledge. At the same time, however, the experiments also show that the Spanish speakers suffered more from increasing levels of noise than the native English speakers, presumably due to the lack of good knowledge of English. The non-native speech perception experiment suggests that recognition models can provide schema-based constraints and the problems of source separation and recognition should be tightly coupled.

One convenience of the decoupling is that the CASA front-end and the ASR back-end can be developed independently. This assumes that, however, source separation is an easier task than recognition and can be achieved without top-down speech schemas (i.e. the recognition models). Nearly all successful CASA systems so far have been based on data-driven processes. This ignorance of top-down schema-based processes is itself in conflict with the auditory scene analysis principles we have learnt from human speech perception. Current data-driven CASA systems can separate sources across frequency with a reasonable performance, based on the pitch cue or a combination of many cues. However, most of them failed to give acceptable temporal groupings. Most researchers make use of cues such as $F0$ continuity [50] and spectral continuity [48]. These constraints can produce robust local temporal grouping (i.e. within a short period), but lack the power to handle long-term grouping. For example, the $F0$ continuity cue is unable to deal with ambiguous pitch tracks when two voices are present simultaneously. This long-term constraint required must come from the top-down knowledge of a sound.

2.5 Other Related Approaches to Robust ASR

While the performance of missing-data ASR on small vocabulary tasks is significantly better than many other conventional robust ASR approaches, its constraint on using spectral features greatly limits its application to large-vocabulary recognition tasks. It is well known that recognition using orthogonal features, such as cepstral coefficients, produces a superior performance in quiet conditions. Attempts to apply missing-data techniques to the cepstral domain have focused on reconstruction or imputation of the missing features in the spectral domain, followed by transformation to the cepstral domain [e.g. 61, 156, 161, 158]. However, potential errors introduced in the feature reconstruction stage are not accommodated, which limits the ASR performance. In this section, we will review some recent developments in robust ASR which are related to missing-data ASR and work in the cepstral domain.

In feature-based robust speech recognition noisy speech is typically preprocessed to remove noise before it is fed into the speech recognition system. However, the noise removal process often introduces errors. This is analogous to missing-data ASR using a discrete mask (see Section 2.4.2), where errors in the mask estimation are made concrete and irreversible. It has been demonstrated that missing-data ASR greatly improves when the hard decision to exclude unreliable features is softened by a continuous weighting [11, 12]. Recently, 'uncertainty decoding' techniques have emerged from a generalisation of this 'soft missing-data' approach, which effectively allows uncertainty in the noise removal process to be explicitly expressed [138, 3, 59].

The uncertainty decoding approach parametrically expresses the conditional probability $p(y|x;\hat{\lambda})$ as a measure of the uncertainty in the noisy speech y as a noisy estimate of clean speech x, where $\hat{\lambda}$ is the noise model. Such techniques have been shown to work well under assumed stationary noise conditions. Most techniques work directly in the cepstral domain and are coupled with denoising algorithms that work in the domain. For example, Droppo et al. [59] used the SPLICE technique [55, 58] to estimate the conditional probability associated with speech enhancement. In order to compute feature uncertainty, Droppo et al. [59] used so-called 'stereo data', in which noisy and clean data are simultaneous recorded and artificially mixed. A third-order polynomial was used to approximate the mapping function. A similar approach was also used in [3, 116]. The need of stereo training data was removed

in [56] by using an approach based on a parametric model of speech distortion to statistical feature enhancement.

In all the techniques discussed above the conditional probability is approximated by using an N-component GMM modelling the acoustic space. However, in low SNR conditions the GMM may generate very low conditional probability for all recognition models, which greatly reduces their discriminative ability [115]. To overcome this problem, Liao and Gales [115] link the conditional distribution with the recognition model components, similar to using regression classes in adaptation schemes such as MLLR [112]. Recently, Srinivasan and Wang [174] proposed an approach to the problem of estimating the uncertainty of cepstral features derived from a missing-data mask defined in the spectral domain. When evaluated on a subset of the Aurora 4 task [145] using oracle masks (i.e. ideal speech/noise segregation), the uncertainty decoding approach produced comparable results to missing-data ASR and significant reductions in WER compared to conventional recognition using enhanced cepstra.

When compared to model-compensation techniques such as PMC, the compensation cost in uncertainty decoding depends on the number of components used in the recogniser to model the feature space. This is significantly less than the computational cost required by PMC. Although uncertainty decoding is very similar to soft missing-data ASR, their motivations are different. While the former is focused on accommodate errors introduced in feature-compensation, the latter grew out of work on auditory scene analysis with motivation from human speech perception. It is interesting to directly compare the performance of these two approaches on common tasks in future.

2.6 Summary

There have been many years of research on engineering approaches to robust ASR [104, 77]. In this chapter we give an introduction of traditional robust speech recognition techniques and contrast them with some ASA-inspired approaches to ASR. Traditional ASR approaches to achieving noise robustness exploit the differences that are assumed to exist between the training and operating environments and try to minimise the mismatch typically using engineering solutions. Therefore they are typically designed to work well in highly predictable and narrowly specified noise conditions. Their performance often decreases dramatically if

the operating environment is not carefully controlled [54]. More importantly, they do not directly address the fundamental issue posed by complex auditory scenes.

Listeners, on the other hand, are adept at coping with a wide range of noisy environments. Largely motivated by its counterpart in the field of computer vision, computational auditory scene analysis emerged as a promising field to tackle the sound organisation problem with inspiration from extensive research of human speech perception. The appeal of CASA is that it does not consider speech as a different source from others and therefore the same ASA principles can be applied equally well to all the sources. This motivated many researchers to build ASR systems which can benefit from CASA studies. One successful approach is the missing-data ASR technique. However, as we have seen in the previous section, the decoupling of source separation and recognition ultimately limits such a simple 'left-to-right' strategy and more sophisticated solutions are needed. In this next chapter we will discuss the 'speech fragment decoding' (SFD) technique [15], which combines segregation and recognition in a tightly coupled process.

Chapter 3
Speech Fragment Decoding

This chapter will examine the 'speech fragment decoding' framework. A 'fragment' is a spectro-temporal region where energy from a single sound source dominates. SFD employs techniques developed from knowledge about the auditory system to identify fragments. A decoding process using statistical speech models is applied to the fragment representation to simultaneously identify speech evidence and recognise speech.

3.1 Introduction

The success of CASA has inspired research into developing more general solutions to the robust speech recognition problem for natural listening conditions where competing sounds are often present. One popular approach to link CASA with ASR is 'missing-data' ASR [35] reviewed in Section 2.4. The typical application of CASA in the missing-data paradigm is via the use of missing-data masks (see Section 2.4.2) – CASA models are employed as a front-end to identify reliable spectro-temporal regions for the target speech. Although missing-data ASR provides a statistical framework to handle the missing data identified by the CASA front-end, there are several limitations to such a left-to-right strategy.

First, missing-data ASR requires segregation of the target source (i.e. producing a missing-data mask). This is a challenging problem when dealing with highly unpredictable noise sources (e.g. simultaneous speech where the masker is also speech). Recognition performance is often poor if the missing-data mask is not correctly identified. Although CASA is appealing for generating missing-data masks, most CASA systems have only achieved success in spectral

grouping. They are relatively unreliable when organising sound components sequentially. Sequential grouping cues commonly employed include continuity in fundamental frequency and spectral shape, and spatial location [46]. For example, two successive segments with close pitch estimates tend to be grouped together. However, most of these cues are considered valid only over short time periods. They are less robust over longer periods due to the great variability of the speech properties over time. Therefore, instead of complete segregation, it is more natural and more reliable for primitive grouping systems to produce local spectro-temporal regions where the energy is likely to have originated from a single source. These local spectro-temporal regions are referred to as *fragments* or *coherent fragments* in this study. The identities of the fragments do not have to be decided until high-level top-down constraints (e.g. lexicons and semantics) are available (i.e. in the recognition stage).

Secondly, in missing-data ASR and many other robust ASR approaches, the processes of sound source segregation and recognition are decoupled. The problems of CASA and ASR were traditionally addressed separately, often by different research communities and with different research goals. A typical application in ASR is to employ CASA as a front-end which either resynthesises clean speech or provides a missing-data mask. However, nearly all CASA studies have focused on bottom-up signal-driven processing. As a result, prior top-down knowledge available in recognition models is ignored when the segregation hypothesis is formed. There is strong evidence that listeners also employ processes driven by learnt models of sounds and the two mechanisms work interactively to form a logical explanation of the present sound scenes [21]. It is highly likely that segregation is a by-product of recognition. In order to make significant progress in building a *robust* ASR system, the two processes have to be tightly coupled.

Inspired by the ASA account of auditory organisation, Barker et al. [15] proposed a 'speech fragment decoding' (SFD) technique which treats source segregation and recognition as tightly coupled problems. As speech energy is sparsely distributed [31], primitive ASA grouping techniques, such as multipitch analysis [126], are employed to segregate a spectro-temporal representation (e.g. the cochleagram) of the mixture into a set of fragments. Statistical model-driven processes then employ speech recognition models to simultaneously search for the most likely word sequence and foreground/background segregation.

3.2 Coupling Segregation with Recognition

In the statistical framework the automatic speech recognition problem is typically formalised as a search for the most probable word sequence, $\hat{W}$, given the acoustic speech signal, X (Eq. 2.3):

$$\hat{W} = \arg\max_{W} P(W|X)$$

In a multisource environment the speech signal X is not directly observed when the speech is mixed with noise from another source. Instead the acoustic mixture, Y, is observed. Many robust ASR approaches including missing-data ASR essentially require a segregation hypothesis, S, of the observed noisy signal, Y, which attempts to recover the speech energy from the mixture. Given this segregation hypothesis the recognition problem can be formulated as

$$\hat{W} = \arg\max_{W} P(W|S,Y) \tag{3.1}$$

Eq. 3.1 is a precise description of the left-to-right robust ASR strategy. In the missing-data approach the segregation hypothesis, S, is represented by a missing-data mask which marks spectro-temporal regions as either being dominated by the target source or masked by background. Section 2.4 has shown how the term $P(W|S,Y)$ can be evaluated by building statistical models of clean speech and partially matching the regions of Y that are marked as foreground to these recognition models.

However, Eq. 3.1 only considers a single proposed segregation which may be incorrect. To fully couple the segregation problem with recognition, the best segregation can be treated as a by-product of recognition. A better description is to search for the word sequence and segregation [1] that together are most probable given the noisy signal, Y:

$$\hat{W}, \hat{S} = \arg\max_{W,S} P(W,S|Y) \tag{3.2}$$

$$= \arg\max_{W,S} P(W|S,Y)P(S|Y) \tag{3.3}$$

where $P(W|S,Y)$ is equivalent to missing-data decoding and $P(S|Y)$ is the segregation model. Eq. 3.3 is the essence of the SFD technique. The search is now being conducted over the joint space of word sequences and segregation hypotheses. In practice, given each segregation S, the word sequence dimension of the search can be efficiently performed using missing-data

[1] In Eq. 3.3 max over segregation S is really approximating sum over S.

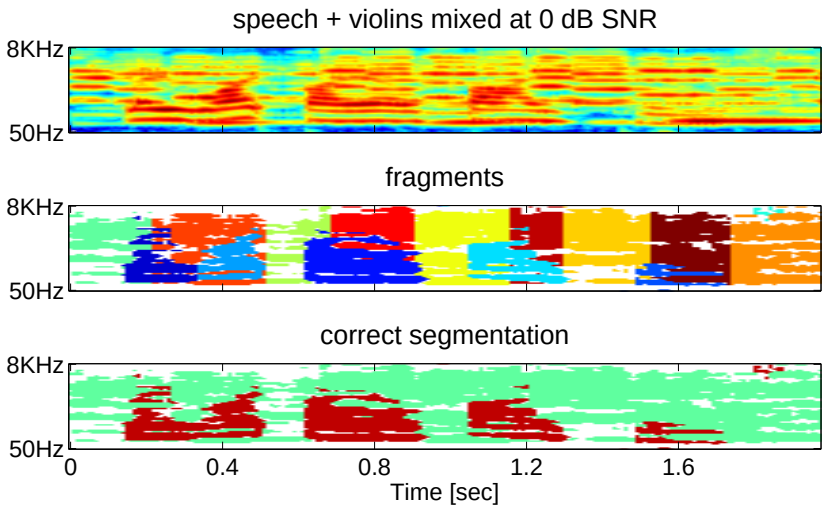

Figure 3.1: *An example of fragments for a speech/violin mixture (SNR = 0dB), together with correct source segmentation. In this example the fragments were generated using prior information of pre-mixed signals.*

techniques. The segregation search is then equivalent to selecting the missing-data mask that gives the best likelihood score from missing-data decoding. An exhaustive search is clearly not practical. If the observation sequence is composed of T frames and each acoustic vector consists of F frequency bands, then the acoustic mixture contains $T \times F$ time/frequency (T/F) components. There are potentially 2^{TF} segregation hypotheses to evaluate as each T/F component can be variously labelled as part of either foreground or background. A typical computer model employs 64 frequency bands and a frame rate of 10 ms. Therefore for a 2-second utterance there will be $2^{200\times64}$ ways of dividing the noisy mixture between foreground and background.

Fortunately, most of the segregation hypotheses do not need to be evaluated. Primitive grouping principles [21] can be employed to group T/F components according to the correlations of their characteristics. For example, T/F components may be grouped if they form continuous pitch tracks. The process results in the acoustic mixture being divided into multiple local fragments in the spectro-temporal plane. The middle panel in Fig. 3.1 shows an example of fragments for a speech/violin mixture. In this example the fragments were generated using prior information of pre-mixed signals. Each colour represents a fragment in which all the T/F components are dominated by energy from a single source.

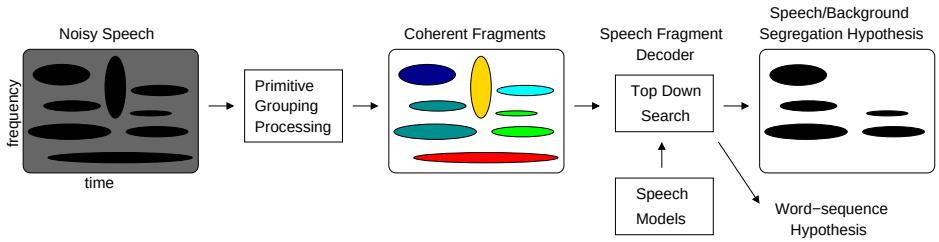

Figure 3.2: *An overview of the speech fragment decoding system (reproduced from [15]). Bottom-up processes are employed to identify spectro-temporal regions where each region is likely to have originated from a single source (fragments). A top-down search with access to speech models is then used to search for the most likely combination of fragment labelling and speech model sequence.*

By dividing the mixture into initial fragments the computation cost is significantly reduced. If the signal is segregated to M fragments, then the total number of segregation hypotheses becomes 2^M, as for each fragment both possible labels (foreground and background) are considered. The concept of fragments is consistent with the underlying principles of auditory scene analysis [15]. Each fragment (or a group of fragments) may correspond to an auditory source or event. Generating fragments is an easier task than completely recovering the speech source, as their identities do not have to be decided before the recognition stage. Therefore the primitive grouping principles can focus on correlations between local T/F components rather than long-term grouping constraints which are less reliable given only the signal properties. The final segregation is performed when top-down information encoded in the recognition models is available.

3.3 Fragment-Driven Speech Recognition

An overview of the SFD system is provided in Fig. 3.2. The technique works by considering all possible fragment labellings and all possible word sequences. Each fragment may be variously labelled as either being a fragment of the target (foreground) or of the masker (background). A foreground/background segregation hypothesis is defined by a unique selection of fragment labels which can be represented by a missing-data mask, m_{tf} – a spectro-temporal map of binary values indicating which T/F components are being considered to be dominated by the target source, and which are being considered to be masked by the competing sources. Given

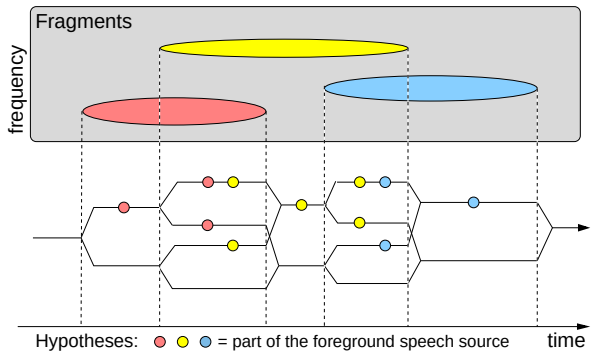

Figure 3.3: *Illustration of an efficient implementation of SFD (reproduced from [15]). The shaded dots indicate which ongoing fragments are being treated as part of the speech foreground. The absence of the dots means the fragments are being considered as part of the background. When a new fragment begins, the hypotheses split in order to consider both labellings of the newer fragment. When a fragment finishes, the hypotheses are merged and the best labellings continue to be propagated.*

such a segregation hypothesis, the decoder employs missing-data techniques to evaluate the likelihood of each hypothesised word sequence. A dynamic programming algorithm is used to find the most likely combination of labelling and word sequence.

3.3.1 An Efficient Decoder Implementation

Although by grouping T/F components into initial fragments the huge search space is significantly reduced, the number of segregation hypotheses under consideration still grows exponentially with time. An utterance of 2-second long can be typically divided into around 20 fragments. This means there are 2^{20} segregation hypotheses to be considered (i.e. missing-data decoding needs to be performed 2^{20} times). Barker et al. [15] demonstrated that this exponential growth may be prevented. Consider the segregation hypotheses frame by frame. A pair of hypotheses will become identical after the offset of the last fragment by which they differ. At this point, the two competing segregation hypotheses are compared and the less likely one can be rejected without affecting the admissibility of the search [15]. As a result, the number of segregation hypotheses under consideration at each frame is 2^N, where N is the number of fragments in parallel at that frame, which is typically less than 4 [126]. Note that the number of fragments being decoded simultaneously varies with time.

The search process is illustrated in Fig. 3.3 where three fragments (shown using the shaded regions) are being decoded. Each time a new fragment starts, all ongoing segregation hypotheses are split so that in each pair one hypothesis labels the fragment as speech while the other assigns it to the background. When a fragment finishes, pairs of hypotheses are merged if they differ only in their labelling of the particular fragment. This process continues until the end of the utterance.

3.3.2 Decoding with Confidence Maps

One weakness of the SFD technique, in the form described above, is that it produces 'hard' segregation, i.e. segregation in which each time/frequency component is marked categorically as either part of the foreground or background. If early processing has incorrectly grouped elements of the foreground and background into a single fragment, then there will be incorrect assignments in a missing-data mask which cannot be recovered in later processing. These problems can be mitigated by using missing-data techniques that use 'soft masks' containing a value between 0 and 1 to express a degree of belief that the element is either foreground or background [11]. Such masks can be constructed in the SFD framework by introducing a spectro-temporal map to express the confidence that the T/F component belongs to the fragment to which it has been assigned.

The confidence map, c_{tf}, uses values in the range 0.5 (equal belief in foreground and background) to 1.0 (high belief in foreground). The values are generally estimated based on fragment generation techniques (see Section 6.6). Given a confidence map, c_{tf}, each hypothesised fragment labelling can be converted into a soft missing-data mask, m_{tf}, by setting m_{tf} to be c_{tf} for T/F components that lie within foreground fragments, and to be $1 - c_{tf}$ for T/F components within missing fragments.

3.3.3 Deploying the SFD Framework

SFD requires identification of fragments. The preference is for larger fragments as the complexity of the decoding process scales as the number of simultaneous fragments. Finding coherent fragments is a simpler problem than identifying reliable speech evidence in the missing-data approach. Techniques based on computational auditory scene analysis can be

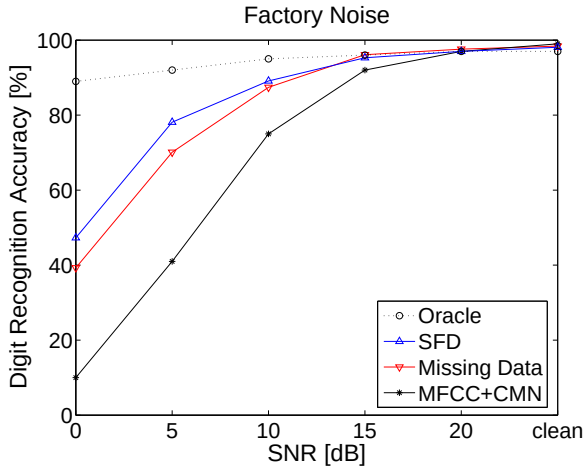

Figure 3.4: *Recognition results comparing the performance of a baseline system using cepstral mean normalisation (MFCC+CMN), regular missing-data, and the speech fragment decoder. The 'oracle' curve shows the result expected when perfect speech/noise segregation is achieved (after [15]).*

used to find cues (both bottom-up and top-down) to form coherent fragments. Short term cues include harmonicity, frequency proximity and common onsets/offsets. Long term cues include F0 continuity, speaker identity and the use of a lexicon and semantics.

Barker et al. [15] have shown that SFD provides significant improvements over a missing-data decoder. Their experiments employed mixtures of TIDigits utterances [113] and NOISEX factory noise [183] at various SNRs. They first identified speech regions based on local SNR estimation [35] (see also Section 2.4.4). The speech regions were represented as missing-data masks and employed by a missing-data decoder as part of the recognition experiment. Because the factory noise has highly unpredictable components such as hammer blows, some of the speech regions identified using the SNR-estimation technique will in fact be due to the noise. To allow the speech fragment decoder to improve on the missing-data results, the missing-data masks were dissected. Barker et al. first divided each mask into four frequency subbands. Each contiguous region within a subband was defined to be a separate fragment.

Fig. 3.4 shows recognition results at various SNRs. The SFD technique yields significant improvement in accuracy over the missing-data approach at the lower SNRs, even though the fragment generation processing employed is still somewhat rudimentary. Analysis of the results showed that noise fragments, which were identified as part of speech regions using the

SNR-estimation technique, were successfully rejected by SFD. The missing-data decoder does not have the ability to recover such errors and therefore produced lower recognition accuracy.

3.4 Possibilities for Improving SFD

Speech fragment decoding is a novel framework that takes advantage of heuristics developed from knowledge about the human auditory system. This power is combined with statistically trained schema-driven processes to give better performance than systems based on either taken in isolation. Therefore SFD provides the foundation for a statistical approach to coupling computational auditory scene analysis and automatic speech recognition. It should be noted that the SFD techniques can be applied to general source recognition (i.e. a multisource decoding framework), given that detailed models of the target source are available.

The recognition errors produced by a SFD system can be due to various reasons. SFD employs missing-data ASR at its core, and the decoding process can produce word-matching errors even given the correct foreground/background segregation hypothesis defined by an oracle missing-data mask (see Section 2.4.2). The errors can also be due to a result of selecting wrong fragments. At last, poor quality of fragments provided to SFD may also lead to recognition errors. With the implementation by Barker et al. [15], there are several possibilities for improving the current system, which will be investigated in this thesis.

First, traditional HMMs have weak duration constraints. This is not a problem when the operating environment is similar to that of HMM training. The corruption of acoustic features, however, often causes word matches with unrealistic durations to be produced by ASR in noisy conditions. SFD employs missing-data techniques at its core, which base speech recognition on partial acoustic evidence. To combat noise corruption, explicit duration modelling may need to be introduced into the decoding process. This will be investigated in Chapter 4.

Second, in the current implementation SFD assumes that each fragment is part of either the speech foreground or the noise background with equal probability. This essentially applies a uniform distribution to all the segregation hypotheses defined by the term $P(S|Y)$ in Eq. 3.3. The uniform distribution is a very crude approximation. Some fragments may 'look' more like speech and others may 'look' more like noise. For example, a thin fragment in the high

frequency region with a long span in time is unlikely to be from the speech source. Knowledge that may help distinguish speech fragments from noise fragments can be exploited to assist the decoder in the choice of fragments. This will be investigated in Chapter 5.

Finally, the recognition performance of SFD also depends on the quality of fragments. If the fragments contain too much energy that belongs to different sources, the performance is likely to be poor. The fragment generation technique employed by Barker et al. [15] is very simple and crude. It is likely that improved fragment generation techniques will result in significant performance gains. This study will therefore investigate ways that coherent fragments can be formed. This will be investigated in Chapter 6.

By investigating techniques for improving SFD in multisource environments we hope that some progress can be made towards finding a general solution to the robust ASR problem.

3.5 Corpora and Experimental Setup

In this section the corpora and common experimental setup employed in this doctoral work will be presented. Where appropriate, in each chapter the experimental setup may be briefly repeated to make the thesis easier to understand. Speech recogniser setup related to each experiment will be given at the beginning of each chapter.

3.5.1 Corpora

Two corpora were employed in this thesis. The Aurora 2 connected-digit database [150] was used for investigating duration modelling in noisy environments. This corpus takes the TIDigits database [113] as basis, which is mixed with various environmental noises. The original 20 kHz data are downsampled to 8 kHz with a low-pass filter extracting the spectrum between 0 and 4 kHz. Aurora 2 has a vocabulary of 11 words ('1'–'9', 'oh' and 'zero'), silence and inter-digit short pauses. It has a free grammar and each utterance may contain one or more digits. The number of words in each utterance is unknown.

The Grid corpus [36] was used in the rest experiments. The corpus consists of utterances spoken by 34 native English speakers, including 18 male speakers and 16 female speakers. The vocabulary contains 51 words. The utterances are short sentences of the form <COMMAND>

<COLOUR> <PREPOSITION> <LETTER> <NUMBER> <ADVERB>, as indicated in Tab. 3.1, e.g. 'lay white by L 5 please'. Each utterance lasts about 2.2 seconds. All signals are sampled at 25 kHz.

Table 3.1: *Structures of the sentences in the Grid corpus*

Verb	Colour	Prep.	Letter	Digit	Adverb
bin	blue	at	A–Z	1–9	again
lay	green	by	(excluding 'W')	and zero	now
place	red	in			please
set	white	with			soon

The Grid corpus provides a challenging ASR task and forms the base for the Interspeech 2006 Speech Separation Challenge [2]. Many techniques presented in this thesis were originally developed for the challenge. However, the Grid corpus is not suitable for investigating duration modelling. Because each Grid utterance has the same number of words (with a fixed grammar of the form '<COMMAND> <COLOUR> <PREPOSITION> <LETTER> <NUMBER> <ADVERB>'), ASR experiments on the Grid corpus will not produce any insertion or deletion errors, which duration modelling may help combat (see Chapter 4). Therefore the Aurora 2 database is used for investigating duration modelling.

3.5.2 Acoustic Feature Representation

All experiments presented in this thesis were performed using spectral features so that missing-data techniques based on marginalisation (which SFD employs at core) can be applied. Cochlear frequency analysis was simulated via a bank of overlapping gammatone filters with centre frequencies spaced uniformly on the equivalent rectangular bandwidth (ERB) scale [75]. The instantaneous Hilbert envelope is computed at the output of each gammatone filter. This is smoothed by a first-order low-pass filter with an 8 ms time constant, sampled at 10 ms intervals, and finally log-compressed to give an approximation to the auditory nerve firing rate – the 'ratemap' representation, or the cochleagram (see Section 1.3.1).

The number of gammatone filters employed is normally decided based on signal sampling rate and the frequency range covered. Having more filters can offer a higher frequency resolution

[2] http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

but bring more computational cost. For the Aurora 2 database, in which data are sampled at 8 kHz, the frequency range employed in cochlear frequency analysis is from 50 Hz to 3850 Hz, following [12]. The number of filters (channels) employed in the gammatone filterbank is 32. For the Grid corpus, in which data are sampled at 25 kHz, the frequency range employed in cochlear frequency analysis is from 50 Hz to 8 kHz, following [16]. The number of filters (channels) employed in the gammatone filterbank is 64.

These spectral features were supplemented with their temporal derivatives to form the final feature vector.

Chapter 4
Explicit Duration Modelling

4.1 Introduction

Hidden Markov models (HMMs) provide a powerful framework for modelling time-varying signals and speech recognition based on them has achieved great success. However, in the presence of noise ASR performance often degrades significantly. One reason is that HMMs do not directly characterise some important information such as duration modelling [153]. In a standard HMM employing a left-to-right no-skip topology, the probability of occupancy duration in each HMM state decreases geometrically with time. The first-order Markov process implicitly imposes a geometric state duration distribution which may be inappropriate for speech signals [155]:

$$d_i(\tau) = a_{ii}^{\tau-1}(1 - a_{ii}) \tag{4.1}$$

where a_{ii} is the self-transition probability of state i and τ is state occupancy duration. Furthermore, the geometric state duration density also produces a skewed (to the short duration side) word duration distribution with a variance much wider than empirical ones [80] and there is no hard limit on word durations. The implicit weak duration modelling may cause problems in the process of decoding, especially if there is a mismatch between the training and testing environments. For example, given models trained on clean speech an ASR system often produces word matches with unrealistic durations in noisy conditions. Word strings where the associated word models have short durations tend to be favoured over competing strings with fewer words but longer durations. This effect can be observed in a connected-digit recognition task with no grammar constraints, where the number of insertion errors greatly exceeds that of deletions and substitutions in noisy conditions [151].

In the SFD framework, acoustic evidence is often incomplete and the missing-data technique is employed to deal with the partial observations. In the lack of complete acoustic evidence, incorporating duration constraints may be particularly useful. This chapter will examine duration modelling in the context of missing-data ASR [35]. As the SFD system employs the missing-data theory at its core, techniques useful for missing-data ASR would also be useful for an SFD system. This allows us to employ many existing missing-data ASR systems developed at Sheffield university.

All experiments presented in this Chapter were performed using the Aurora 2 connected-digit corpus [150]. Spectral features were used so that missing-data techniques can be applied. Feature vectors were obtained via a 32-channel gammatone filterbank distributed in frequency between 50 Hz and 3850 Hz on the equivalent rectangular bandwidth (ERB) scale [75]. The features were supplemented with their temporal derivatives to form a 64-dimensional feature vector. Some of the work reported in this chapter has previously appeared in [122, 125]. Section 4.2 will discuss experiments of modelling state durations. Section 4.3 will present techniques modelling word durations. Particularly, a significant effect on word durations, the 'prepausal lengthening effect' [41], is investigated. Section 4.4 concludes and presents future directions.

4.2 State Duration Modelling

4.2.1 Overview

Many approaches have been proposed to model state duration information in an HMM based speech recognition system. The work by Ferguson [65] pioneered the use of an explicit state duration model. His model associates each HMM state i with a non-parametric state duration probability $d_i(\tau)$, $\tau = 1, 2, \cdots \tau^i_{\max}$, where $\tau^i_{\max}$ is the longest duration allowed for state i. The estimation of $d_i(\tau)$ was incorporated into the Baum-Welch re-estimation algorithm. Due to the introduction of $\tau^i_{\max}$ duration parameters for each state, the Ferguson model suffers from excessive computational load and more importantly, sufficient training data may not be available to estimate all the duration parameters.

Researchers have suggested using parametric state duration distributions to address the

problem of insufficient training data. For example, Russell and Moore [166] proposed using the Poisson distribution in addition to the Gaussian distribution, while Levinson [114] replaced the geometric state duration distribution with the Gamma distribution. Burshtein [26] showed that state durations are most accurately described by the Gamma distribution. In most research these parametric distributions were applied in the hidden semi-Markov model (HSMM) framework where temporal properties are incorporated into the HMM framework. HSMMs have a more appropriate state duration distribution: the state transition probability depends on the amount of time that has elapsed in the current state, whereas in HMMs the state transition probability is constant. As the Markov assumption no longer holds, HSMMs need to be trained using a modified Baum-Welch algorithm [166].

Although with parametric distributions the need for large amounts of training data is reduced, the problem of excessive computational cost still remains. The loss of the simplifying Markov assumption seriously degrades the efficiency of model re-estimation and decoding algorithms [154]. To incorporate duration modelling in a computationally efficient way, Juang et al. [105] suggested a post-processor approach in which candidate word matches output by the Viterbi algorithm with unreasonable durations are eliminated. Burshtein [26] incorporated parametric state and word duration modelling into a modified Viterbi algorithm which keeps track of the duration of each state at various time. The modified algorithm has essentially the same computational requirements of the conventional Viterbi algorithm. Mitchell et al. [136] demonstrated a technique to reduce the complexity of training a semi-Markov model with explicit duration modelling.

Another state duration modelling approach is the expanded-state HMM (ESHMM) technique [165]. In this approach each state in a standard HMM is replaced with another HMM, referred to as a 'sub-HMM' in their study, whose states share the original state observation probability density. The expanded HMMs can then be re-estimated [1] using a standard Baum-Welch algorithm. The correct state duration distribution is realised as the overall duration *p.d.f.* of the sub-HMM, which is determined by its topology as well as transition probabilities. Various topologies have been examined [e.g. 165, 141, 40, 151]. It is important to note that the overall duration *p.d.f.* of a sub-HMM depends on all possible state sequences, whereas in recognition the standard Viterbi algorithm finds only the most likely state se-

[1] Usually only the transition probabilities are updated in re-estimation.

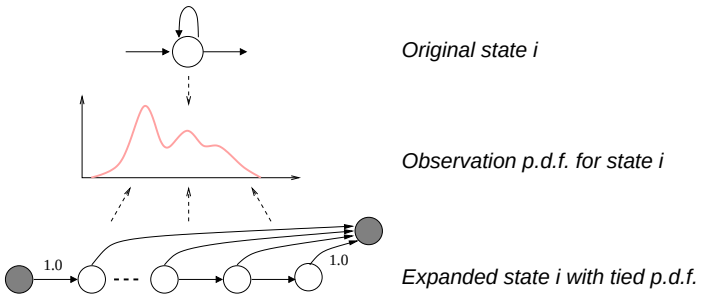

Figure 4.1: *Illustration of the expanded-state duration model. Note the sub-HMM states share the original observation probability density.*

quence. Therefore the forward-backward search algorithm is needed to give the true duration *p.d.f.*, which restricts usable topologies for the sub-HMM. An exception is to employ a special topology for the sub-HMM where the maximum duration in each state of the sub-HMM is one [165]. Fig. 4.1 illustrates such an expanded-state HMM. The sub-HMM states have no self-transitions and instead a transition to the final non-emitting state is added. With such a topology the overall duration *p.d.f.* of the sub-HMM depends on only one possible state sequence.

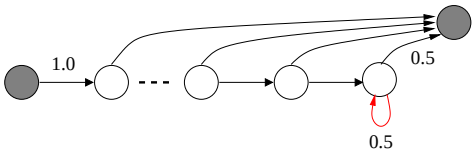

Figure 4.2: *Illustration of the expanded-state HMM topology with a self-transition in the last state, as suggested by Noll and Ney [141]. The self-transition at the last state is used to model the geometric duration distribution beyond a maximum duration.*

ESHMM is simple as duration distributions can be directly included to the HMM-based ASR framework without modification of existing decoding algorithms. However, it imposes a hard restriction on the duration range of a state. The maximum state duration is determined by the number of sub-HMM states and the minimum duration is one frame. In practice, the maximum state duration can be decided heuristically based on state duration statistics obtained during the model training process. Rabiner and Juang [154] suggested that 25 is reasonable for many word-level HMM based speech processing problems. Expanding each HMM state to a sub-HMM with such a large number of states introduces many free parameters to be

estimated and therefore requires more training data and extra computation. Noll and Ney [141] suggested adding a self-transition to the final sub-HMM state. This new topology is displayed in Fig. 4.2. By adding the self-transition the explicit restriction on the maximum state duration is removed. The geometric distribution of the final sub-HMM state is used to model the tail of the state duration distribution, which is a reasonable approximation.

In this section we first examine some factors affecting state duration statistics. The effect with various numbers of states is then investigated. The expanded-state HMM (ESHMM) technique is applied to model state duration constraints.

All experiments presented in this Chapter were performed using the missing-data ASR system based on the marginalisation techniques. Soft missing-data masks are identified by estimating local SNRs using the first 10 frames where speech is absent [13]. Instead of 16 states suggested in the Aurora 2 corpus [150], whole-word HMMs with 10 states were employed. This is because 16-state HMMs imply a minimum word duration of 160 ms (16 frames). When modelling short words such as digits, the most typical state duration can be as short as 10 ms (1 frame) and this prevents meaningful duration statistics. Each state was modelled by 10 Gaussian components instead of 3 components as suggested to compensate the correlation across dimensions of the spectral feature used.

4.2.2 State Duration Statistics

To investigate the empirical state duration distribution, histograms of duration counts were first obtained for each HMM state. The state durations were obtained by forced-aligning a set of well-trained HMMs with the Aurora 2 training data (clean speech) using a standard Viterbi decoder [2]. Around 2500 duration instances per state for each digit in the vocabulary were collected to compute the histograms with a bin width of 10 ms. The dashed lines in Fig. 4.3 show state duration distributions of digit 'seven' obtained from the histograms.

It has been reported that the empirical state duration distribution has a normal-like shape [114]. Surprisingly, in Fig. 4.3 the first 7 states have a geometric distribution and only the last 3 states have a bell-like shape. The average occupation time in each state increases toward the last state – less than 30 ms in the first few states while over 150 ms in the

[2]The word error rates on clean speech in this task are lower than 1%.

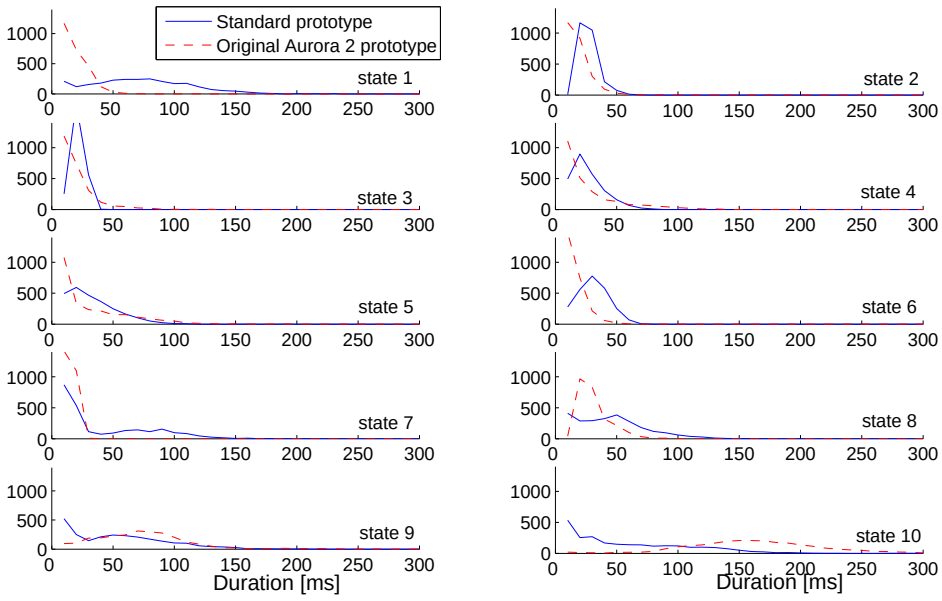

Figure 4.3: *Empirical state duration distributions of the digit 'seven' obtained from forced-alignments using 10-state HMMs. 'Standard prototype' (solid line) refers to models trained with a normal prototype; 'Original Aurora 2 prototype' (dashed line) indicates to HMMs trained with the prototype in the original Aurora 2 disc, in which the self-transition probability of the last state is set to 0.9.*

last state. This suggests that in the process of forced-alignment the decoder had a tendency to quickly move toward the last few states when discovering the most optimal state sequence. Histograms for other word models demonstrate a similar pattern. The occupation duration in the last state is far too long.

This happened because of the prototype file included in the Aurora 2 corpus. A prototype defines the structure of an HMM and also gives the initial values of all transition probabilities (normally all set to 0.5). However, in the Aurora 2 prototype the initial self-transition probability of the last state is set to 0.9 and the other self-transition probabilities are set to 0.6. Fig. 4.4 shows this prototype where white circles represent emitting states and grey circles represent non-emitting states. Due to the strange initial transition probabilities, in each training iteration more observations would be assigned to the last state than other states. As a result, Gaussian mixtures of the last state were trained to match a larger portion of a digit than any other states.

To validate this point a separate set of HMMs were trained with exactly the same setup

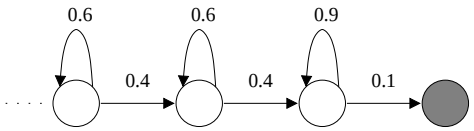

Figure 4.4: *Illustration of the original HMM prototype file supplied in Aurora 2 database discs. Note the initial self-transition probability is set to 0.9 in the last emitting state.*

except for a standard prototype in which all the initial transition probabilities were set to 0.5. The models were employed to perform the same forced-aligning experiment. The computed histograms are shown as solid lines in Fig. 4.3. The 'standard prototype' state duration distributions look more reasonable. The distributions for the middle states are no longer geometric and the average duration of the last state is much shorter.

The prototype problem also has an impact on recognition performance. The two set of HMMs were separately employed by the same missing-data recognition system to decode the Aurora 2 test set A. Fig. 4.5 shows that the models trained with the 'standard prototype' gave consistently better performance across various noise conditions and signal-to-noise ratios (SNRs). Among the recognition word errors, there were more insertion errors produced with the 'wrong prototype' HMMs compared to those with a 'standard prototype'. The insertion errors were primarily due to the fact that the 'wrong prototype' HMM set was trained with a tendency to match the most observations with the last few states. This is analogous to having fewer states. Therefore during decoding the optimal state path will quickly move toward the last few states and it is more likely to jump out of the current word model (i.e. a shorter word duration).

All the experiments discussed from now on employed the 'standard prototype'.

4.2.3 Modelling State Durations using ESHMMs

To examine if duration modelling at the state-level can bring significant improvements in missing-data speech recognition, in this section we model state duration distribution using the expanded-state HMM (ESHMM) technique [165, 40, 151]. Whole-word HMMs (with the standard prototype) were trained using clean Aurora 2 training data. Each digit was modelled by 10 states with 10 Gaussians per state. These HMMs will be employed in the

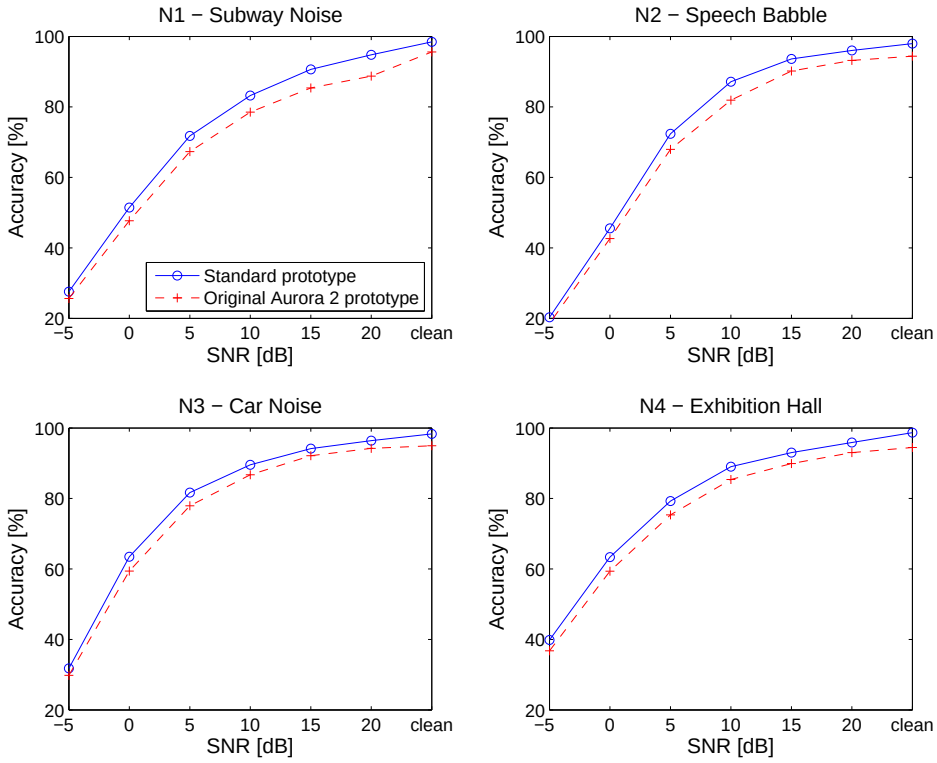

Figure 4.5: *Recognition results for Aurora 2 Test Set A comparing the models trained with different prototypes. 'Standard prototype' refers to HMMs trained with a standard prototype. 'Wrong prototype' indicates HMMs trained with the prototype in Aurora 2 corpus discs.*

baseline system.

ESHMMs were obtained by replacing each original emitting state by an 8-state sub-HMM (see Section 4.2.1) with a topology shown in Fig. 4.2, where only the last state in the sub-HMM has a self-transition. All the states in a sub-HMM share the same Gaussian mixtures of the original HMM state which the sub-HMM replaces. All the transition probabilities of the sub-HMMs were initially set to 0.5, and then updated using the standard Baum-Welch algorithm.

In the ESHMM the implicit duration distribution (of an original HMM state) is no longer geometric. Let M_i be the sub-HMM for the original state i and a_i^n denote the transition probability from the n^{th} sub-state of M_i to the non-emitting state at the end. Then the

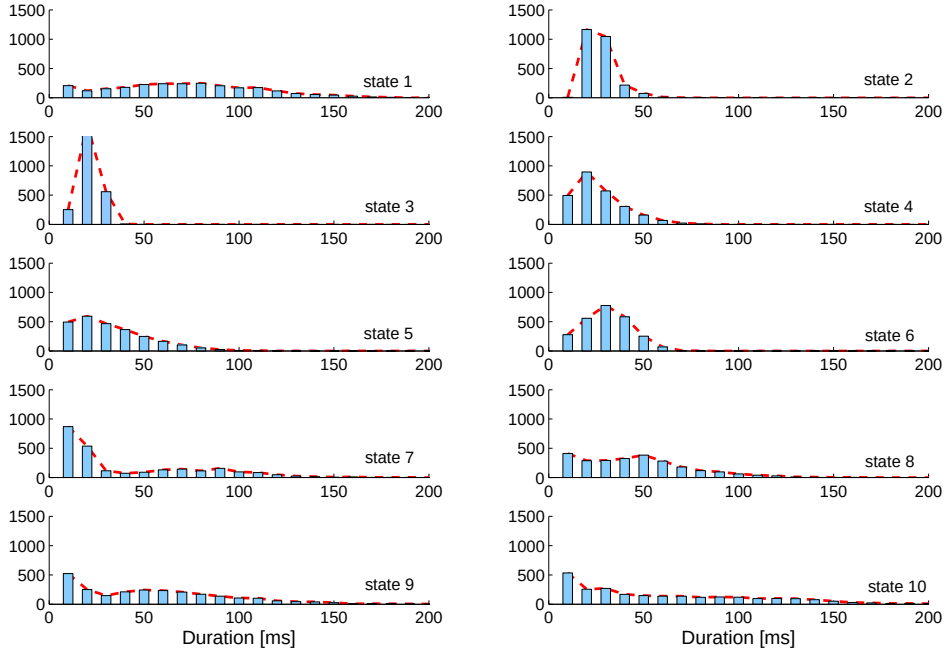

Figure 4.6: *State duration distributions of the digit 'seven' implicitly modelled by the expanded-state HMMs (dashed lines), along with the empirical distribution (histograms) determined from Viterbi forced-alignment using the original 10-state HMMs.*

duration distribution of original state i is:

$$\hat{d}_i(\tau) = \begin{cases} \prod_{1 \le n < \tau} (1 - a_i^n) \cdot a_i^\tau, & \tau \le N \\ \prod_{1 \le n < N} (1 - a_i^n) \cdot (1 - a_i^N)^{\tau - N} \cdot a_i^N, & \tau > N \end{cases} \tag{4.2}$$

where τ is state duration and N is the number of sub-states in each sub-HMM (8 in this experiment). The implicit duration distributions can be computed using Eq. 4.2 from re-trained state transition probabilities. The distributions are plotted as dashed lines in Fig. 4.6. The empirical duration histograms are also shown in Fig. 4.6. They were determined from forced-alignments produced using the original HMMs. It is clear that the duration distributions implicitly modelled by the expanded HMMs closely match the empirical distributions. Other word models demonstrated similar matches.

Two recognition systems employing the same missing-data decoder with soft SNR masks [12] were evaluated on the Aurora 2 task. The baseline system employed standard 10-state HMMs. The ESHMM system employed the expanded HMMs after re-training their transition prob-

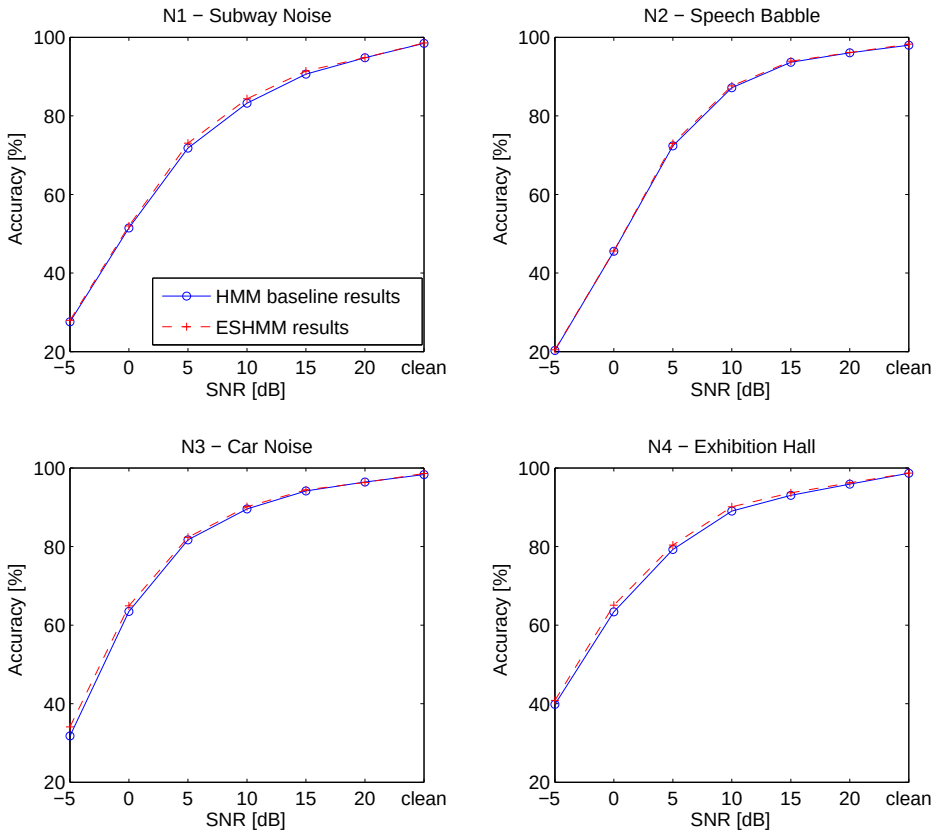

Figure 4.7: *Both recognition systems employed missing-data techniques with soft SNR masks [12] in the Aurora 2 task. In the baseline system standard HMMs were used. In the ESHMM system, each state of the standard HMMs was replaced with an 8-state sub-HMM.*

abilities. Fig. 4.7 compares their recognition accuracy rates at various noise conditions. Although the ESHMMs encode more reasonable state duration constraints, the improvements are not significant ($p > 0.2$) throughout various types of noise and SNR levels. As in ESHMMs state durations were modelled by transition probabilities, this is not surprising considering the little impact that transition probabilities normally have on overall ASR performance. The recognition errors by both systems were still subject to a large number of insertion errors [3].

Let us consider standard HMMs. Interestingly the Aurora 2 paper [150] suggested using 16-state whole-word HMMs for the connected-digit recognition task. 16 states seem to be far

[3] Various insertion penalties were tried and the best performance was reported here for both model sets.

too many for modelling many short digits in the vocabulary, e.g. 'oh'. With too many states the duration in each state will decrease which results in its distribution being geometric. However, as the HMM topology is left-to-right and no-skip, employing more states means longer minimum word durations. The 16-state whole-word HMMs have a minimum word duration of 160 ms (10 ms per frame) while the minimum word duration for 10-states HMMs is 100 ms.

The number of HMM states had an impact on recognition performance. Our preliminary experiments showed that consistent improvements over the 10-state HMM baseline were achieved by employing 16-state HMMs. This is still the case even when the total number of model parameters was the same (i.e. the 10-state HMMs employed more Gaussian mixtures). The improvements were largely due to reduced insertion errors with the 16-state HMMs which have a longer minimum word duration.

The meaning of modelling state durations seems to be obscure. It may be more useful to model whole-word durations. The next section will investigate word duration modelling.

4.3 Word Duration Modelling

Much research on duration modelling only addressed the issue at the state level and the minor performance advantage produced in recognition of clean speech often does not justify the extra complexity introduced [136]. The previous section showed that modelling state durations does not much benefit missing-data ASR. Instead, ASR improvements resulted from a longer minimum word duration in HMMs were observed. While the meaning of modelling state-level durations is obscure, modelling word-level duration constraints is potentially more effective for improving ASR in noise. The spectral representation of speech may change significantly in noisy conditions, but the duration structures of speech are, despite of the influence of the Lombard effect on speech [106], relatively insensitive to moderate noise levels [77]. Hochberg and Silverman [90] found that there is a strong correlation between recognition performance and the variance of modelled word duration.

Our goal in this section is to use word duration constraints to combat the corruption of acoustic features in noisy conditions. With the Markov state independence assumption,

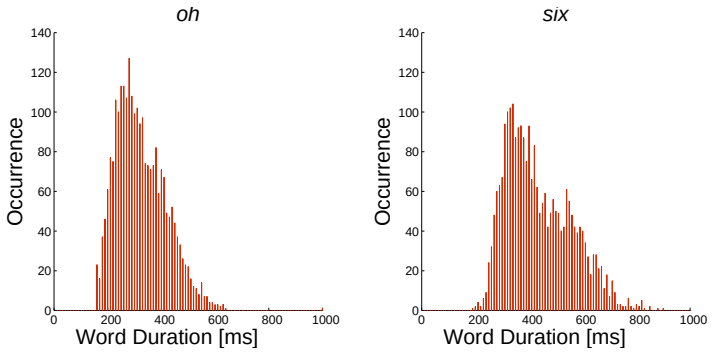

Figure 4.8: *Word duration histograms of digits 'oh' and 'six' in Aurora 2 produced by Viterbi forced-alignment.*

modelling state durations does not necessarily produce a good model of word durations [80]. In this section we first investigate some characteristics of word durations. Techniques for explicit duration modelling at the word level are then proposed. Recognition experiments using the Aurora 2 corpus are described and discussed in Section 4.3.4.

4.3.1 Word Durations Statistics and Modelling

The empirical word duration distribution was determined from an automatic Viterbi forced-alignment of the Aurora 2 training data with well-trained HMMs. Each HMM consists of 16 states with a left-to-right no-skip topology and 7 Gaussian components were employed in each state. Word durations of about 2500 instances per digit from the training data were used to compute the histograms with a bin width of 10 ms. Fig. 4.8 shows the word duration histograms for digits 'oh' and 'six' from the Aurora database. Word duration distribution does not have a normal-like shape, as Fig. 4.8 illustrates, and the histograms of both digits have a skewed shape. As word durations are themselves discrete, this makes a discrete distribution very attractive for a small vocabulary task such as Aurora 2. For a large (or even medium) vocabulary task it may become intractable to get sufficient training data for such a discrete duration model, thus a parametric model (e.g. Gaussian mixture model) may be required but can be used in a similar manner.

In this study a discrete distribution based on histograms is used to model word durations. The duration histograms are smoothed using a 5-point median filter and as an example,

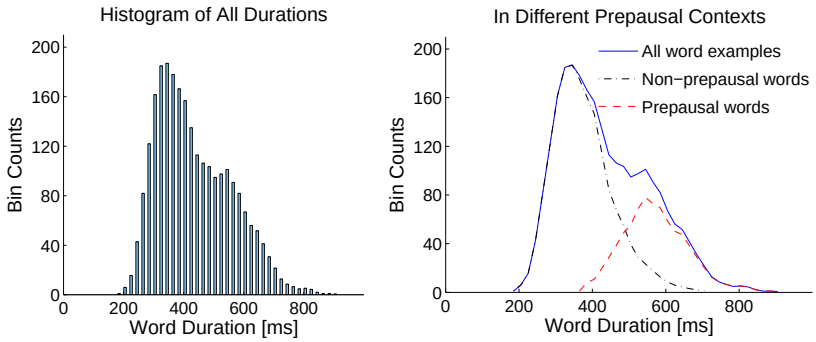

Figure 4.9: *Word duration histograms for digit 'six'. Left: histogram computed from all duration examples. Right: a comparison of duration histograms in different prepausal contexts.*

the smoothed histogram for digit 'six' is shown as solid lines on the right panel in Fig. 4.9. Let $P(d|w)$ denote the probability of word w having a duration d. To evaluate $P(d|w)$, the histograms are normalised to have unit area so that they are equivalent to probabilities. Because of the high dimensionality of the feature vectors typically used, a scaling factor is introduced to control the impact of the duration probability during decoding, forming the word duration penalty $D(d|w)$:

$$D(d|w) = P(d|w)^{\gamma} \tag{4.3}$$

where γ is the empirical scaling factor on word durations.

The smoothed duration histogram for digit 'six' shown in Fig. 4.9 has a bimodal distribution: there is one peak around 340 ms and another around 570 ms. Furthermore, the distribution is a very wide covering a duration of around 750 ms. The wide bimodal distribution is observed in the word duration histograms of many other digits. This is clear evidence that some instances of a digit in Aurora 2 database are significantly lengthened while the others are not.

Crystal and House [40] performed a set of experiments analysing segmental durations in connected-speech signals in an effort to apply duration information to the automatic analysis of speech. Among many factors that may influence segmental durations for an individual speaker, the stress patterns of a language are a primary factor. Speakers tend to lengthen syllables (or words) when stressing them. For example, in the Crystal and House experiments the mean duration of stressed vowels is found to be 70 ms greater than the average for

Table 4.1: *Mean durations (Mn.) and standard deviations (s.d.), in milliseconds, of various digits in Aurora 2 corpus. Prepausal context as indicated. N = number of cases; Mn. Inc. = relative mean duration increase in the prepausal context.*

	All examples			Non-prepausal			Prepausal				
word	N	Mn.	s.d.	N	**Mn.**	s.d.	N	**Mn.**	s.d.	**Mn.**	**Inc.**
one	2545	357	94	1784	325	80	761	432	80	33%	
two	2531	338	99	1757	302	85	774	421	75	40%	
three	2521	349	92	1769	314	75	752	433	72	38%	
four	2539	373	97	1748	338	83	791	450	79	33%	
five	2491	407	111	1698	360	83	793	509	96	41%	
six	2545	436	123	1756	374	79	789	572	87	53%	
seven	2525	426	88	1778	397	77	747	493	74	24%	
eight	2515	323	95	1752	280	64	763	420	83	50%	
nine	2492	406	99	1715	369	80	777	488	87	32%	
oh	2500	324	94	1769	291	79	731	402	80	38%	
zero	2523	448	100	1761	419	91	762	515	88	23%	

unstressed vowels. Crystal and House also discussed a strong prepausal lengthening effect on vowel durations, an effect in which vowels followed by syntactic pauses (e.g., sentence markers) are longer than the others. In a connected digits database like Aurora 2 high-level linguistic cues are minimised so the effect of lexical stress is not obvious. Our experiments have shown that the word duration statistic of a digit spoken at the beginning or in the middle of an utterance is not affected by the following digit [122]. For example, in digit strings '*one* two' and '*one* three', the digit 'one' has very similar duration statistics. In fact the bimodal word duration distribution found in this study is due to the prepausal lengthening effect.

To further examine this effect we divide the duration instances of each digit into two sets: those preceding a digit and those preceding a long pause. In Aurora 2 there is a long pause at the end of each utterance and our experiments show that the brief inter-digit pauses in some long multi-digit utterances do not give a strong prepausal lengthening effect. Therefore in this study only the sentence-final words are considered as prepausal instances. For each digit two duration histograms were computed and smoothed for the two sets. The histograms of digit 'six' are shown in the right panel of Fig. 4.9, along with the bimodal duration histogram determined from all duration instances. It is clear that the bimodal distribution consists of

the two uni-modal distributions determined from the two duration instance sets. Tab. 4.1 lists the mean and standard deviations of the duration of each digit in both the prepausal and non-prepausal contexts, in milliseconds. Prepausal instances consistently demonstrate longer durations and the relative increase of mean duration is up to 53% (digit 'six'). The duration standard deviations of each duration set are also narrower than those of all instances. The prepausal lengthening effect is observed for all digits but is less strong for multi-syllable digits (e.g. 'seven' and 'zero'). Single-syllable digits are more often shortened in continuous speech due to the effect of co-articulation than longer multi-syllable digits. Therefore the prepausal lengthening effect may be more significant. The prepausal lengthening effect has also been found in large vocabulary, spontaneous speech corpus, e.g. the Switchboard I corpus [76]. See Appendix B for examples.

To model the prepausal lengthening effect, we estimate $P(d|w,c)$ – the probability of word w having a duration d in context c. In our case $c =$ (prepausal | non-prepausal). By applying a scaling factor, we can compute the context-dependent word duration penalty:

$$D(d|w,c) = P(d|w,c)^{\gamma} \tag{4.4}$$

In connected word recognition the Viterbi algorithm is widely used to find the most probable path through a probabilistically scored time/state lattice [140]. We wish to apply the word duration penalty to word sequence hypotheses as they leave word-final states. This cannot be done directly with a standard Viterbi decoder because it does not keep a record of durations of different hypothesis paths. Competing paths may have different duration histories for the word now terminating. Fig. 4.10 shows two competing paths through template T_1 reaching the final state of the template at the same time t_3. Assume the solid line $h1$ is the best hypothesis path through T_1 from the time frame t_1 to t_3, discovered by the Viterbi algorithm. The dashed line $h2$ is another path terminating at t_3 with a less likelihood score than h_1. With a different word duration penalty hypothesis path $h2$ may in fact has a higher overall likelihood score than $h1$, but the word duration information is not available when comparing paths in the Viterbi process which only keeps the most likely path leading to each state.

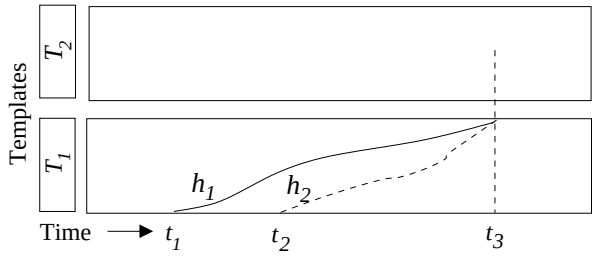

Figure 4.10: *Two competing Viterbi paths through the template T_1 reaching the template end at the same time frame t_3 but starting at different time.*

4.3.2 Duration Modelling with a Multistack Decoder

To incorporate word duration constraints a decoding algorithm based on the idea in Renals and Hochberg's *NOWAY* decoder [160] was employed. *NOWAY* is a stack decoder [149, 7] which sets up a separate stack for word sequence hypotheses that end at each time frame and processes these stacks time-synchronously from left to right. Newly created hypotheses are added to stacks and this process is continued until a complete hypothesis is determined. Since all word hypothesis paths between two stacks have the same word duration, word duration penalties can be safely applied in a multistack decoder. To keep record of word durations the items on each stack are word sequence hypotheses $H(t, W(t), P(t))$ which consist of:

1. The reference time t at which the hypothesis ends.
2. The word sequence $W(t) = w(1)w(2)\ldots w(n)$ covering the time from 1 to t.
3. Its overall likelihood $P(t)$.

The decoder extracts the most likely hypothesis from the stack based on its overall likelihood at time t, computes one-word extensions, applies word duration constraints for the word, and places all the new hypotheses into corresponding stacks. When the search finishes, the most likely hypothesis path on the last stack is the optimal path.

To make the multistack search more efficient, some heuristic pruning can be applied to reduce the computation cost. For example, when the top hypothesis of each stack is extended for one more word w, we need only consider extensions between a minimum word duration and

a maximum duration (D_{min} and D_{max}), obtained by examining word duration statistics from the training data. This word duration boundary itself seems to be able to improve the recognition performance as hypotheses with very short or very long words can be pruned out of the search. A typical duration range for a non-terminating digit in Aurora 2 is 200-700 ms. For a prepausal digit a typical duration range is 300-900 ms.

Let T denote the length of the utterance in frames and let $H(t, W^*(t), P^*(t))$ be the most likely hypothesis on the stack at time t, where $W^*(t)$ is the best word sequence finishing at time t and $P^*(t)$ is its likelihood. A Viterbi search $V(t, u, v)$ can be used to find the best-matching single words starting from a given time t and finishing at each time within a range u to v. The full algorithm uses $V(\ldots)$ as follows:

1. *Initialisation*:
   Run $V(1, D_{min}, D_{max})$ to find initial one-word matches for $t = D_{min} \ldots D_{max}$, and place initial hypotheses $H(t, W(t), P(t))$ in the stacks at $D_{min} \ldots D_{max}$.

2. *Iteration*:
   For $t = D_{min}$ to $T - D_{min}$

   (a) Select $H(t, W^*(t), P^*(t))$ after applying word duration penalties $P(d|w^*(n), w(n+1))^\gamma$ from the stack at t, where d is the duration of the best-matching word $w^*(n)$ finishing at time t, and $w(n+1)$ is the next extending word. Note, with different extending words the penalty is different.

   (b) Run $V(t, t + D_{min}, t + D_{max})$; form extended hypotheses and add them to each stack respectively.

3. *Termination*:
   Find $H(T, W^*(T), P^*(T))$ from the stack at time T and the final result $W^*(T)$.

As we keep the best word sequence in each stack, there is no need to do backtracking to find the global optimal word sequence.

This algorithm is illustrated in Fig. 4.11. The most likely word sequence hypothesis ('*Seven One*') is extended by the most probable one-word extension '*Eight*' finishing at time t_2. When the decoder continues to process the stack at time t_2, a word duration penalty $P(D =$

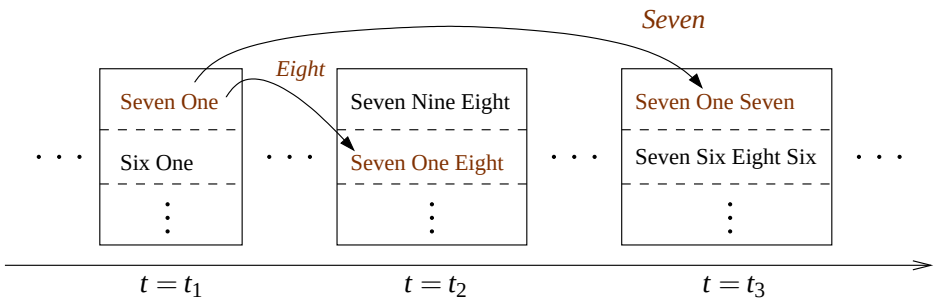

Figure 4.11: *Illustration of the multistack decoding algorithm (adapted from [160]). The stack at time t_1 is being processed. 'Eight' is the best-matching single word starting at t_1 and finishing at t_2. See text for more details.*

$t_2 - t_1 | w_1 = \text{`}eight\text{'}, w_2)$ is first applied to the likelihood score of hypothesis '*Seven One Eight*', and w_2 should be decided by the next searching word. If the search goes into an HMM for silence, the penalty will be different from that if the search goes into an HMM for a digit. Since in Aurora 2 database an individual digit has a maximum duration of 900 ms (90 frames), although the search space is increased by a factor of 90, the computational load increases by a much smaller factor because most of the calculation is in the observation probability computation which does not scale up.

4.3.3 Duration Modelling with Unrolled HMMs

Previous section reported a multistack decoding algorithm to incorporate word duration constraints. However, most HMM-based ASR systems employ a Viterbi decoder and it may not always be feasible to incorporate a stack decoder. This section proposes a more generic technique which amounts to little more than expanding the HMM topologies so that word duration penalties can be incorporated. This technique is fully compatible with existing ASR systems and is theoretically equivalent to the multistack decoding technique.

Assuming a no-skip, left-to-right HMM with N states $q_1, q_2, \ldots, q_N$ for word w is being expanded. We can compute corresponding duration penalties using Eq. 4.3 with an expected duration range d^w_{min} to d^w_{max}. Each emitting state q_i is then duplicated d^w_{max} times, the duplicates sharing the same Gaussian parameters with q_i. The duplicated states q_i form a sequence and the self-transition of each state is replaced by a one-way transition between two

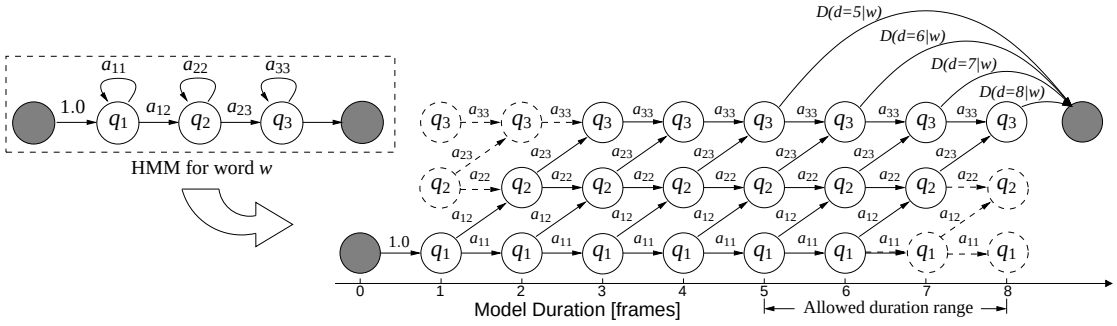

Figure 4.12: *Unrolling a standard no-skip, left-to-right HMM with word duration penalties.*

adjacent states. For N states in the old HMM we get N state sequences. Except for the last state sequence q_N, the j^{th} state in the sequence q_i is connected to the $(j+1)^{th}$ state in the next state sequence q_{i+1}, with the transition probability the same as that from q_i to q_{i+1} in the old HMM. The beginning non-emitting state is connected to the first state in the first state sequence q_1 with a transition probability of 1.0 as an entry point. Each state in the last sequence is connected to the non-emitting state at the end as one of the terminating points in the expanded model. When word sequence hypotheses leave a model from any of the final states q_N, different paths within the model to the leaving state are guaranteed to have the same duration. Therefore the duration penalty can be safely applied in the expanded HMM. We use the corresponding duration penalty to replace the transition probability from each state in the last sequence q_N to the terminating non-emitting state.

Fig. 4.12 illustrates this procedure using an example of a 3-state no-skip, left-right HMM with 2 non-emitting states (dark circles) at the two ends. The states that will not be visited are marked with dashed circles. To make the expanded HMMs more efficient, we do not supply the terminating transition to the states before the $(d^w_{min})^{th}$ state in the last state sequence. In this example the allowed duration range for word w is 5 to 8 frames, therefore there are no such transition for the first four q_3 states. The allowed duration ranges are determined by examining the duration statistics obtained from the training data. A typical duration range for a non-prepausal digit in the Aurora corpus is 200 – 700 ms and for a prepausal digit a typical duration range is 300 – 900 ms. With a 10 ms frame shift although the state space is expanded roughly by a factor of 90, the computational load increases only by a much smaller factor in a small vocabulary task. Most of the computation is for the observation

probabilities, which remain constant because the Gaussian mixtures are tied up.

For word duration modelling in the prepausal context, we expand two sets of HMMs using Eq. 4.4: *NPPdigit* – HMMs for non-prepausal digits; and *PPdigit* – HMMs for prepausal digits. The decoder then employs the 2 HMM sets with a grammar in an EBNF format: (sil {$NPPdigit} ($PPdigit) sil).

4.3.4 Results and Discussions

Gender-dependent word-level HMMs were trained on the Aurora clean speech training set. Digit models ('1'–'9', 'oh' and 'zero') consist of 16 no-skip, left-right states with observations modelled by 7-component diagonal GMMs. A 3-state silence model was used to model the long pauses before and after an utterance and an additional 1-state silence model was used to model the brief inter-digit pauses that may occur during long digit strings.

The *NPP-WD* system represents the recognition system with pause-context-free word duration penalties calculated according to Eq. 4.3. The penalties can be employed using either the proposed multistack decoder or the unrolled HMMs. As the two techniques are theoretically equivalent, their recognition results are the same. *PP-WD* represents the ASR system employing duration penalties according to Eq. 4.4. The scaling factor γ was set to 10 for all noise conditions in this study, which was tuned based on a small set of developing data. The baseline system was a strong 'missing data' recognition system described in [12], which employed combined masks estimated from harmonicity and SNR-based cues.

To reduce the number of digit insertions, a grammar was used in the baseline system to constrain all hypotheses to start and end with the silence model. To further guard against too many insertion errors, an empirical word insertion penalty was introduced for all missing data recognition experiments [4].

Fig. 4.13 shows the word error rates (WERs) of the four different noise types in Aurora 2 test set A at various SNR levels. The word error rate of various systems in the 'subway' noise condition is shown in Table 4.2, together with the number of substitutions (S), deletions (D) and insertions (I), respectively. At low SNR levels both systems with duration models clearly

[4] A word insertion penalty of -25 in negative-logarithm domain was used and this was optimised to give the best performance for the Aurora 2 task.

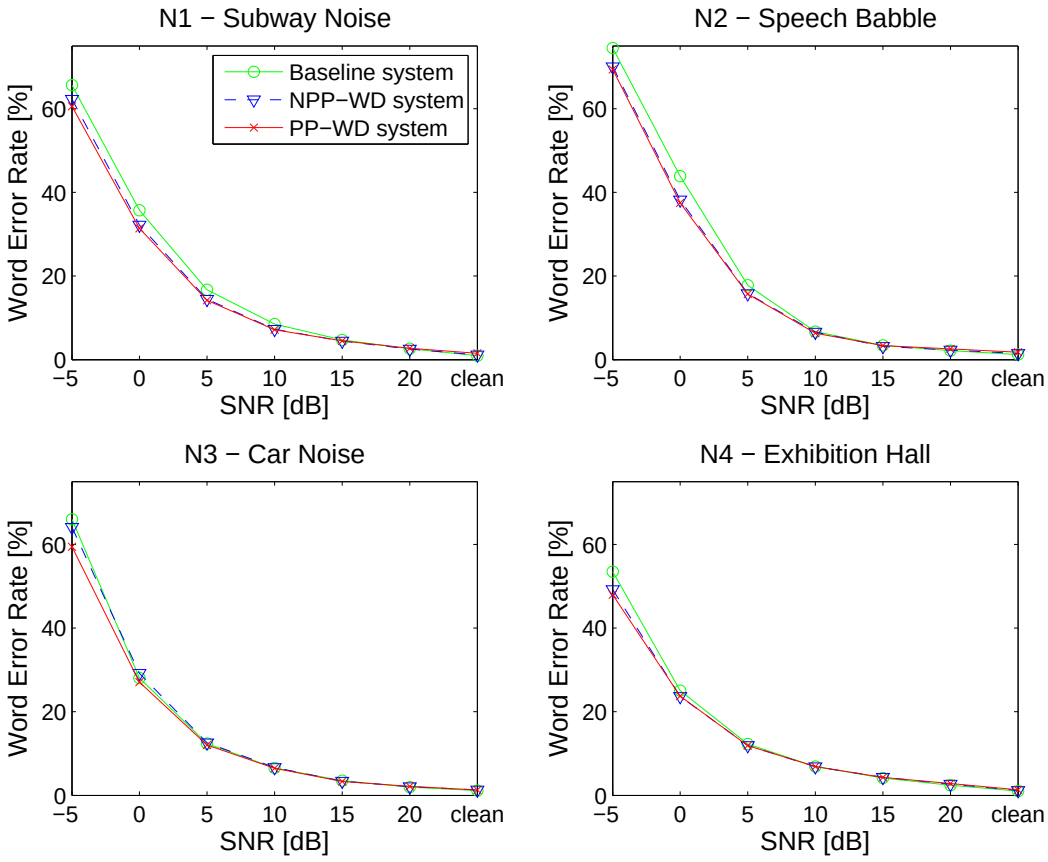

Figure 4.13: *Word error rates for test set A in Aurora 2 at various SNR levels.*

outperformed the baseline system, which is a strong missing-data ASR system [12]. The average relative improvements over the baseline system are shown in Tab. 4.3. Significance testing (the matched-pairs test [74]) showed that at SNRs lower than 10 dB the improvements are statistically significant ($p < 0.05$). No significant difference was found in WERs at SNRs above 10 dB ($p > 0.1$).

It should be noted that the baseline system was an optimised missing-data ASR system for the Aurora 2 task, which performed well in competitive evaluations [12]. It was capable of achieving high recognition accuracy in quiet conditions. When SNRs are low, estimation of missing-data masks was difficult and more data was corrupted. Duration constraints are more important in these conditions. This is analogous to increasing the contribution of the language model when the acoustic model is poor.

Table 4.2: *Word error rate (%)of various systems in the 'subway' noise condition. The respective numbers of substitutions (S), deletions (D) and insertions (I) are shown in ().*

	Baseline	NPP-WD	PP-WD
20 dB	4.6 (S=85 D=41 I=24)	2.5 (S=51 D=23 I=8)	2.7 (S=46 D=25 I=18)
15 dB	6.6 (S=115 D=53 I=48)	4.5 (S=80 D=34 I=31)	4.5 (S=74 D=32 I=40)
10 dB	13.0 (S=209 D=79 I=134)	7.3 (S=128 D=50 I=59)	7.2 (S=116 D=50 I=67)
5 dB	23.8 (S=407 D=162 I=207)	14.5 (S=258 D=118 I=95)	14.2 (S=245 D=115 I=102)
0 dB	44.8 (S=777 D=431 I=250)	32.2 (S=586 D=316 I=146)	31.4 (S=582 D=289 I=151)
-5 dB	70.2 (S=984 D=1045 I=257)	62.2 (S=965 D=867 I=195)	60.6 (S=1069 D=700 I=203)

Table 4.3: *Average relative improvements with duration modelling over the baseline in word error rates [%].*

System	SNR (dB)						
	-5	0	5	10	15	20	clean
NPP-WD	5.4	7.1	7.4	4.5	2.5	0.0	0.0
PP-WD	8.6	9.8	9.0	6.4	1.7	0.0	0.0

The PP-WD system achieves the lowest WERs, but the results of the NPP-WD system are close. This is because without considering the pause context the 'prepausal' portion of the duration distribution is still well modelled by the histogram-based model. The performance gain using the prepausal context is mainly due to the emphasis of the prepausal duration distribution and a better estimate of the allowed duration ranges.

To examine the impact of the word duration constraints on the recogniser, we also compared the duration statistics produced during the decoding process. Word duration examples were collected in the back-tracing stage of various recognition systems. Histograms of these duration examples are then computed and compared to those obtained by forced-aligning the training data. Fig. 4.14 shows these duration histograms for digit 'six' at the SNR level of 0 dB. Panel (A) is the duration histogram obtained from forced-aligning the clean Aurora training data (same as in the left panel of Fig. 4.9). Panels (B–D) show the histograms produced by the baseline recogniser, the NPP-WD system and the PP-WD system, respectively. When decoding noisy speech, the baseline recogniser generates many word matches with too short or too long durations and fails to demonstrate the second peak in the distribution around 572 ms. The proposed word duration model forces the recogniser to focus on word matches with more realistic durations and with the prepausal context it produces a duration distribution

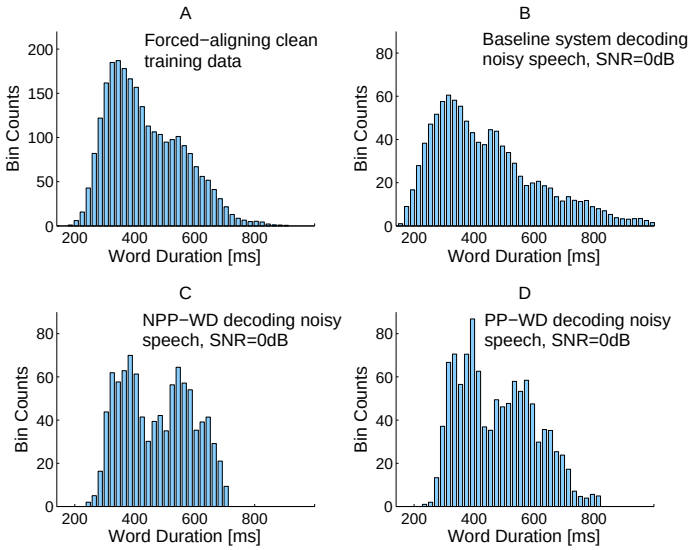

Figure 4.14: *Comparisons of word duration statistics produced by various ASR systems for digit 'six'.*

more similar to that from training data.

Note that the 16-state no-skip, left-right topology imposed a minimum word duration of 160 ms on any word matches produced by the decoder. This boundary can help ASR as word hypotheses with a unrealistic duration shorter than 160 ms will be ruled out. Ma and Green [122] also reported that lower WERs can be achieved by only applying constraints on minimum/maximum word duration (i.e. a uniform duration distribution was applied). In adverse acoustic environments these hard duration boundaries can be effective in reducing WERs for ASR.

4.4 Summary

This chapter presented several experiments that together examine the duration information at both state-level and word-level in an HMM-based missing data decoding framework. The state duration distribution in a conventional HMM is examined and the expanded-state HMM approach was employed to model state duration. Although experiments show that this method is capable of capturing the state duration characteristics, little improvement was achieved. Experiments suggest that it is more effective to model duration constraints at word-level. Word durations are relatively insensitive to moderate noise levels. We present

two methods to explicitly employ word duration constraints in different pause contexts. Experiments show that the techniques are able to offer significantly lower WERs over a strong missing-data baseline system in noisy situations.

Although the duration modelling techniques presented in this chapter are developed for missing-data ASR, they can be applied to any system that employs word-level HMMs. Especially with the unrolled-HMM technique, a standard Viterbi decoder can be used. However, whole-word HMMs are limited to small vocabulary tasks. Large vocabulary speech recognition tasks typically employ phone-level HMMs. Although similar techniques can be applied to model phone durations instead of word durations, it is certain that more complex models need to be considered in order to get any improvement in accuracy. Gadde [69] proposed to form a duration feature vector for each word composed of durations of the phone sequence in the word. The word duration features were modelled by using Gaussian mixture models (GMMs). These word duration GMMs were employed during recognition to re-score recognition hypotheses in an N-best list. Evaluated on a large vocabulary recognition task Gadde [68] reported significant reduction in word error rates over their baseline system.

4.4.1 Further Development

The duration statistics used to build duration models were obtained from clean speech recorded in a quiet environment. Therefore the duration modelling techniques proposed assume that word durations remain constant in various noise conditions. This is in general unrealistic as speakers tend to make effort to have better articulation in noisy conditions, i.e. the Lombard effect on speech [106]. This may cause different loudness, rhythms and therefore phone durations from those in a quiet condition. In a realistic situation, the duration model should be adapted based on feedback from on a speaking rate detector. The proposed techniques, however, provide a method to incorporate a duration model if one is available.

Work is also underway to model the prepausal lengthening effect on conversational speech tasks. Gadde attempted to model the effect in their large vocabulary systems, but the improvement was limited. Conversational speech such as the Switchboard corpus also demonstrates a strong lengthening effect on durations of phones preceding a pause. Some examples demonstrating the prepausal lengthening effect from the Switchboard can be found in Ap-

pendix B [5]. Currently the possibility of using dynamic Bayesian networks [102] to model the prepausal lengthening effect is being investigated. Strategies include employing a pause detector which influences the selection of a state transition matrix. When a pause in the near future is detected, the state transition matrix which favours self-transitions is selected. This will result in a preference for word hypotheses with longer durations. By far no significant difference in WERs has been observed, partially because the transition probabilities often have little impact on recognition results. Phones in prepausal words often have not only longer durations but also stretched spectral shapes across time. Future work will address these points.

[5] Listening examples are available at http://www.dcs.shef.ac.uk/~ning/research/prepausal/.

Chapter 5
A 'Speechiness' Measure to Improve SFD

The performance of a speech fragment decoding (SFD) system relies on two factors. First, the speech recognition models need to be fairly detailed as they are also employed to select speech fragments. Errors may occur if the top-down information in speech models is insufficient. Secondly, the fragments need to be fairly coherent (i.e. each fragment should contain only energy from the same source). The quality of fragments is not a trivial issue. This chapter will investigate the fragment selection issue. Some of the work has previously appeared in [123]. The next chapter will investigate the fragment generation problem.

5.1 Introduction

The strength of SFD is that it is designed to operate without strong assumptions about the interfering noise. Unlike techniques such as HMM decomposition [182] which needs a precise noise model, SFD employs only top-down information in models of the target source (i.e. the speech recognition models). This is partially motivated by the observation that perception of a particular sound in an acoustic mixture benefits far more from listeners' familiarity with the sound than from their familiarity with the noise. Dowling [57] demonstrated this point in perceptual experiments with overlapping melodies. Listeners found it easier to segregate a tone from a set of distracting tones if the foreground melody was familiar, but familiarity with the background melody did not reduce the interfering effect. Cooke et al. [37] also demonstrated that listeners with better language-specific knowledge will perceive the language more easily (see 2.4.5 for details). These observations suggest that although a CASA-based ASR system requires detailed models of the target source, it perhaps does not

need detailed models of the background.

5.1.1 Are Speech Models Sufficient for SFD?

The minimum assumption SFD makes about the background is well-motivated and important as it clearly limits the search space of plausible models. However, there are two additional factors to be considered.

First, although detailed noise models may not be necessary and are difficult to estimate, there is often some knowledge about the noise available (or at least the difference between speech and noise). This happens in daily listening environments. For example, when hearing approaching footsteps it may be difficult to identify the coming person, but clear that the sound is from footsteps and not speech. This suggests that listeners may form 'simple' models of various sounds and these models may help them to confirm the selection of speech evidence.

Secondly, SFD uses top-down information in speech models to recruit correct speech fragments but these constraints may be insufficient. For example, Barker et al. [14] reported that SFD often failed to choose enough speech fragments and therefore produced poor recognition results. This is confirmed by control experiments described in Section 5.4. In the current implementation SFD considers each fragment as being part of either the speech foreground or the noise background with equal probability. This essentially gives an equal prior probability to all the segregation hypotheses defined by the term $P(S|Y)$ in Eq. 3.3:

$$\hat{W}, \hat{S} = \arg\max_{W,S} P(W|S,Y)P(S|Y)$$

The uniform distribution is a very crude approximation. Some fragments may 'look' more like speech and others may not. Therefore segregation hypotheses that include more speech-like fragments should be given more weight. Basing the choice of fragments purely on speech models means these models have to be fairly detailed (i.e. having very 'peaky' distributions) so that noise fragments can be reliably rejected. One can build speaker-dependent models. This, however, is not a general solution and does not address the fundamental issue. Each segregation hypothesis is matched against the recognition models using missing-data techniques. SFD compares different segregation hypotheses by looking at their matching scores. As the matching process is based on 'missing data', a subset of speech fragments may match some (incorrect) speech models better than that the full set of speech fragments matches

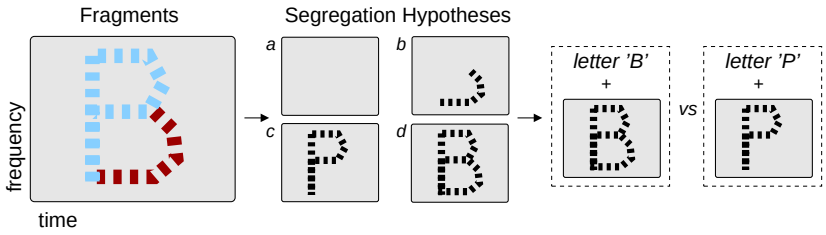

Figure 5.1: *Illustration of the fragment selection problem in the SFD system. Two fragments (shown using the shaded regions) are being decoded here. Both fragments are speech fragments and together correspond to the speech evidence of letter 'B'.*

correct speech models. As a result, only partial speech evidence is selected and incorrect words are reported by the decoder.

The problem of selecting speech fragments is illustrated in Fig. 5.1 where two fragments are being decoded and shown using shaded regions. Assume both fragments are true speech fragments and together they correspond to the acoustic evidence of letter 'B'. Also assume the fragment in light grey by itself matches the acoustics of letter 'P' [1]. There are four various speech/background segregation hypotheses to be considered by SFD, shown as (a–d) in the figure. In theory, the recognition model of letter 'B' should produce the highest likelihood score with the segmentation hypothesis (d) as the most likely segregation, i.e. both fragments are selected as part of the speech foreground. However, since the fragments represent only partial acoustic evidence, it is possible that the model of letter 'P' produces a higher score with the segregation hypothesis (c) than that of the model 'B' with the segregation hypothesis (d). Therefore the letter could be recognised incorrectly as 'P'.

5.1.2 Introducing 'Speechiness'

Extra top-down information can be exploited to assist the decoder in the choice of fragments. In this chapter we propose a 'speechiness' measure for each fragment – a value in $[0, 1]$ expressing a degree of confidence that the fragment is part of the speech foreground. The measure can be used to steer the decoder towards preferring reliable speech evidence in adverse conditions.

[1] The illustration of the process, in which the fragments are drawn in the shape of letters 'B' and 'P', uses a visual analogy.

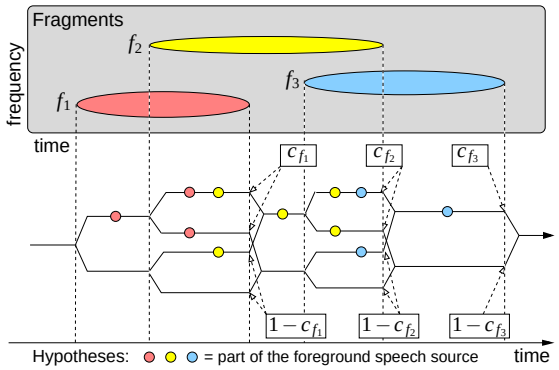

Figure 5.2: *Evolution of parallel segregation hypotheses with speechiness c_f being applied (adapted from Fig. 3.3). The dots indicate which ongoing fragments are being treated as a part of the speech foreground.*

In Section 5.2 we first describe techniques to incorporate speechiness measures into the SFD framework. Section 5.3 describes the noise materials and the experiment setup used in this study. Section 5.4 presents control experiments which demonstrate the problem of fragment selection in SFD. In Section 5.5 a modulation filtering technique is proposed as a speechiness measure. Section 5.6 concludes with future research directions.

5.2 Fragment Decoding with Speechiness

Speechiness can be incorporated into the decoding process. As described in Section 3.3, the search process can be efficiently implemented as illustrated in Fig. 5.2 where three fragments (shown using the shaded regions) are being decoded. Each time a new fragment starts, all ongoing segregation hypotheses are split so that in each pair one hypothesis labels the fragment as speech while the other assigns it to the background. When a fragment finishes, pairs of hypotheses are merged if their labelling only differs with regard to the fragment. Fig. 5.2 includes an additional term – the speechiness, c_f before pairs of hypotheses are merged. If the speechiness, c_f, of the finishing fragment, f, can be estimated, then in each pair c_f is added to the log-probability of the hypothesis that labels the fragment f as speech and $1 - c_f$ to that of the other one.

The speechiness values range in $[0, 1]$. Values 1 or 0 respectively represent that the fragment is definitely part of the speech foreground or the background. Value 0.5 gives equal weight to

either hypothesis. Because of the high dimensionality of the feature vectors typically used, a scaling factor is needed to control the impact of c_f. A natural candidate is the fragment size, s_f, i.e. the number of T/F pixels included in the fragment f, since the observation probability calculation involves the same amount of data. Therefore in log-domain $s_f \cdot \log(c_f)$ and $s_f \cdot \log(1 - c_f)$ are applied to each hypothesis pair.

Incorporating speechiness is an approach to approximating the segregation model $P(S|Y)$ in the SFD equation (Eq. 3.3). In a segregation hypothesis each fragment is independently determined to be part of the foreground. This is a crude assumption, but as we will see in Section 5.4, it has proved very effective in practice. Let $\mathcal{F}_S$ denote the subset of fragments labelled as the speech foreground in the segregation hypothesis, S. Then $P(S|Y)$ can be approximated as:

$$P(S|Y) = \prod_{f \in \mathcal{F}_S} c_f \prod_{f \notin \mathcal{F}_S} 1 - c_f \tag{5.1}$$

Eq. 5.1 ensures the same numbers of terms in the calculation for every segregation hypothesis (i.e. the total number of fragments).

5.2.1 Speechiness Measure

The basic approach to the speechiness measure is to use a set of features extracted from *fragments* to estimate a speech model, M_s, and a background model, M_b. The background model can be trained using various non-speech sounds. Given the feature x_f extracted from an unknown fragment f, its speechiness c_f can be derived from the posterior probability of the speech model M_s (e.g. using a sigmoid function):

$$c_f \leftarrow P(M_s|x_f) \tag{5.2}$$

If we assume $P(M_s|x_f) + P(M_b|x_f) = 1$ and the priors $P(M_s) = P(M_b)$, then using Bayes' rule Eq. 5.2 becomes:

$$c_f \leftarrow \frac{P(x_f|M_s)P(M_s)}{P(x_f)} \tag{5.3}$$

$$= \frac{P(x_f|M_s)}{P(x_f|M_s) + P(x_f|M_b)} \tag{5.4}$$

In this case estimating speechiness is similar to solving a speech/non-speech classification task, on which there is a substantial literature. For example, Scheirer and Slaney [168]

proposed using 13 different features to discriminate speech from music. However, the features used to estimate speechiness should be extracted from the time/frequency (T/F) components included in a fragment rather than the full-band.

If the background model is not available, speechiness can be measured by converting values derived from the fragment features to scores in the range of $[0, 1]$. These features, however, should be designed to reflect the characteristics of speech fragments so that noise fragments will score low. The preference is for measures based on the difference between speech and noise as the requirement of a noise background model is often difficult to meet.

5.3 Experimental Setup

5.3.1 Speech and Noise Materials

Experiments were performed using a corpus of speech artificially mixed with various types of noise at a range of signal-to-noise ratios (SNRs). 600 end-pointed utterances were randomly drawn from the Grid corpus [36] as the test set. Another 300 utterances were drawn as the development set. All the mixtures are single-channel signals. Utterances were normalised to have equal root-mean-square (rms) energy and sampled at 25 kHz.

Six types of noise with various characteristics were selected: *violins*, *piano*, *singing voice*, *drums*, *speech babble* and *factory noise*. Their details can be found in Appendix C. All noise signals were resampled to 25 kHz. The length of each noise signal is around 30 seconds. Noisy mixtures were produced by artificially adding each of the noise signals to the 600 test utterances at a range of global SNRs: -6 to 6 dB with an interval of 3 dB. The starting point in the noise signals was randomly set each time. All the mixtures were single-channel signals.

5.3.2 Fragment Generation

For each mixture a set of *oracle* fragments was generated by making use of *a priori* knowledge of pre-mixed signals. The process of oracle fragment generation is illustrated in Fig. 5.3. First, cochleagrams of the pre-mixed signals were compared against each other pixel-by-pixel to produce two spectro-temporal binary masks for the two sources, respectively. Each binary mask labelled the regions where the energy of one source exceeds that of the other

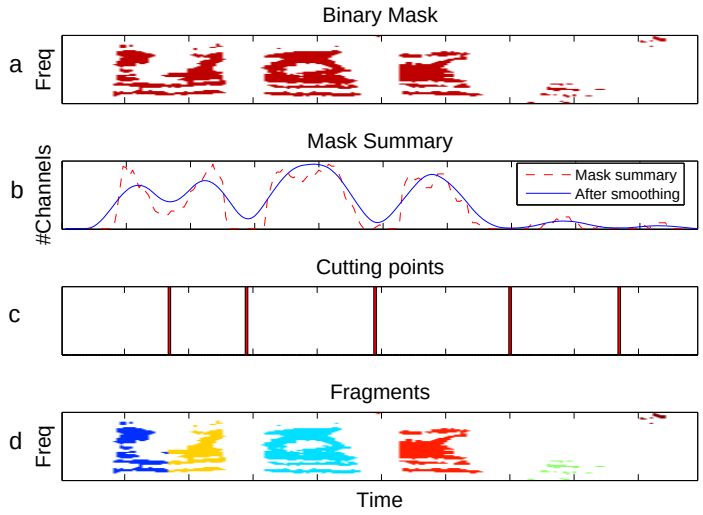

Figure 5.3: *Illustration of oracle fragment generation for a single source. See text for details.*

more than 1 dB and the mixed energy exceeds a background threshold (2.0 was used for the log-compressed cochleagram).

As an example, Fig. 5.3a shows the mask for speech in the speech/babble mixture. Each mask was then summed across frequency (dashed line in Fig. 5.3b) and smoothed by convolving the summary with a sinusoid. The smoothed summary is shown in Fig. 5.3b as a solid line. Local minima were selected as cutting points (Fig. 5.3c) to dissect a binary mask into a set of fragments (Fig. 5.3d) for the single source. Fragments generated for both sources were combined together and each fragment was given a unique label. Examples of the oracle fragment generated in this way for various speech/noise mixtures are shown in Fig. C.2.

The oracle fragments provide a reasonable match to genuine fragments generated using CASA models. Common CASA grouping cues include harmonicity and onset synchrony. Since most speech energy concentrates at harmonics and formants, which are emphasised in the cochleagram, genuine fragments will have significant harmonic energy (see Chapter 6) and clear onset boundaries. These properties are well reflected in the process of oracle fragment generation. Using such oracle fragments allows us to study the fragment selection problem in isolation from the fragment generation problem, which will be examined in Chapter 6.

5.3.3 Speech Recogniser Setup

The speech recognition task was to identify the letters and digits in the Grid utterances. Speaker-independent word-level HMMs were trained using 500 clean utterances per speaker (different from the test set). Each word was modelled using two states per phoneme with 16 diagonal-covariance Gaussian mixture components per state and a left-to-right no-skip topology. The number of HMM states for each word was decided based on 2 states per phoneme. They were trained using 500 utterances from each of the 34 speakers. The speech fragment decoding (SFD) system employed spectral features (the cochleagram) which were supplemented with their temporal derivatives to form 128-dimensional feature vectors.

5.4 Control Experiments

Control experiments were first performed to investigate the decoder's ability to recruit fragments. The fragment decoder is forced to recruit a fragment as part of the foreground if it is given speechiness 1 and forced to reject a fragment if it is given speechiness 0. Speechiness 0.5 indicates that the fragment will be considered part of either the speech foreground or the noise background with equal probability.

Oracle fragments were employed by the SFD system in four experimental setups:

- (a) 'SFD baseline' – normal SFD with no speechiness.
- (b) 'correct segment' – SFD with access to the correct segmentation so that only true speech fragments were employed. This is equivalent to missing-data decoding with a mask formed using only the true speech fragments.
- (c) 'speechiness 1/0.5' – forcing SFD to recruit all the true speech fragments (given speechiness 1) but treat noise fragments with equal probability (given speechiness 0.5), i.e. all the hypotheses in which true speech fragments had been labelled as background were pruned.
- (d) 'speechiness 0.5/0' – forcing SFD to reject all the true noise fragments (given speechiness 0) but treat speech fragments with equal probability (given speechiness 0.5), i.e.

all the hypotheses in which true noise fragments had been labelled as foreground were pruned.

Experiments (a–b) examined the performance difference due to fragment selection errors and (b) provided the best possible SFD performance. Control experiments (c–d) investigated the decoder's ability to select fragments. The fragment identities were revealed using prior knowledge so that controlled speechiness could be assigned.

5.4.1 Results and Discussions

Fig. 5.4 shows the keyword recognition accuracy using oracle fragments in various noise conditions (SNR = 0dB). Comparing results (a) and (b), the SFD baseline system (a) produced significantly lower accuracy than the correct segmentation results (b). The recognition errors were mainly due to the fact that top-down information in SFD were insufficient to recruit correct fragments. A similar observation was also reported in [14].

The performance difference occurred at various SNR levels, as shown in Fig. 5.10. The gap was particularly large for noise that masked significant speech energy at low SNRs, such as the 'speech babble' and 'factory' noise. In these conditions the unmasked glimpses of speech [31] were very small (i.e. the frequency channels included in the speech fragments were much fewer than those in the noise fragments). Therefore the likelihood scores based on the small speech glimpses may not be high enough to justify correct fragment selection. In fact, in these conditions most time noise fragments were selected. Therefore the SFD baseline (curves (a) in Fig. 5.10) reported recognition rates at the chance level. However, results using correct segmentations (curves (b) in Fig. 5.10) show that if the speech fragments can be identified, recognition accuracy was reasonable high even at low SNRs.

It can be seen in Fig. 5.4 that the performance gap was largely due to the failure to recruit enough speech fragments, rather than the failure to assign enough noise fragments to the background. In experiment (c) where the fragment decoder was forced to employ all the true speech fragments, most noise fragments were correctly assigned to the background, which leads to recognition accuracies close to the best possible performance (result b). However, when all the true noise fragments were assigned to the background in experiment (d), the decoder failed to select enough speech evidence. The noise corruption caused a strong mis-

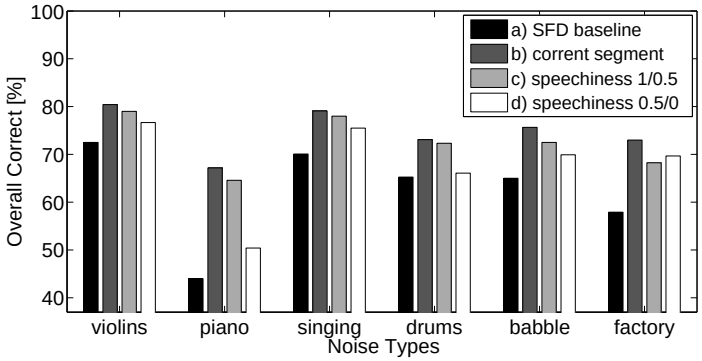

Figure 5.4: *Keyword recognition results of SFD with controlled speechiness using oracle fragments. SNR = 0 dB.*

match between speech acoustics and recognition models and therefore it is more likely for SFD to assign some speech fragments to the background in experiment (d).

Speaker-Dependent Modelling

The 34 speakers in the Grid corpus were modelled using speaker-independent HMMs with 16 Gaussian mixtures per state. With these models SFD has been shown to be incapable of recruiting enough speech fragments. One question one may ask is whether these speech models are detailed enough to make a good choice of fragments. We therefore increased the number of mixtures per state to 32 but very similar results were produced.

To further investigate this question, 34 sets of speaker-dependent (SD) models were employed. Each set of SD models was trained using 500 Grid utterances spoken by a single speaker. Seven Gaussian mixtures were employed in each HMM state. The keyword recognition experiments were then performed again. For each test utterance, the speaker identity was also revealed to the decoder so that the corresponding HMM set could be used. This is an unrealistic situation, but it allows us to examine the fragment selection problem in full. The keyword recognition accuracies of the four experimental setups are shown in Fig. 5.5.

Compared to the speaker-independent results the performance of the baseline SFD is significantly improved. More importantly, the performance gap between the baseline setup and that with the correct segmentation is reduced. This suggests that more detailed speech models can increase the decoder's ability to recruit correct fragments. However, the performance

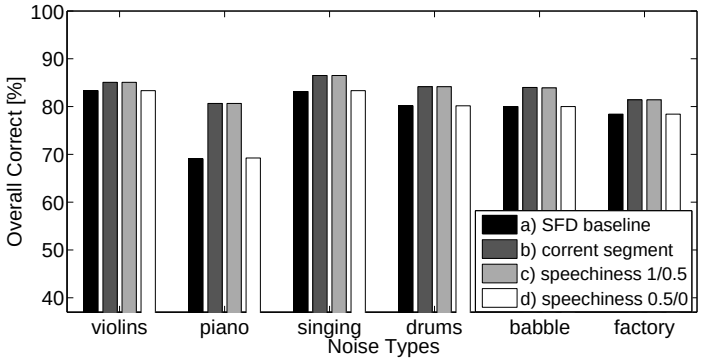

Figure 5.5: *Recognition results using Speaker-Dependent HMMs, with controlled speechiness. SNR = 0 dB.*

gap is still significant in many noise conditions. Experiments with controlled speechiness also demonstrated a similar behaviour to that using speaker-independent models – SFD tends to favour hypotheses in which more fragments are labelled as background. The true reason for this behaviour is perhaps due to the fact that the top-down information in the speech models was insufficient. As discussed in Section 5.1.1, hypotheses in which only a subset of speech fragments are labelled as foreground may produce a higher likelihood score than hypotheses which label all the speech fragments as foreground. The key point is that there is no extra mechanism to stop the decoder from rejecting true speech fragments. The speechiness measure therefore is introduced to help combat this problem by steering the decoder toward selecting more fragments that are likely to have originated from the speech source.

5.5 Speechiness by Modulation Filtering

5.5.1 Introduction

Speech has characteristic low-frequency amplitude modulations which are associated with the syllabic rate of around 4 Hz [93]. Perceptual experiments have shown that modulations at rates above 16 Hz are not required for human speech perception and significant intelligibility remains even if modulations at rates of 6 Hz and above are removed [94, 60]. Inspired by these studies Kingsbury et al. [109] proposed a modulation filtering technique that can be used to derive a speech representation which emphasises this low-frequency temporal structure – the 'modulation filtered spectrogram' – for automatic speech recognition in adverse conditions. Their experiments have demonstrated that when combined with other ASR features such as log-RASTA-PLP [87], the modulation filtered spectrogram can give WER reductions in certain noisy and reverberant conditions.

Palomäki et al. [143] also reported a similar modulation filtering technique to identify regions dominated by direct speech rather than reverberation. Reverberation usually has clean beginnings with more noisy reverberation tail to follow. Palomäki et al. therefore employed a filter with a shape of the impulse response such that resulting filtered signal captures the onsets of strong speech modulations. The technique resulted in a 'reverberation mask' which indicates regions dominated by direct speech energy, with which missing data ASR can be applied. Significant recognition improvement has been reported in reverberant conditions over ASR systems which use acoustic features derived from perceptual linear prediction (PLP) and the modulation filtered spectrogram.

5.5.2 Fragment Modulation Energy

In this study the modulation filtering technique is employed in a different way. A modulation filter is applied to each frequency band and periods where significant energy gets through are identified. The energy levels are averaged over pixels in a fragment to judge its speechiness.

To design a filter that reflects the speaking rate in the Grid corpus, 17000 utterances from the training set were forced-aligned using a set of well-trained models. Fig. 5.6a shows the word duration histogram calculated from the forced-alignments. It is clear that the word

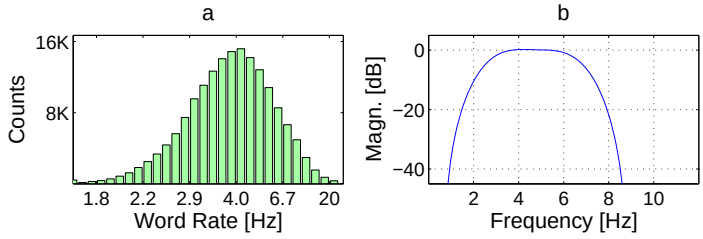

Figure 5.6: *(a) Histogram showing the speaking rate in the Grid utterances. (b) Frequency response of the modulation filter.*

rate has a peak around 4 Hz. Since most words in the Grid corpus are single-syllable, this is approximately the syllabic rate.

We therefore designed a bandpass modulation filter, $h(t)$, with a pass-band of 2.5–6.67 Hz to emphasise the 4 Hz energy. The finite impulse response (FIR) filter was designed using the frequency sampling method. Fig. 5.6b shows its frequency response. The spectral energy in each frequency band (see Section 5.3.2) was filtered with $h(t)$. Following a similar technique used by Kingsbury et al. [109], the peak level over all bands was set to 0 dB and levels more than 40 dB below the peak were set to -40 dB. The process produces a representation similar to the modulation filtered spectrogram [109]. An example for a mixture of speech and speech babble is shown in Fig. 5.7d.

The modulation energy for each fragment was calculated by averaging energy levels of the modulation filtered spectrogram over T/F components that were included in the fragment. Fig. 5.8 plots average modulation energy of fragments against their sizes. The fragments shown here were oracle fragments generated from the 300 mixtures in the development set (see Section 5.3.1) mixed at a SNR of 0 dB.

It is clear that true speech fragments (shown as the red dots) had significantly higher modulation energy than noise fragments (the yellow dots). The mean level across all the speech fragments was around -20 dB while the mean level for noise fragments was below -30 dB. This distribution pattern was fairly stable across various noise conditions. Noise fragments with a slow temporal structure, such as the piano music, had particularly low modulation energy, while noises with a rhythm closer to the speech syllabic rate range, such as the drums, had fragments with higher modulation energy. Although larger fragments demonstrate more

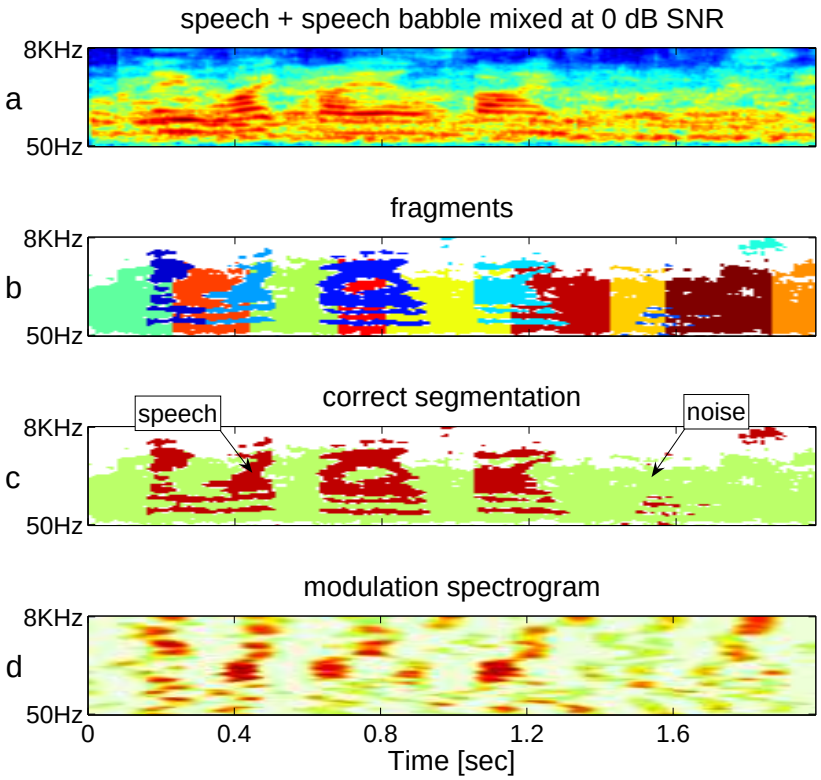

Figure 5.7: *modulation filtered spectrogram of the speech and speech babble mixture at 0 dB SNR.*

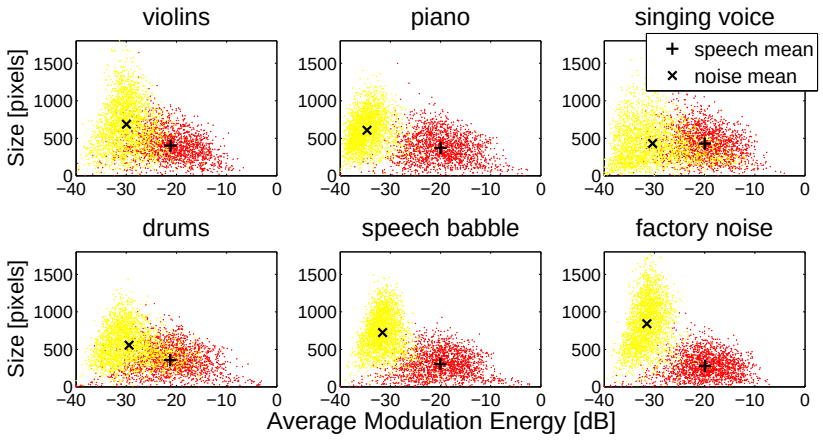

Figure 5.8: *Modulation energy of oracle fragments, SNR = 0 dB. The red dots represents speech fragments and the yellow dots represent noise fragments.*

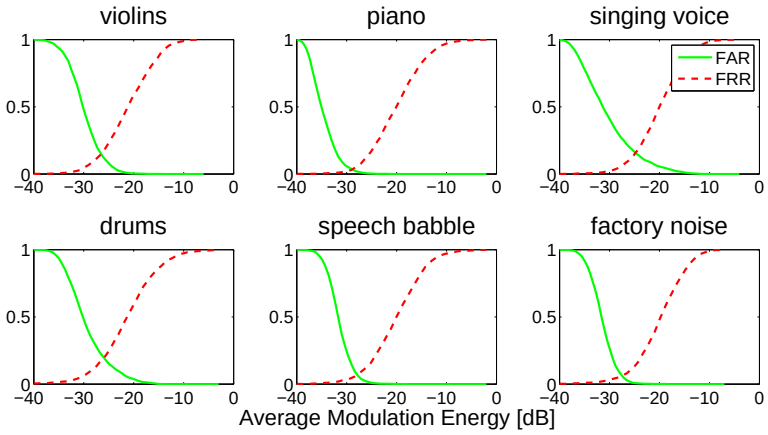

Figure 5.9: *False acceptance rates (FARs) and false rejection rates (FRRs) of fragment classification based on the modulation energy, SNR = 0dB.*

consistent modulation energy (i.e. smaller variance), there is no strong correlation between modulation energy and fragment size.

The speechiness c_f of fragment f was judged by compressing its modulation energy e_f with a sigmoid function:

$$c_f = \frac{1}{1 + \exp(-\alpha(e_f - \beta))} \tag{5.5}$$

where $\alpha = 0.02$ is the sigmoid slope and $\beta = -27$ is the sigmoid centre. The parameters were derived via a classification experiment using the development dataset and fixed for all the noise conditions. Equal error rate (EER) analysis was performed at 0 dB SNR. Fig. 5.9 shows false acceptance rates (FARs) and false rejection rates (FRRs) of the fragment classification. EERs occur around -27 dB across various noise conditions.

5.5.3 Experiments and Discussions

SFD was run with speechiness measured using the modulation filtered spectrogram. In this experiment no prior knowledge of fragment identities was used. The measured speechiness scores were applied to each unknown fragment during decoding. Fig. 5.10 shows the keyword recognition results of SFD employing speechiness estimated using the modulation filtering technique (circles), along with the SFD baseline results (crosses) and those with correct

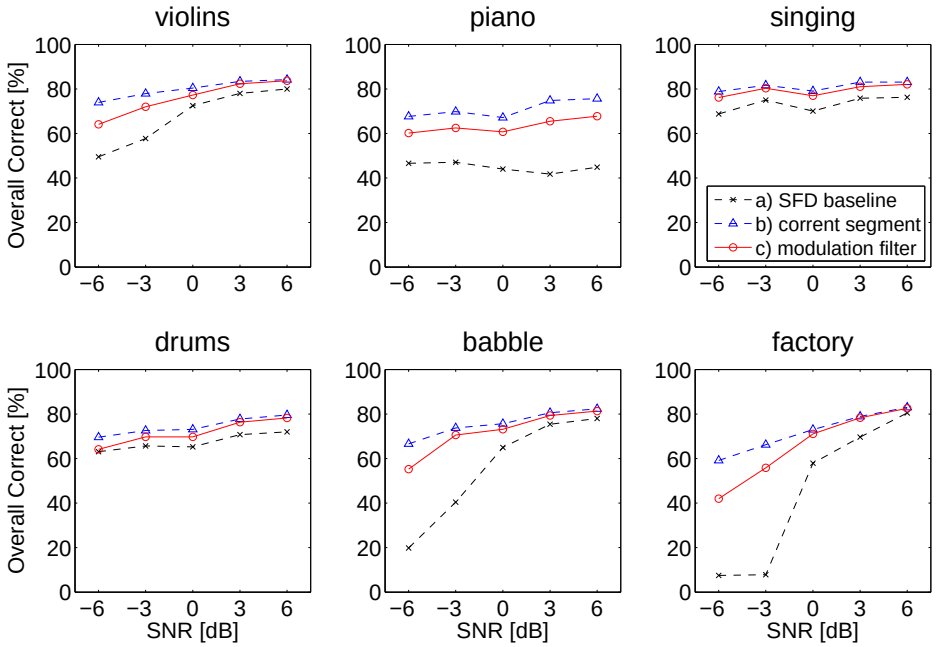

Figure 5.10: *Keyword recognition results of SFD with speechiness measured by using the modulation filtering technique.*

segmentation (triangles) [2].

When employing the estimated speechiness SFD produced recognition accuracies close to the best possible performance (i.e. SFD with correct segmentation, see Section 5.4 for more detail), which was consistent throughout various SNR levels and noise conditions. Although the performance gap became larger when the SNR was lower, the current recognition results suggest that the speechiness measure is relatively insensitive to noise levels and noise types. The larger performance gap at low SNR levels were due to less reliable speechiness measures. In low SNR conditions speech energy was less dominated and therefore the mixtures would have less modulation energy around 4 Hz. The current parameters used to estimate speechiness values were fixed based on EER analysis at 0 dB SNR. Ideally they should be optimised for each SNR condition. This may not be feasible, however, without prior information of SNR levels.

We are particularly interested in the 0 dB SNR because in this condition the level difference

[2]See Section 5.4.1 for discussions on comparisons between the baseline and the 'correct segment' systems.

cue that can be exploited by SFD to recruit fragments is minimised. At this SNR level, improvements in recognition accuracy were large in noise conditions (such as 'piano') that have a temporal structure significantly different from that of speech. In these conditions the speechiness measure was more reliable. Note the speechiness measure does not have to be 100% accurate as it is presented as a degree of confidence in $[0, 1]$. Inappropriate values can still be offset if the acoustic evidence strongly matches recognition models.

5.6 Summary

Speech fragment decoding provides a general framework to couple source segregation and recognition. However, recognition experiments demonstrate that the top-down information in speech models is insufficient to select enough speech evidence and extra constraints are necessary. The SFD implementation by Barker et al. [15] bases the selection of fragments purely on speech recognition models. However, there is plenty of other information about the fragments that is not represented in these models. For example, two fragments from different directions may both match speech models well but should not be assigned to the same source.

In this study we integrated a speechiness measure into the SFD framework to bias the decoder towards selecting fragments that are more likely to be part of the speech source. The proposed technique provides a way of exploiting extra top-down constraints in SFD. A modulation filtering technique which emphasises the characteristic low-frequency modulation energy of speech has been proved an effective speechiness measure. Recognition experiments show that the speechiness measure can help the decoder employ more reliable speech evidence and therefore produce significant improvement in accuracy.

5.6.1 Further Development

The modulation filtering technique appears to be a good way of measuring speechiness, but it has several limitations. First, these measures emphasising 4-Hz modulations are not temporally very precise, therefore they may not be very informative over short time periods such as a 10 ms frame commonly employed in speech processing. Although the current system overcomes this by averaging modulation energy over all the T/F components included in a fragment, these measures are less reliable for small fragments. Second, the speechiness

measuring technique will be less effective if the background noise has a rhythm similar to the syllabic rate. Therefore, in order to produce robust speechiness estimates, various properties that may distinguish speech from noise, such as pitch dynamics and location cues, should be combined together.

The current system does not employ noise models when estimating speechiness. There is evidence that listeners quickly form simple noise models [21] in speech perception. Employing noise models (although not necessarily as detailed as the speech HMMs) will bring more reliable speechiness estimates as speech fragments can be 'pushed away' from noise models in classification. This is crucial if present noise sources have many similarities to speech.

The speechiness measure assumes that the target source is the only speech source and therefore will not apply to conditions such as simultaneous speech. However, in a broader view the framework can be seen a way to employ knowledge about the target. Each fragment can have a weight representing a confidence of being part of the *target foreground*. Therefore the technique could be generalised to represent properties that the target source possesses. For example, in a simultaneous speech recognition task we can employ known properties of the target speaker to measure how likely it is that a fragment may have originated from the speaker.

Future work will also include investigating methods of applying speechiness measures to each segregation hypothesis which may include multiple fragments. The current method assumes fragments are independently determined to be foreground or background. This assumption provides an effective and efficient implementation for the segregation model. However, fragment independence may not always occur because some fragments do not 'look' very like speech individually, but will do when combined with other fragments. For example, a fragment dominated by energy of an unvoiced consonant may be very like noise on its own. In the current implementation hypotheses in which the fragment is labelled as speech foreground would be given a low score. However, if such a fragment is followed by a fragment in which energy matches that of a vowel, they may together form a complete syllable.

Employing oracle fragments allows investigation of the fragment selection problem in isolation from the fragment generation problem. The next chapter will focus on techniques that can be used to construct genuine fragments. Speaker-dependent models will be employed, which

have been demonstrated to significantly reduce the fragment selection problem.

Chapter 6
Improved Fragment Generation for SFD

The performance of SFD depends on the quality of fragments. If fragments contain too much energy that belongs to different sources, the recognition accuracy will be poor. The number of fragments also influences the computation cost. Having too many fragments means there are more segregation hypotheses to be considered. The fragment generation technique employed by Barker et al. [15] is very simple and crude. Improved fragment generation techniques will result in significant performance gains. This chapter will present fragment generation techniques using pitch cues derived from structure in the correlogram. Some of the work reported in this chapter has previously appeared in [124, 126, 17].

6.1 Introduction

6.1.1 The Correlogram

One important representation of auditory temporal activity that combine both spectral and temporal information is the autocorrelogram (ACG). The autocorrelogram, or simply correlogram, is a visual display of sound periodicity. It is normally defined as a three-dimensional volumetric function, mapping a frequency channel of an auditory periphery model, temporal autocorrelation delay (or lag), and time to the amount of periodic energy in that channel at that delay and time. The periodicity of sound is well represented in the correlogram. If the original sound contains a signal that is approximately periodic, such as voiced speech, then each frequency channel excited by that signal will have a high similarity to itself delayed by the period of repetition. Primarily because it is well-suited to detecting signal periodicity,

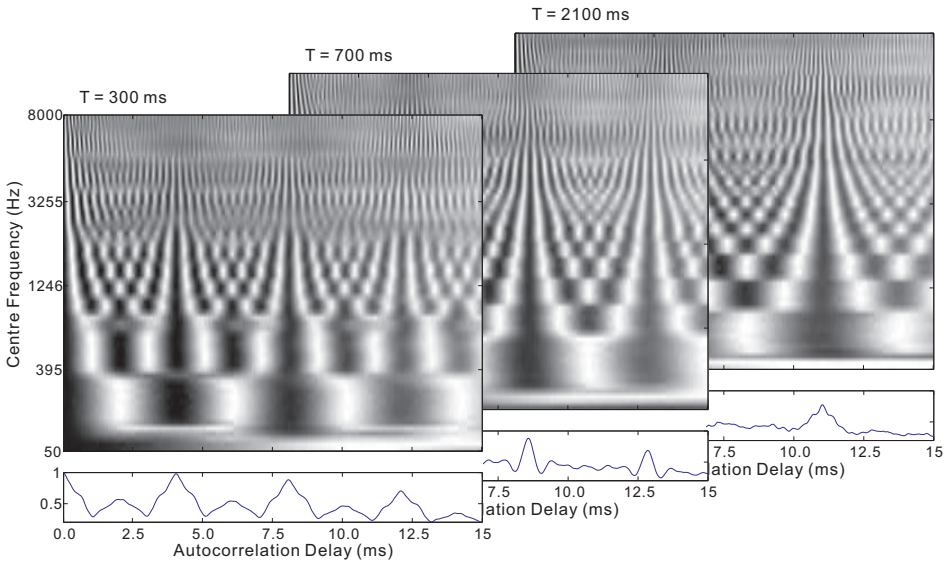

Figure 6.1: *Three correlograms of a clean speech signal uttered by a female speaker, produced for time frame 300, 700 and 2100 ms, respectively. Each correlogram has been normalised and plotted as an image. A corresponding summary ACG is shown at the bottom of each correlogram.*

the ACG model is widely considered as the preferred computational representation of early sound processing in the auditory system.

Correlograms are normally sampled across time to produce a series of two-dimensional graphs, in which frequency and autocorrelation delay are displayed on orthogonal axes. Fig. 6.1 shows three correlograms of clean speech spoken by a female speaker. Each correlogram has been normalised and plotted as an image for illustration. All the ACG frequency channels respond to the fundamental frequency ($F0$) and this can be emphasised by summing the ACG over all frequency channels, producing a 'summary ACG' (the bottom panels in Fig. 6.1). The position of the largest peak in the summary ACG corresponds to the pitch of the periodic sound source.

6.1.2 Correlogram-Based CASA Models

The correlogram was first suggested as a model for pitch perception by Licklider [117] in his neural auto-coincidence model, where the concept of sub-band periodicity detection was discussed. The model was then reintroduced by Slaney and Lyon [173] and others (e.g. [5, 131]),

as a computational approach to pitch detection. Slaney and Lyon employed the correlogram computed from the output of a cochlear model to model how humans perceive pitch. The pitch was estimated based on locating the peaks in the summary correlogram. The ACG model has subsequently been extended as a popular mechanism for segregating concurrent periodic sounds and the primary methods have been based on inspection of the summary correlogram. Assmann and Summerfield [5] reported a place-time model on a concurrent vowel segregation task. The model estimated the pitch of each vowel as corresponding to the autocorrelation delays with the two largest peaks in the summary correlogram. Meddis and Hewitt [132] proposed a residual-driven approach. They first selected the largest peak in the summary ACG, the delay of which corresponds to the $F0$ of the stronger sound source. Frequency channels that respond to this $F0$ were grouped and removed from the correlogram. The rest of the channels were integrated together and the largest peak in the residue corresponds to the $F0$ of a second (and weaker) source. More recently, neural oscillator models have been successful at providing accounts of the interaction of cue combinations, such as common onset and proximity [23, 186], in which the summary correlogram model was also employed as a front end.

One limitation of the methods which are based on the summary correlogram is that when speech is corrupted by competing sounds, locating peaks in the summary is often difficult. The position of the largest peak in the summary would not always correspond to the pitch of the target speech and peaks indicating pitches of different sound sources may be correlated. Another limitation is that these models cannot account for the effect of harmonic components of the weaker source being dominated by the stronger source, where all correlogram channels will be assigned to the stronger source [51]. To address these limitations, Coy and Barker [39] proposed to keep the four largest peaks in the summary ACG as pitch candidates for each time frame and then employed a multi-pitch tracker to form smooth pitch tracks from these candidates. Frequency channels that respond to pitch values in the same pitch track are grouped together. By keeping multiple pitch candidates they show that better sound segregation can be achieved. However, their system relies on a robust multipitch tracker and keeping an arbitrary number of pitch candidates is not effective when dealing with different competing sources.

The summary ACG is not the only method to reveal pitch information. The methods based

on the summary ACG discard the rich representation of the spectral content and time structure of a sound in the original correlogram. Visually there are clear pitch-related 'dendritic structures' in the correlogram. The 'dendrites' are tree-like structures whose stems are centred on the delay of multiple pitch periods across frequency channels. Slaney and Lyon [173] discussed this dendritic structure in their perceptual pitch detector. They convolved the correlogram with an operator to emphasise the structure before integrating all ACG channels together. Summerfield et al. [178] also proposed a convolution-based strategy for the separation of concurrent synthesised vowels with F0 not harmonically related in the correlogram. By locating the dendritic structure in the correlogram they demonstrated that multiple fundamentals can be recognised.

6.1.3 Organisations of the Chapter

In this chapter we are concerned with the use of primitive CASA models to address the problem of separating and recognising speech in monaural acoustic mixtures. The dendritic correlogram structure was exploited to separate a spectrogram representation of the acoustic mixture into 'fragments' – spectro-temporal regions such that the acoustic evidence in each region is likely to have originated from a single source of sound (see Section 3.1). Some of these fragments will arise from the target speech source while others may arise from noise sources. These coherent fragments are passed to the speech fragment decoder to identify the best subset of fragments as well as the word sequence that best matches the target speech models. We evaluate the system using a challenging simultaneous speech recognition task [1].

The remainder of this chapter is organised as follows: in Section 6.2 the overall structure of our system is briefly reviewed. Section 6.3 describes the techniques used to integrate spectral components in each frame based on the ACG. Section 6.4 presents methods for sequential integration which produces harmonic fragments. Techniques to produce fragments for inharmonic regions are discussed in Section 6.5. Section 6.6 introduces a confidence map to soften the discrete decision of assigning a pixel to a fragment. In Section 6.7 we evaluate the system and discuss the experimental results. The results will show the SFD is lack of a mechanism to attend to words known *a priori* to be spoken by the target speaker, which results in poor recognition results around 0 dB SNR. Section 6.8 presents a fragment-

[1] http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

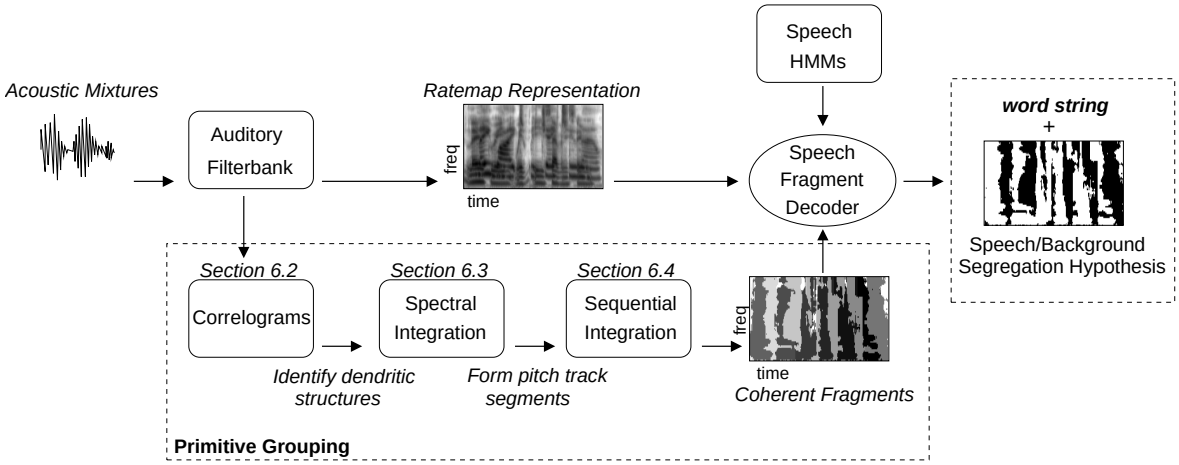

Figure 6.2: *Schematic diagram of the proposed fragment generation system.*

based speaker identification technique that can help SFD combat this problem. Section 6.9 concludes and presents future research directions.

6.2 System Overview

Fig. 6.2 shows the schematic diagram of our system. The input to the system is a mixture of target speech and interfering sounds, sampled at a rate of 25 kHz. In the first stage of the system, cochlear frequency analysis is simulated by a bank of 64 overlapping bandpass gammatone filters, with centre frequencies spaced uniformly on the equivalent rectangular bandwidth (ERB) scale [75] between 50 Hz and 8000 Hz. The output of each filter is then half-wave rectified.

Spectral features are then computed in order to employ the speech fragment decoder [15]. The instantaneous Hilbert envelope is computed at the output of each Gammatone filter. This is smoothed by a first-order low-pass filter with an 8 ms time constant, sampled at 10 ms intervals, and finally log-compressed to give an approximation to the auditory nerve firing rate – the 'ratemap' representation [23].

The output of the auditory filterbank is also used to generate the correlograms. A running short-time autocorrelation is computed on the output of each cochlear filter, using a 30 ms Hann window. At a given time step t, the autocorrelation $A(i, t, \tau)$ for channel i with a time

lag τ is given by

$$A(i,\tau,t) = \sum_{k=0}^{K-1} g(i,t+k)w(k)g(i,t+k-\tau)w(k-\tau) \tag{6.1}$$

where g is the output of the Gammatone filterbank and w is a local Hann window of width K time steps. Here $K = 750$ corresponding to a window width of 30 ms. The autocorrelation can be implemented using the efficient fast Fourier transform (FFT), but has the disadvantage that longer autocorrelation delays have attenuated correlation owing to the narrowing of the effective window. We therefore use a scaled form of Eq. 6.1 with a normalisation factor to compensate for the effect:

$$A(i,\tau,t) = \frac{1}{K-\tau} \sum_{k=0}^{K-1} g(i,t+k)w(k)g(i,t+k-\tau)w(k-\tau) \tag{6.2}$$

The autocorrelation delay τ is computed from 0 to $L-1$ samples, where $L = 375$ corresponding a maximum delay of 15 ms. This is appropriate for the current study, since the $F0$ of voiced speech in our test set does not fall below 66.7 Hz. We compute the correlograms with the same frame shift as when computing the ratemap features (10 ms), hence each one-dimensional (frequency) ratemap frame has a corresponding two-dimensional (frequency and autocorrelation delay) correlogram frame.

In the stage of spectral integration the dendritic structure is exploited in the correlogram domain to segregate each frame of the mixture into spectral groups, such that the partial spectra in each group is entirely due to a single sound source in that frame. In the next stage local pitch estimates are computed for each group and a multipitch tracker links these pitch estimates to produce smooth pitch tracks. Spectral groups are integrated temporally based on these pitch tracks. The processes separate the spectro-temporal representation of the acoustic mixture into a set of coherent fragments, which are then employed in the speech fragment decoder, together with clean speech models, to perform automatic speech recognition.

6.3 Spectral Integration

Spectral integration involves organising time/frequency (T/F) components of acoustic mixtures across frequency. Typically this is performed in each time frame.

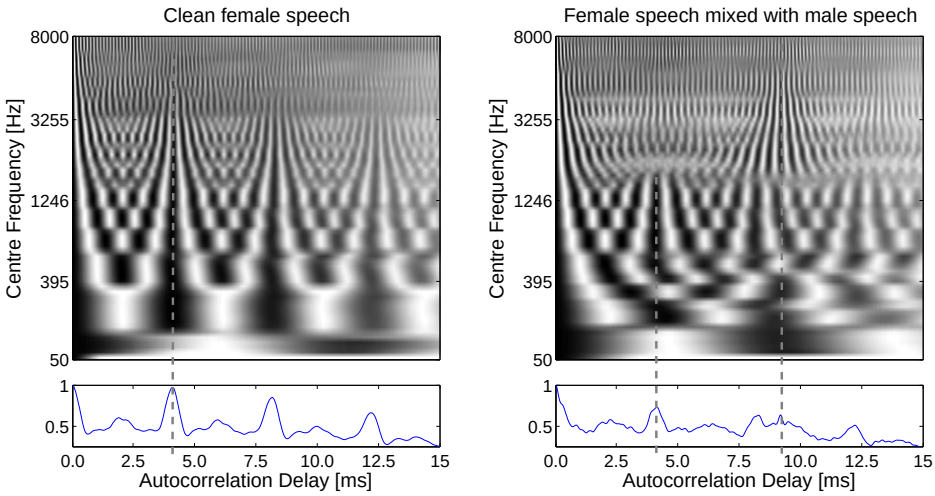

Figure 6.3: *A comparison of correlograms in clean and noisy conditions. Left – a correlogram and its summary of clean female speech, taken at time frame 60; right – taken at the same frame when the female speech is mixed with male speech at a target-to-masker ratio (TMR) of 0 dB. The dendritic structures that correspond to the F0 of different speech sources are marked using vertical dashed lines. It can be clearly seen that in the noisy condition the dendrites do not extend across the entire frequency range.*

6.3.1 The Dendritic ACG Structure

For a periodic sound source all autocorrelation channels respond to $F0$ (i.e. the energy reaches a peak at the same frequency), forming vertical stems in the correlogram centred on the delays corresponding to multiple pitch periods. Meanwhile, because each filter channel also actively responds to the harmonic component that is closest to its centre frequency (CF), the filtered signal in each channel tends to repeat itself at an interval of approximately 1/CF, giving a succession of peaks at approximately the frequency of the CF of each channel in the correlogram. This produces symmetric tree-like structures appearing at intervals of the pitch period in the correlogram (dendritic structures). When only one harmonic source is present, the stem of each dendritic structure extends across the entire frequency range (see the left panel in Fig. 6.3). The one with the shortest autocorrelation delay is located at the position of the pitch period of the sound source. When a competing sound source is also present, some ACG channels may be dominated by the energy that has arisen from the competing source, causing a gap in the stem of the dendritic structure corresponding to the target source's pitch. If the competing source is also periodic, channels dominated by its energy may also

form part of a dendritic structure on the delay of its pitch period.

Fig. 6.3 compares two correlograms taken at the same time frame of a female speech utterance in either a clean condition (left panel) or when mixed with male speech at a target-to-masker ratio of 0 dB (right panel). The summary ACGs are also shown correspondingly. The dendritic structures which correspond to the $F0$ of sound sources are marked using dashed lines. In the clean condition it is visually clear that the dendritic structure extends across the entire frequency range except those ACG channels whose centre frequency is much below the female speaker's $F0$ (the bottom 5 channels). In the ACG on the right, the dendritic structures corresponding to the two competing speech sources both fail to dominate the whole frequency range. The one extending from 400 Hz to 1300 Hz on the delay of 3.9 ms indicates that there exists a harmonic source with an $F0$ of 256 Hz and the energy of the channels within this range has originated from the female speaker source. The rest of channels form part of another dendritic structure on the delay of 9.0 ms which indicates a second harmonic source with an $F0$ of 111 Hz (the male speaker). This information can be used to separate the two sound sources but is lost in the summary ACG.

6.3.2 Initial Spectral Grouping

ACG channels are pre-grouped before the dendritic structures are extracted in the correlogram. Gammatone filters have overlapping bandwidth and respond to the harmonic with the highest energy. Therefore, ACG channels which are dominated by the same harmonic share a very similar pattern of periodicity [170]. Fig. 6.3 illustrates this phenomenon. For example, in the left panel channels with a CF between 100 Hz and 395 Hz demonstrate a very similar pattern of periodicity. This redundancy can be exploited to effectively pre-group ACG channels. We employ a cross-channel correlation metric [186] where each ACG channel is correlated with its adjacent channel as follows:

$$C(i,t) = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{A}(i,\tau,t)\hat{A}(i+1,\tau,t) \tag{6.3}$$

where L is the maximum autocorrelation delay and $\hat{A}(i,\tau,t)$ is the autocorrelation function of Eq. 6.2 after normalisation to zero mean and unit variance. The normalisation ensures that the cross-channel correlation is sensitive only to the pattern of periodicity of ACG channels,

and not to their energy. Channel i and $i+1$ are grouped if $C(i,t) > \theta$. We choose $\theta = 0.95$ to ensure that only ACG channels with a highly similar pattern are grouped together.

A 'reduced ACG' is obtained by summing pre-grouped channels across frequency. Each set of grouped channels is referred to as a 'subband' in the reduced ACG. The pre-grouping significantly reduces computational cost as the average number of ACG subbands is 39 compared to 64 ACG channels originally. Preliminary experiments also show that the process can effectively reduce grouping errors in the later stages.

6.3.3 Extracting the ACG Structure

The essential idea in this study is to make use of the dendritic structure in the full correlogram for the separation of sound sources. The technique of extracting the pitch-related structure used here is derived from work by Summerfield et al. [178]. For each subband in the reduced ACG, a two-dimensional cosine operator is constructed, which approximates the local shape of the dendritic structure around the subband. The operator consists of five Gabor functions applied to adjacent reduced ACG subbands, in which the middle Gabor function is aligned with the subband it operates on (see Fig. 6.4c). The Gabor function is a sinusoid weighted by a Gaussian. If the sinusoid is a cosine, the Gabor function is defined as:

$$gabor_c(x; T, \sigma) = e^{-x^2/2\sigma^2} \cos(2\pi x/T) \tag{6.4}$$

where T is the period of the sinusoid and σ is the standard deviation of the Gaussian. The frequency of each sinusoid used by Summerfield et al. is the centre frequency of the channel with which it is aligned, and the standard deviation of the Gaussian is 1/CF. This works well with the synthesised vowels in their study. However, speech signals are only quasi-periodic and a filter channel responds to a frequency component that is only an approximation to its CF. Therefore the repeating frequency of the filtered signal in each ACG channel is often off its CF depending on how close the nearest harmonic is to the CF, and sometimes the shift is significant. Therefore in our study we compute the actual repeating period p_i in each ACG subband i by locating the first valley (v_i) and the first and second peaks (p'_i and p''_i) of the autocorrelation function. The repeating period p_i of subband i is approximated as:

$$p_i = \frac{2v_i + p'_i + p''_i/2}{3} \tag{6.5}$$

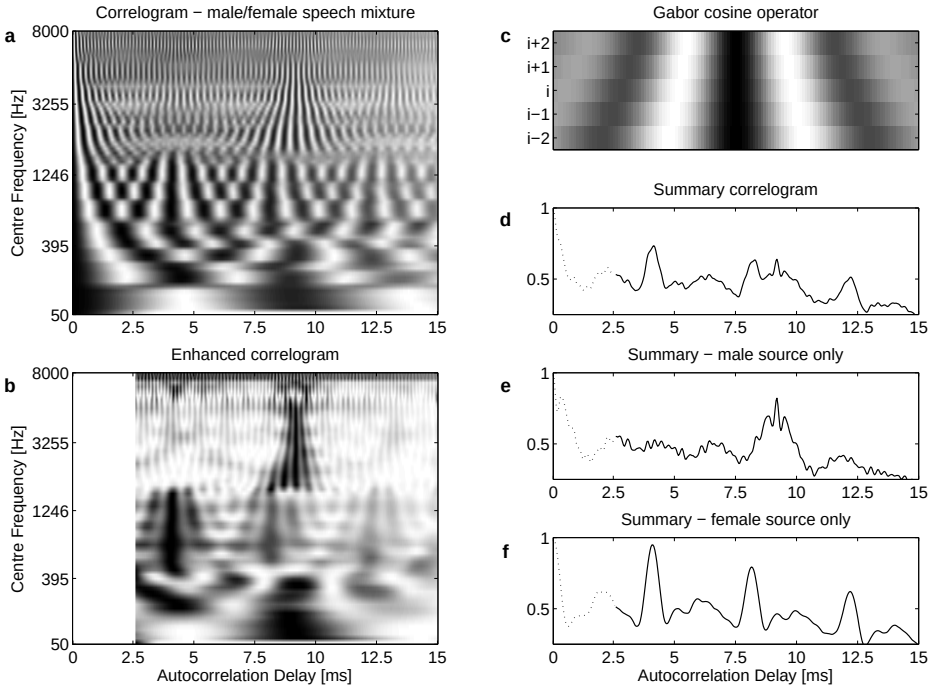

Figure 6.4: *(a) A correlogram of a mixture of male and female speech. (b) Enhanced correlogram after the 2-D convolution. The region with delays less than 2.5 ms (corresponding to regions with F0s higher than 400 Hz) is not computed. (c) An example of a Gabor cosine operator. (d) Summary correlogram. (e–f) Summaries of spectral components in the correlogram dominated by energy from respective speaker sources (male or female). Dotted line represents the high F0 region which is not computed.*

To further enhance the dendrite stem $p_i/2$ is used as the standard deviation in the Gabor function, a value roughly half that used by Summerfield et al.. These changes have been very effective with realistic speech signals.

The autocorrelation function $A(i, \tau, t)$ for each subband i, with support of its four adjacent subbands (two above and two below), is convolved with its corresponding two-dimensional cosine operator after zero-padding, producing an initial enhanced autocorrelation function $A_c(i, \tau, t)$:

$$A_c(i, \tau, t) = \sum_{m=-2}^{2} \sum_{n=1}^{L} A(i+m, \tau+n, t) gabor_c(n; p_{i+m}, p_{i+m}/2) \tag{6.6}$$

where L is the maximum autocorrelation delay. The central part of the convolution is saved for each subband. When the operator is aligned with the stem of a dendrite, the convolution gives

a large product, and the product is smaller if misaligned. Unfortunately, ripples will occur as the cosine operator will also align with peaks other than the stem. Following Summerfield et al. [178], these ripples are removed using a sine operator constructed by substituting the cosine function in Eq. 6.4 for a sine function:

$$gabor_s(x; T, \sigma) = e^{-x^2/2\sigma^2} \sin(2\pi x/T) \tag{6.7}$$

The original correlogram is convolved with the sine operators to generate a function $A_s(i, \tau, t)$ in the same manner as in Eq. 6.6. At each point the results of the two convolutions are squared and summed, producing a final autocorrelation function $A_e(i, \tau, t)$ with the peak in each subband located on the stem of the dendritic structure:

$$A_e(i, \tau, t) = A_c(i, \tau, t)^2 + A_s(i, \tau, t)^2 \tag{6.8}$$

In the enhanced correlogram, A_e, the stems of dendritic structures are greatly emphasised, as illustrated in Fig. 6.4b. The correlogram is computed for a frame in which a female speaker source is present simultaneously with a male speaker source. The two black vertical lines in the enhanced correlogram (one around 3.9 ms and the other around 9.0 ms) are the stems of two dendritic structures which correspond to the two speaker sources. To reduce computational cost, regions with autocorrelation delays less than 2.5 ms (corresponding to regions with $F0$ higher than 400 Hz outside the speech $F0$ range) are not computed.

The largest peak in each subband in the enhanced correlogram is selected and a histogram with a bin width equivalent to 3 Hz is computed over these peak positions. The two highest-counting bins indicate the locations of two possible dendrites corresponding to two harmonic sources. A bin is ignored if its count is less than an empirically determined threshold (5 in this study), therefore in each frame 2, 1 or 0 dendritic structures are found[2].

6.3.4 Final Spectral Grouping

Once the dendritic structures are extracted from the correlogram, the frequency bands can be divided into partial spectra: the ACG subbands with their highest peak at the same position in the enhanced ACG are grouped together. Each group of subbands therefore form

[2]This technique can be extended to handle more sources provided the maximum number of simultaneous periodic sources at each frame is known.

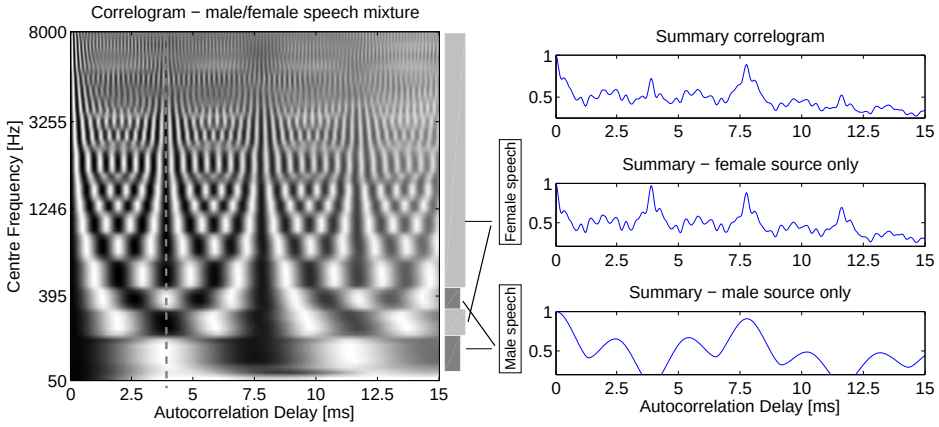

Figure 6.5: *Correlogram of a mixture of male and female speech. The F0 of the male speaker is half of that of the female speaker. Subbands dominated by the energy from different speaker sources are indicated using different shades of grey. The dendritic structure with the shortest delay caused by the female source is marked using a dashed vertical line. Summary of all ACG subbands and those dominated by energy from the female and the male speaker source are shown respectively on the right.*

an extracted dendritic structure. The number of simultaneous spectral groups depends on the number of dendrites identified. If no such structures appear in the correlogram (e.g. for an unvoiced speech frame), the system skips the frame and no spectral group is generated.

After this grouping it is still possible that some ACG subbands remain isolated. Although this is rare, it could happen because a subband may respond to a different dendrite from the one formed by its adjacent subbands. Therefore the subband will not be emphasised in the enhanced ACG. When only one spectral group is formed, an isolated subband is assigned to the group only if it matches the periodicity of the subband within a threshold of 5% in the original ACG. When two spectral groups are formed, an isolated subband is assigned to the group which better matches its periodicity within the threshold of 5%.

This spectral integration technique has the ability to deal with the situation where the fundamentals of two competing speakers are correlated. Fig. 6.5 shows a correlogram computed for a frame in which a male speaker source with a pitch period of 7.8 ms is present simultaneously with a female speaker source with a pitch period approximately half of that (3.9 ms). Since the subbands dominated by the energy from the female source have peaks at an interval of 3.9 ms in the ACG, all the subbands have peaks at the delay of 7.8 ms, causing

the largest peak in the summary ACG to occur at that delay. When the summary ACG is inspected, it is difficult to group subbands as they all respond to the largest peak. However, the female speech subbands will form a partial dendritic structure (marked using a dashed vertical line). The white gaps in its stem clearly indicate that subbands within these gaps do not belong to the female source as otherwise the dendrite would extend across the entire frequency range. Those subbands are actually dominated by the energy from the male speaker source. By exploiting the dendritic structure, a more reliable separation of sources with correlated fundamentals can be performed. Fig. 6.5 also shows the summary of ACG subbands dominated by female and male speaker sources, respectively. The position of the largest peak in each summary clearly indicates the pitch period of each source.

6.4 Sequential Integration

After the spectral integration in the correlogram domain, spectral groups that are likely to belong to same source can be linked together across time to form fragments. In each frame we refer to the source that dominates more frequency channels as the 'stronger' source. If the stronger source were constant from frame to frame, the problem of sequential integration would be solved by simply combining the spectral groups associated with the greater number of channels in each frame. However, due to the dynamic aspects of speech, the dominating source will change as the relative energy of the two sources changes over time. Although a speaker's pitch varies over a considerable range, and pitches from simultaneous speakers may overlap in time, within a short period (e.g. 100 ms) the pitch track produced by each speaker tends to be smooth and continuous. We therefore use this cue to generate harmonic fragments.

Spectral groups produced in the spectral integration stage are combined across time if their pitch estimates form a smooth pitch track segment. Each fragment corresponds to one pitch track segment. This process is illustrated in Fig. 6.6. The upper-left panel shows integrated spectral groups for five frames. Regions with different shades of grey represent different spectral groups in each frame. Pitch estimates for each group in each frame are shown in the lower-left panel. The lower-right panel shows two smooth pitch track segments that are formed. The two corresponding spectro-temporal fragments are shown in the upper-right

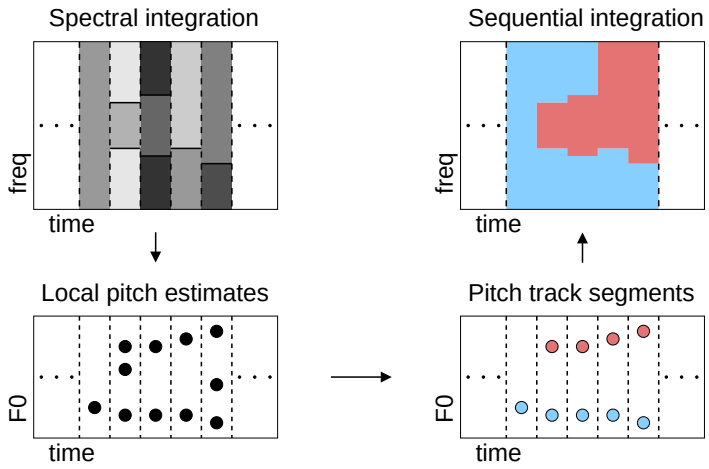

Figure 6.6: *In anticlockwise sequence, upper-left panel: Different shades of grey represent different spectral groups in each frame. Lower-left panel: dots are local pitch estimates for the spectral groups. Lower-right panel: two pitch track segments are produced by linking the local pitch estimates. Upper-right panel: two fragments are formed corresponding to the two pitch track segments.*

panel.

6.4.1 Multipitch Tracking

The original ACG channels grouped in the spectral integration stage are summed and the largest peak in each summary is selected as its local pitch estimate. As shown in Fig. 6.4 (e–f), it is easier to locate the largest peak after spectral integration. The peak that corresponds to the pitch period of each source is very clear in each summary, while locating them in the summary of all ACG channels (panel d) is a more challenging problem. For the stronger source the largest peak is selected as its pitch estimate. For the weaker source (if one exists) up to three peaks are selected as its pitch candidates. Although this is rare, there are situations where the position of the largest peak in the summary of the weaker source does not correspond to its pitch period, due to lack of harmonic energy or errors made in the spectral integration stage. In this case the second and third largest peaks may be just slightly lower than the largest peak and it is very likely that the position of one of them represents the pitch period. Keeping three pitch estimates for the weaker source has been proved beneficial in reducing this type of error. The pitch estimates are then passed to a multipitch tracker to form smooth pitch track segments. The problem is to find a frame-to-frame match for each

pitch estimate. Here we compare two different methods.

Model-Based Multipitch Tracker

Coy and Barker [38] proposed a model-based pitch tracker which models the pitch of each source as a hidden Markov model (HMM) with one voiced state and one unvoiced state. When in the voiced state the models output observations that are dependent on the pitch of the previous observation. Gender dependent models of pitch dynamics are trained from clean speech by analysing the pitch of the utterances in the Aurora 2 training set [150]. In order to track two sources in a pitch space which contains several candidates, two models are run in parallel along with a noise model to account for the observations not generated by the pitch models. The Viterbi algorithm is employed to return the pitch track segments that both models are most likely to generate concurrently. In this study, the model-based tracker is employed in a manner that does not make assumptions about the genders of the speech sources that were made in [124] and [16]. In those papers the two simultaneous speakers were always assumed to be different genders and therefore two HMMs for different genders were used. This manner of application is inappropriate as the genders of concurrent speakers are not known. Therefore in this study three different model combinations (male/male, female/female and male/female) are compared and the hypothesis with the highest overall score (obtained using the Viterbi algorithm) is selected.

Rule-Based Multipitch Tracker

McAulay and Quatieri [129] proposed a simple 'birth-death' process to track rapid movements in spectral peaks. This method can be adapted to link pitch estimates over time to produce smooth pitch track segments. A match is attempted for a pitch estimate p_t in frame t. If a pitch estimate p_{t+1} in frame $t+1$ is the closest match to p_t within a 'matching-interval' Δ and has no better match to the remaining unmatched pitch estimates in frame t, then it is adjoined to the pitch track associated with p_t. A new pitch track is 'born' if no pitch track is associated with p_t and both p_t and p_{t+1} are added into the new track. Analysis of $F0$ trajectories in clean Grid speech [36] showed that in 90% of the voiced frames the inter-frame (10 ms frame-shift) pitch changes did not exceed 5% of the pitch value in the previous frame.

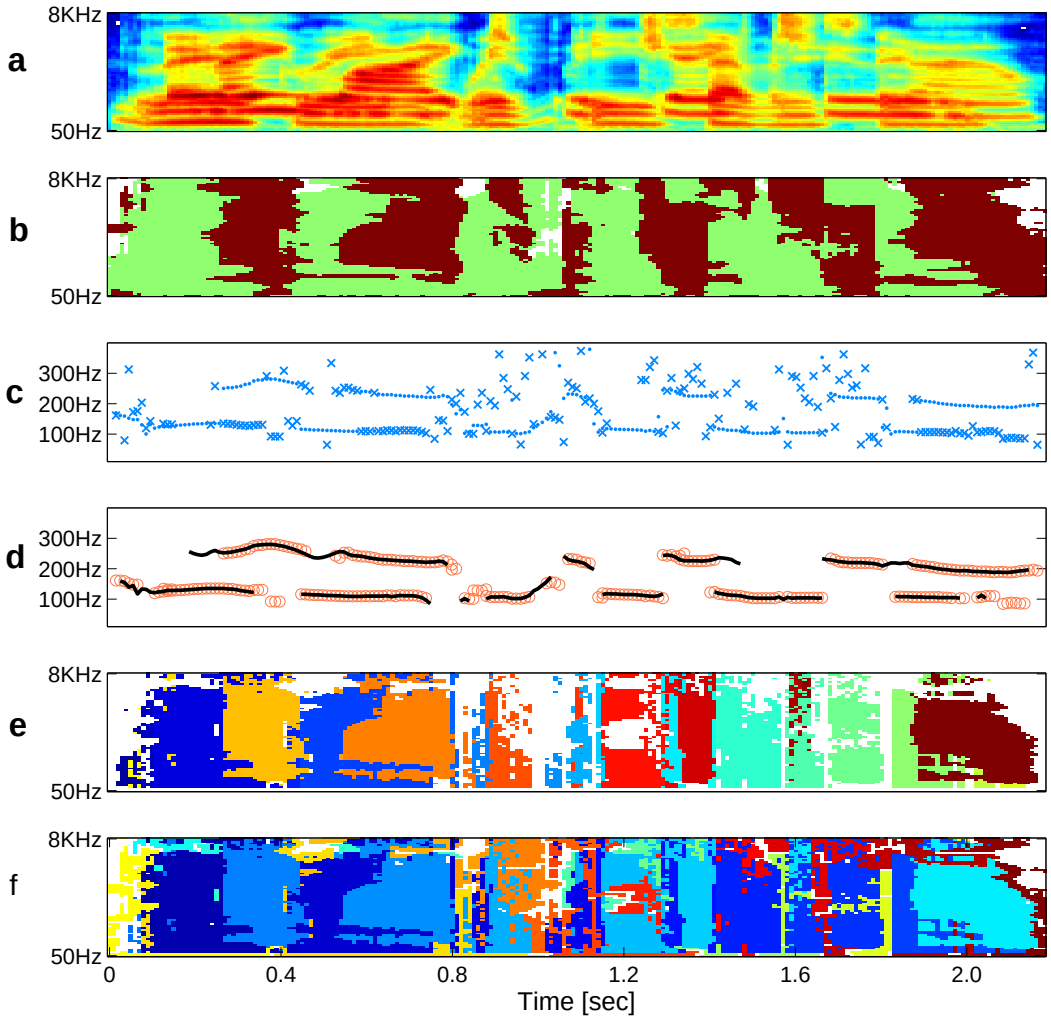

Figure 6.7: *(a) Cochleagram of the mixture of 'lay white with J 2 now' (female) and 'lay green with E 7 soon' (male), TMR = 0dB. (b) The 'oracle' segmentation. Brown region: pixels where the value in the mixture is close to that of the female speaker; green region: the mixture value is close to that of the male speaker; white: low energy regions. (c) Pitch estimates of simultaneous sources. Dots represent pitch estimates of the stronger source in each frame and crosses represent the weaker source. (d) Circles are pitch tracks produced by the multipitch tracking algorithm; solid lines are the ground-truth pitch tracks. (e) Harmonic fragments after sequential integration. (f) Combining inharmonic fragments.*

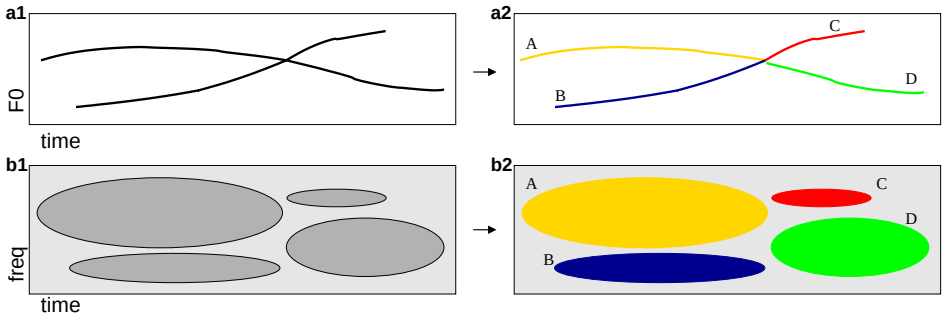

Figure 6.8: *Two intersecting pitch tracks can be represented as four pitch track segments. Four corresponding spectro-temporal fragments can be formed allowing a later decision on fragment combination (e.g. {AD, BC} or {AC, BD}) during the recognition process.*

Therefore, the matching-interval Δ used here was 5% of the pitch estimate which the track is trying to match. This rule-based process was repeated until the last frame.

An example of the output of the rule-based multipitch tracker is shown in Fig. 6.7c. The panel shows the pitch estimates for a female(target)/male(masker) speech mixture. Dots represent pitch estimates of the stronger source in each frame and crosses represent those of the weaker source. The smooth pitch track segments are displayed as circles in panel d, with ground-truth pitch tracks [3] of the pre-mix clean signals displayed as solid lines in the background. The concurrent pitch track segments produced show a close match to the ground-truth pitch estimates. The model-based tracker gave very similar output. Fig. 6.7e shows the fragments produced corresponding to the pitch tracks (Fig. 6.7d) in the example of the female(target)/male(masker) speech mixture. Each fragment is represented using a different shade of grey. It demonstrates a close match between the generated fragments and the 'oracle' segmentation (panel b).

This sequential integration step has also the potential to deal with ambiguous pitch tracks caused by a similar pitch range from different sound sources. Consider the situation where two pitch tracks intersect, as illustrated in Fig. 6.8a1. The ambiguous pitch tracks will be represented as four pitch track segments (Fig. 6.8a2) by the system and hence four corresponding spectro-temporal fragments can be formed (Fig. 6.8b2). This allows the decision on combining fragments (e.g. {AD, BC} or {AC, BD}) to be deferred to the recognition stage.

[3]The pitch analysis is based on the autocorrelation method in the 'Praat' program (www.praat.org).

6.5 Generating Inharmonic Fragments

One weakness of the fragment generation technique described above is that it only handles harmonic regions. Unvoiced speech lacks periodicity and thus does not produce dendritic structures in the correlogram domain. The proposed technique which exploits the periodicity cue skips unvoiced regions and as a result spectro-temporal pixels corresponding to these regions are missing (e.g. the white region at about 1.1 second in Fig. 6.7 (e)). The unvoiced regions of the speech signal are important in distinguishing words which differ only with respect to their unvoiced consonants (e.g. /pi:/ and /ti:/). Therefore it is necessary to include some mechanism that can form coherent fragments for these unvoiced regions.

Hu [95] gave a systematic study of unvoiced speech segregation. In the current work, as the focus is on separation of *periodic* sounds, we employ a simple inharmonic fragment generation technique reported in [39]. Harmonic regions are first identified in the 'ratemap' representation of the mixture using the techniques described in Section 6.4. The 'ratemap' of the remaining inharmonic regions is then treated as an image and processed by the 'watershed algorithm' [78]. The watershed algorithm is a standard region-based image segmentation approach. Imagine the process of falling rain flooding a bounded landscape. The landscape will fill up with water starting at local minima, forming several water domains. As the water level rises, water from different domains meets along boundaries (*watersheds*). As a result the landscape is divided into regions separated by these watersheds. The technique can be applied to segregate inharmonic sources under the assumption that inharmonic sources generally concentrate their energy in local spectro-temporal regions, and that these concentrations of energy form resolvable maxima in the spectro-temporal domain. The inharmonic fragments produced using this technique are pooled together with the harmonic fragment as illustrated, for example, in Fig. 6.7f.

6.6 Estimating Confidence Maps

Another weakness of the system is that it produces 'hard' segmentations, i.e. segmentation in which each spectro-temporal element is marked categorically as either foreground or background. If the early processing has incorrectly grouped elements of the foreground and background into a single fragment, then there will be incorrect assignments in the missing-

data mask which cannot be recovered in later processing. These problems can be mitigated by using missing-data techniques that use 'soft masks' containing a value between 0 and 1 to express a degree of belief that the element is either foreground or background [11]. Such masks can be used in the SFD framework by introducing a spectro-temporal map to express the confidence that the spectro-temporal element belongs to the fragment to which it has been assigned. This confidence map, c_{tf}, uses value in the range 0.5 (low confidence) to 1.0 (high confidence). Given a confidence map, c_{tf}, each hypothesised fragment labelling can be converted into a soft missing data mask, m_{tf}, by setting m_{tf} to be c_{tf} for time-frequency points that lie within foreground fragments, and to be $1-c_{tf}$ for time-frequency points within missing fragments.

In harmonic regions, the confidence map is based on a measure of the similarity between a local periodicity computed at each spectro-temporal point, and a global periodicity computed across all the points within each frame in the fragment as a whole. For each spectro-temporal point the difference between its periodicity and the global periodicity of the fragment measured at that time is computed in Hertz, referred to as x. A sigmoid function is then employed to derive a score between 0.5 and 1:

$$f(x) = \frac{1}{1 + \exp(-\alpha(x - \beta))} \tag{6.9}$$

where α is the sigmoid slope, and β is the sigmoid centre. Appropriate values for these parameters were determined via a series of tuning experiments using a small development data set available in the Grid corpus (see Section 6.7). It was found that the values of these parameters are not critical to the overall performance and $\alpha = 0.6$ and $\beta = -10$ were used in this study.

Confidence scores for the inharmonic fragments in our study are all set to 1. These confidence scores were used in our coherence evaluation experiment and also employed along with generated fragments in the SFD system.

6.7 Experiments and Discussions

Experiments were performed using simultaneous speech data constructed from the Grid corpus [36]. The test set consists of 600 pairs of end-pointed utterances which have been artifi-

cially added at a range of target-to-masker ratios (TMRs). All the mixtures are single-channel signals. In the test set there are 200 pairs in which target and masker are the same speaker; 200 pairs of the same gender (but different speakers); and 200 pairs of different genders. The 'colour' for the target utterance is always 'white', while the 'colour' of the masking utterance is never 'white'.

Three sets of coherent fragments were evaluated and compared on the same task: 'Fragments – Coy' are fragments generated by the system reported by Coy and Barker [39]; 'Fragments – model' and 'Fragments – rule' are coherent fragments generated by the proposed system employing the model-based pitch tracker and the rule-based pitch tracker, respectively.

6.7.1 Coherence Measurement Experiment

The fragments are ultimately employed by the speech fragment decoding ASR system and can be evaluated in terms of the recognition performance achieved. However, in addition to ASR performance, a natural criterion for evaluating the quality of fragments is to measure how closely they correspond to the 'oracle' segmentation, obtained with the access to the pre-mix clean signals (see Fig. 6.7b for an example). To do this we derive the 'coherence' of a fragment as follows. If each pixel in a fragment is associated with a weight, the coherence of the fragment is,

$$100 \times \frac{\max(\sum w_1, \sum w_2)}{\sum w_1 + \sum w_2} \tag{6.10}$$

where w_1 are a set of weights for pixels in the fragment overlapping one source and w_2 are a set of weights for those which overlap the other source. The fragments were compared with the 'oracle' segmentation to identify the pixels overlapping each source. When the decision of each pixel being present or missing in the fragment is discrete (1 or 0), these weights are all simply '1'. In this study we use the confidence scores described in Section 6.5 as the weights. This choice of weight has the desirable effect that incorrect pixel assignments in regions of low confidence cause less reduction in coherence than incorrect assignments in regions of high confidence. Note that regardless of the confidence score, some spectro-temporal pixels may be more important for speech recognition than others. For instance, pixels with high energy representing vowel regions may be of greater value than low energy pixels. It is less critical that the latter pixels are correctly assigned, and ideally, the coherence score should reflect this. In the current measurement, in the absence of a detailed model of spectro-temporal

pixel importance, we make the simple assumption that each pixel has equal importance.

A histogram with a bin width of 10% coherence (hence 5 bins from coherence 50% to 100%) is computed over the set of fragment coherence values. Both the harmonic and inharmonic fragments are included in the experiment. The fragments are different in size. As smaller fragments are less likely to overlap different sources, their coherence is inherently higher. For example, at one extreme, a single-pixel fragment must always have a coherence of 100%. Although we can get higher coherence scores by generating more small fragments, this would be at the expense of reducing the degree of constraint that the primitive grouping processes are providing, i.e. a large number of small fragments produces a much greater set of possible foreground/background segmentation hypotheses. Furthermore, the increased hypothesis space leads to an increase in decoding time. This increase can be quite dramatic, especially if fragments are over-segmented across the frequency axis (see Barker et al. [15]). Therefore the aim here is to produce *large and highly-coherent fragments*. With these considerations, in the coherence analysis, we reduce the effect of the high coherence contributed by small fragments, by weighting each fragment's coherence value by its size when computing the histogram, i.e. a fragment is counted S times if its size is S pixels. The histograms for the three sets of fragments in all mixture conditions at a target-to-masker ratio (TMR) of -9 dB are shown and compared in the top three panels of Fig. 6.9. They have been normalised by dividing the count in each bin by the total number of pixels.

The proposed system with either the model-based pitch tracker or the rule-based pitch tracker produces fragments with very similar quality in terms of coherence. When compared with the fragments generated by Coy and Barker's system, proportionally more fragments with high coherence are produced by the proposed system. This is probably because pitch estimates of each source are computed after the sources are separated. The pitch estimates are thus more reliable and multipitch tracking becomes a much less challenging problem. In Coy and Barker's system, however, pitch candidates are formed from the summary of all ACG channels. The multipitch tracker possibly finds more incorrect tracks through the noisier pitch data. Furthermore, unlike the proposed system where spectral integration is performed before temporal integration, in Coy and Barker's system spectral integration relies on the less reliable pitch tracks. Therefore it is more likely to produce fragments with low coherence. Within each system, the best results were achieved in the 'different gender' condition, presumably

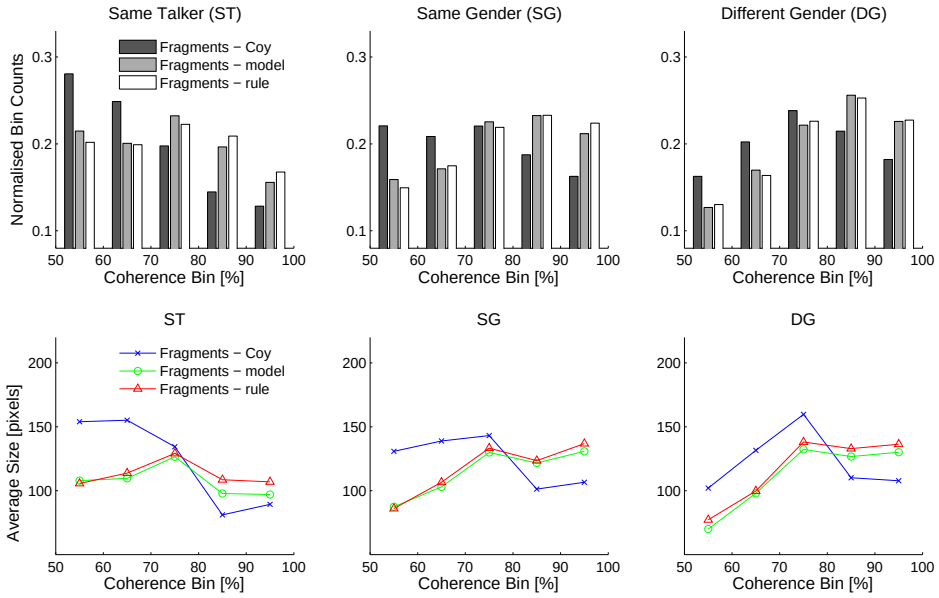

Figure 6.9: *Coherence measuring results for different sets of coherent fragments. Top three panels: histograms of fragment coherence after normalisation (TMR = -9 dB). Each fragment's contribution is weighted by its size when computing the histogram. See text for details. Bottom three panels: average size of fragments in each corresponding histogram bin.*

due to the larger difference in the average $F0$s of the sources.

To examine the impact of fragment sizes on the fragment coherence, we also measured the average size of fragments for each coherence histogram bin, shown in the bottom three panels of Fig. 6.9. Again the two sets of fragments generated by the proposed system give a very similar pattern. In the coherence bins higher than 80% their average fragment size is larger than that of Coy and Barker's system, although in the low coherence bins it is smaller. This is, however, acceptable as there are proportionally less fragments with low coherence in the proposed system.

6.7.2 Automatic Speech Recognition Experiment

The technique proposed here was using the experimental set-up developed in [16] for the Interspeech 2006 Speech Separation Challenge. The task is to recognise the letter and digit spoken by the target speaker who says 'white'. The recognition accuracy of these two keywords were

averaged for each target utterance. The recogniser employed a grammar representing all allowable target utterances in which the colour spoken is 'white'.

In the SFD system a 64-channel log-compressed 'ratemap' representation was employed (see Section 6.2). The 128-dimensional feature vector consisted of 64-dimension ratemap features plus their delta features. Each word was modelled using a speaker dependent word-level HMM in a simple left-to-right model topology, with 7 diagonal-covariance Gaussian mixture components per state. The number of HMM states for each word was decided based on 2 states per phoneme. They were trained using 500 utterances from each of the 34 speakers. The SFD system employs the 'soft' speech fragment decoding technique [39].

The baseline system was a conventional ASR system employing 39-dimensional MFCC features. A single set of speaker independent HMMs with an identical model topology employed 32 mixtures per state. They were trained on standard 13 MFCC features along with their deltas and accelerations.

Following [16], in all experiments, it is assumed that the target speaker is one of the speakers encountered in the training set, but two different configurations were employed: i) 'known speaker' – the utterance is decoded using the set of HMMs corresponding to the target speaker, ii) 'unknown speaker' – the utterance is decoded using HMMs corresponding to each of the 34 speakers and the overall best scoring hypothesis is selected.

We first examine the effect of using soft masks and inharmonic fragments on the recognition performance. The SFD systems with soft masks and inharmonic fragments are then compared to the baseline system and a SFD system using 'Fragments – Coy' with an identical recognition setup.

Effects of Inharmonic Fragments and Confidence Maps

As discussed in Section 6.6, an incorrect decision of a spectro-temporal pixel being present in a fragment cannot be recovered when using discrete masks. This also affects the decoding process in automatic speech recognition as the recogniser will try to match speech models with unreliable acoustic evidence. Therefore we compared the recognition performance using the same set of fragments with discrete masks and soft masks. The soft masks described in

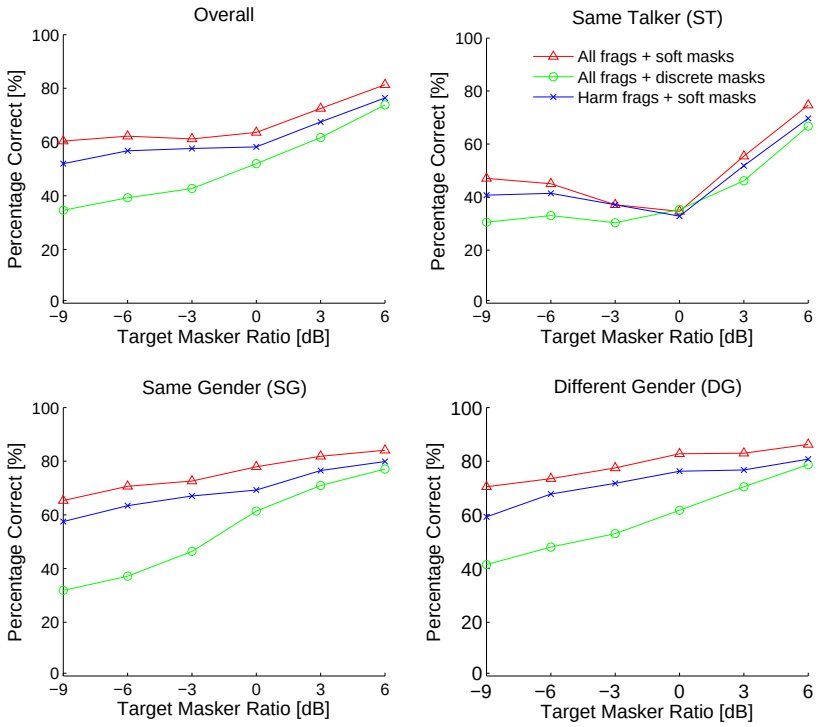

Figure 6.10: *Recognition accuracy performance of the SFD system using 'Fragments – rule' in 'known speaker' configuration. 'All frags + soft masks': all fragments (both harmonic and inharmonic fragments) with soft masks, 'All frags + discrete masks': all fragments with discrete masks, and 'Harm frags + soft masks': harmonic fragments only with soft masks.*

Section 6.6 were employed. The discrete masks were produced by simply replacing all the pixels in the soft masks with '1' if their values are greater than 0.5, and with '0' otherwise. The effect of including inharmonic fragments (Section 6.5) on the recognition performance was also examined. Fig. 6.10 shows recognition results of the SFD system using the set of 'Fragments – rule' in the 'known speaker' configuration. 'All frags + soft masks' represents that both harmonic and inharmonic fragments were used, combined with soft masks. 'All frags + discrete masks' represents results using all fragments but with discrete masks. 'Harm frags + soft masks' is the result with harmonic fragments only using soft masks.

Results show that the soft masks had a considerable effect on the recognition performance. With soft masks the system significantly outperformed that with discrete masks across all conditions. As shown in the coherence measuring experiment many fragments have low

coherence. Some pixels are unreliable and by assigning a confidence score to each pixel the speech fragment decoder is able to weight the pixel's contribution to the decision. Fig. 6.10 also shows that in the 'same talker' condition the SFD system using soft masks did not give any recognition accuracy improvement. One possible reason is that in this condition, as shown in Fig. 6.9, there are more fragments with low coherence and even with soft masks the system could not recover from the errors. Another reason could be that more 'important' pixels were incorrectly assigned in this condition.

Inharmonic fragments also have some impact on the performance in this 'letter + digit' recognition task as many letters are only distinguished by the presence/absence of unvoiced consonants, e.g. letter 'p', 't' and 'e'.

Comparison of Different Fragment Generation Techniques

All the recognition results in this section were obtained with the 'soft' SFD system using both harmonic and inharmonic fragments. Fig. 6.11 shows keyword recognition results of the system using the three sets of coherent fragments discussed before: 'Fragments – Coy', 'Fragments – model' and 'Fragments – rule', in both 'known speaker' and 'unknown speaker' configurations. The 'unknown speaker' results are repeated in Tab. 6.1 (model-based pitch tracker) and Tab. 6.2 (rule-based pitch tracker). Note the 'known – model' and 'unknown – model' results are essentially the same as those published in [16], with minor differences owing to a correction made in the application of the model-based tracker (see Section 6.4.1).

The SFD systems clearly outperform the baseline across all TMRs and across all mixture conditions. They are also able to exploit knowledge of the target speaker identity. The recognition accuracy is significantly higher when the speaker identity is available. Prior knowledge of the speaker identity only fails to confer an advantage in the 'same talker' condition as one would expect. Recognition accuracy results using fragments generated by the proposed system with different pitch trackers are quite similar. This is consistent with the results in the coherence measuring experiment that with different tracks the system produced fragments with similar coherence. The results are significantly better than those produced by Coy and Barker's system, especially at low TMRs. The biggest performance gain was achieved in the 'different gender' condition. This occurs because in this condition the two

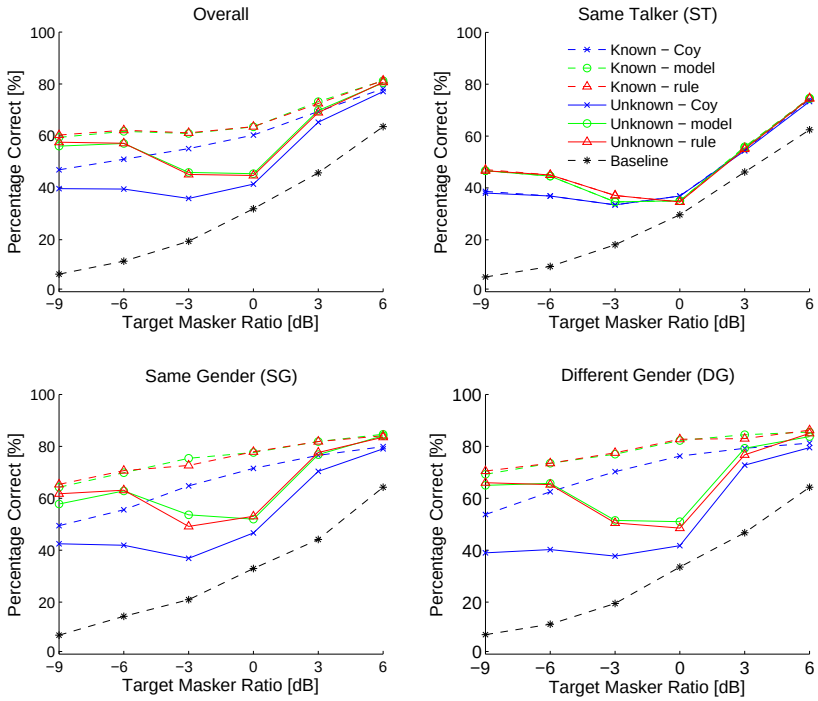

Figure 6.11: *Keyword recognition results of the proposed system with the model-based/rule-based pitch tracker (*model *and* rule*) compared against 'Fragment – Coy' (*Coy*) in both 'known speaker' and 'unknown speaker' configurations. The baseline results are taken from [16].*

sources are more likely to have correlated fundamentals, which is difficult to solve purely based on the summary correlogram as discussed in Section 6.1. The performance improvement in the 'same talker' condition is much less than in the other conditions. This is partially because the target speech and the masker speech are spoken by the same person. With very close $F0$s it is more likely that the pitch-based fragment generation process will group together acoustic evidence from different sources. At low TMRs, same-speaker performance gains may also be reduced because energetic masking is more effective in a same-speaker utterance than in an utterance of different speakers. Many target utterances will be so completely masked at -9 dB that there will be little any system can do to achieve more than chance performance. This effectively reduces the size of the set of utterances on which gains can realistically be made.

It is also instructive to examine the recognition errors. When the decoder is making errors, an interesting question is whether it is incorrectly transcribing the target (due to energetic

Table 6.1: *Keyword recognition correct percentage (%) in unknown speaker configuration using the model-based multipitch tracker.*

	TMR (dB)					
Condition	-9	-6	-3	0	3	6
Overall	56.08	57.08	45.92	45.42	69.75	80.42
ST	46.61	44.57	34.62	35.07	55.43	74.43
SG	57.82	62.85	53.63	51.96	76.82	84.08
DG	65.00	65.75	51.50	51.00	79.25	83.75

Table 6.2: *Keyword recognition correct percentage (%) in unknown speaker configuration using the rule-based pitch tracker.*

	TMR (dB)					
Condition	-9	-6	-3	0	3	6
Overall	57.58	57.17	45.17	44.75	69.00	80.67
ST	46.61	45.02	37.10	34.62	54.98	74.43
SG	61.73	63.13	49.16	53.07	77.65	83.52
DG	66.00	65.25	50.50	48.50	76.75	85.00

masking), or because it is reporting the masker instead (i.e. a failure to 'attend' to the correct source). To investigate this point the recognition output was scored against the correct transcription for the masker utterance.

Tab. 6.3 shows the recognition accuracy results at a TMR of 0 dB when scoring against the target speech (as presented in Fig. 6.11) and when scoring against the masking speech. With the known speaker configuration, the decoder correctly recognised most of the target speech words, without getting confused by the masking speech, in both the 'same gender' and 'different gender' cases. For the artificial 'same talker' condition, however, the reduced performance seems to be explained entirely by the decoder outputting words from the masking utterance. When the simultaneous speech is spoken by the same talker, knowing the identity of the target speaker does not discriminate between fragments of the target and the masker. In fact, at 0 dB there are neither level cues nor speaker identity cues with which to identify the target. For example, when the target speaker says 'a' and the masker (the same speaker) says 'b' concurrently, the two words equally match the known-target speech models and whether 'a' or 'b' is output may be arbitrary.

Table 6.3: *Keyword recognition correct percentage (%) of decoding the target and the masking speech, respectively. TMR = 0 dB.*

	Known speaker			Unknown speaker		
Condition	target	masking	sum	target	masking	sum
ST	34.62	47.06	81.68	34.62	47.06	81.68
SG	77.93	3.07	81.00	53.07	31.01	84.08
DG	82.75	1.25	84.00	48.50	35.25	83.75

In the unknown speaker configuration the decoder exhibits a performance minimum in the range 0 dB to -3 dB. This pattern of results can be broadly explained in terms of the combined effects of two types of masking. Energetic masking occurs in spectro-temporal regions where energy due to the masker dominates that of the target. It prevents the extraction of reliable features. Informational masking [62], on the other hand, is a masking effect that can occur after feature extraction and is partly related to the foreground-background confusion caused by potential similarity between fragments of the target and masker sources.

At the 0 dB TMR level the decoder was unable to use level difference cues to distinguish fragments of the target and the masker. Although the energetic masking effect was weaker than that in lower SNR conditions, the informational masking effect was at its peak. As the TMR fell below -3 dB, the re-introduction of a level difference between the sources compensated for the increased energetic masking and performance initially increased again – at least down to -9 dB.

Although, as discussed earlier, source and target fragments are particularly confusable in the same talker case, the performance dip at 0 dB is also present in the same gender and even in the different gender condition. It appears that the decoder requires level differences to reliably follow the correct source, and speaker differences alone are not enough due to the informational masking effect. This is surprising considering the large acoustic differences that exist between the speaker-dependent models. However, at 0 dB the only cue for distinguishing target and masker is that the target is the person that says ‘white’, in the absence of level cues. So in effect, the system has to solve a speaker identification problem using a single word in the presence of substantial energetic masking. If the word ‘white’ is not heavily masked it will only fit well to one speaker and decoding paths through that speaker model will be the best overall – hence, the target will most often be correctly identified. However, in utterances

where the word 'white' is heavily masked, the fragments masking the word 'white' will be labelled as 'background' and for each speaker there will be a similarly scoring best path. This case is analogous to the word 'white' not being heard, so the cue to the target identity is lost and all speakers become potential targets. In this case, whether the target or the masker is reported may rely on arbitrary factors. In particular, the winning score will depend largely on whether it was the target or the masker who produced an utterance most typical of their average speech patterns, hence leading to the highest likelihood.

6.8 Fragment-Based Speaker Identification

The previous section demonstrated that when the target speaker identity was unknown (i.e. in the 'unknown speaker' configuration), the SFD system was essentially unable to distinguish fragments of the target and masker in the TMR range of -3 to 0 dB (see Tab. 6.3). This resulted in a large drop in performance in all target/masker gender conditions at these TMRs. Unlike listeners, the current decoder seemed to be unable to follow the target source using speaker differences alone. However, when the SFD system was run with prior knowledge of the target speaker identity, results were effectively free from the effects of informational masking in most conditions.

In this ASR task the identifier-word, 'white', is the only word known *a priori* to be spoken uniquely by the target speaker. In order to identify the letter-number keywords spoken by the target speaker, some mechanism is needed to associate the keywords occurring later in the utterance with the identifier-word. This may be achieved in a purely bottom-up manner by tracking low-level properties. For example, if pitch can be tracked from the identifier-word to the letter-number keyword combination, then energy from each region can be incorporated into the same fragment. This does not happen in practice since discontinuities in voicing lead the primitive grouping process to segment the mixture into shorter fragments, typically of the duration of a syllable. Other grouping cues, such as continuity of spatial location, are not available in the monaural mixtures. Even they were, in fact, there is little evidence that the bottom-up processing is reliable for maintaining the perceptual integrity of a sound source over a long period of time.

The grouping of the identifier-word and the letter-digit keywords must be done in a top-

Table 6.4: *Target speaker identification accuracy (%) produced by current SFD system in the unknown speaker configuration. Figures in brackets indicate the percentage of mixtures for which the target is misidentified as the masker.*

	TMR (dB)					
Condition	-9	-6	-3	0	3	6
Overall	94.0	91.7	78.8	76.7	94.2	98.2
ST	98.6	100.0	100.0	100.0	99.1	99.5
SG	89.4 (7.3)	84.9 (15.1)	69.8 (30.2)	65.9 (34.1)	91.1 (8.4)	98.9 (1.1)
DG	93.0 (4.5)	88.5 (11.5)	63.5 (36.5)	60.5 (39.0)	91.5 (5.5)	96.0 (0.0)

down manner by exploiting high-level invariance such as the vocal tract length or accent of the target speaker. In the previous section, top-down tracking was implemented by decoding the fragment set using models for each potential target speaker. The speaker model that gave the highest overall likelihood was selected as the target. However, it is possible that the masker speaker is selected even if it is a poor local match to the identifier-word, since a good fit to masker fragments over the remainder of the utterance can cause it to score better than the target speaker model overall. This is particularly the case if the identifier-word forms only a short portion of the utterance.

Evidence that the SFD system makes this type of error can be seen by examining the speaker identities associated with the ASR hypothesis generated in the 'unknown speaker' configuration. This can be done by investigating the Viterbi backtrace to determine which of the parallel speaker HMMs the winning hypothesis had passed through. Tab. 6.4 presents target speaker identification accuracy computed in this way. For conditions where the target and masker are different speakers, the figures in brackets indicate the percentage of times that the target speaker was misidentified as the masker speaker. At 0 dB the target is being confused for the masker in nearly 40% of cases.

The implementation of top-down tracking in SFD is different from listeners' strategy. In order to minimise potential target/masker confusions, listeners must pay specific attention to the identifier-word 'white'. If a similar top-down tracking model that pays closer attention to the identifier-word could be introduced, then the recognition result around 0 dB TMR could be greatly improved.

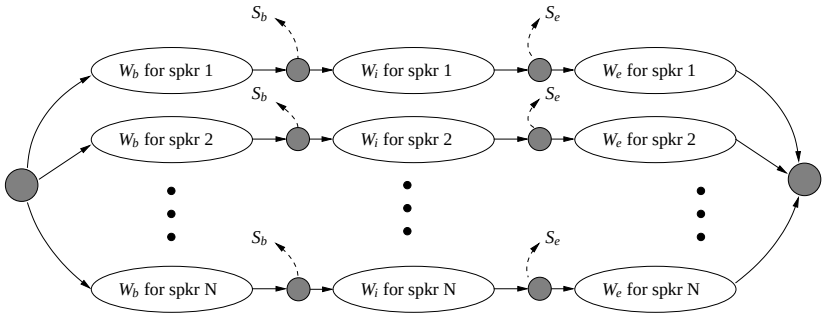

Figure 6.12: *A decoding network for the current SFD system. White ovals represent word-level HMMs, and grey circles are non-emitting nodes connecting the HMMs. S_b and S_e are token scores.*

6.8.1 Attention-Driven Speaker Identification

To tackle the issues raised above, an attention-driven approach to identifying the target speaker is proposed. The technique requires that the target utterances can be represented by a grammar of the form

$$utterance ::= W_b, W_i, W_e \tag{6.11}$$

where W_b (beginning), W_i (identifier) and W_e (end) are three non-terminals grammar items. W_i generates a sequence of *identifier*-words that uniquely identifies the target speaker, i.e. sequences generated by the grammar segment W_i cannot occur in the masker utterance. For the speech separation challenge task, the grammar for the identifier-word sequence, W_i, is simply the word 'white'.

Utterance-level speaker-dependent HMMs are constructed according to the above grammar and placed in parallel as illustrated by the network in Fig. 6.12. The speaker identification mechanism operates by examining token scores generated by the SFD during decoding of the noisy W_i, given a set of fragments generated by the primitive grouping process. Let $S_e(t, seg, n)$ be the scores of tokens that arrive at time t in the non-emitting node at the end of W_i for each foreground/background segregation hypothesis, seg, and for each speaker, n. As the identifier-words will match well to the target model around the time when they finish, tokens through the target model can be expected to have higher likelihoods than those of tokens that have passed through the other speaker models.

To eliminate the contribution to the token score that has been made by word-models in the utterance prior to the identifier-words, (i.e. during W_b), each token maintains a record of the

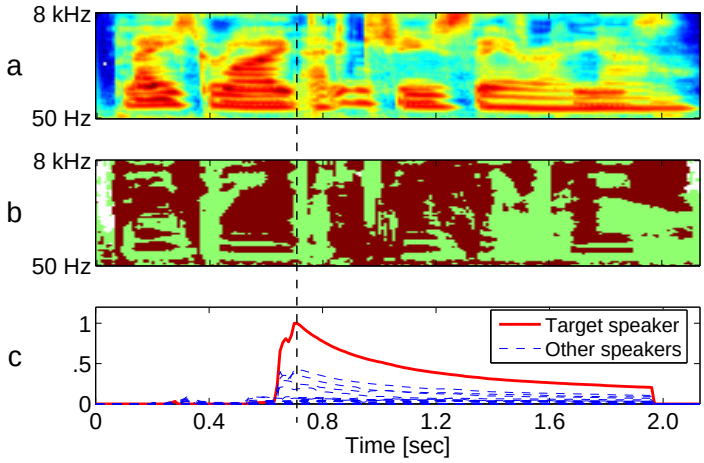

Figure 6.13: *(a) A 'ratemap' representation of the mixture 'place white in G 6 please' (target, female) plus 'lay green at Q 0 now' (masker, female) at 0 dB TMR. The dotted vertical line indicates where the identifier word 'white' finishes. (b) 'Oracle' segmentation: brown region – pixels where the energy in the mixture is close to that of the target speech; green region – mixture energy is close to that of the masker. (c) The score, S_i, computed for each speaker at each frame of the mixture (see Eq. 6.12). The trace for the target speaker is shown as a solid line. The scores have been scaled by dividing by the peak value of S_i.*

score, $S_b(t, seg, n)$, when it first enters into the identifier-word sequence, W_i. Tokens also keep a record of the duration spent traversing the identifier-word sequence model, $D(t, seg, n)$. The score S_b is then removed from the end score S_e in the logarithmic domain and normalised by dividing by $D(t, seg, n)$ to reveal the average score accumulated during W_i alone. For each token in the n^{th} speaker model of segmentation, seg, arriving at the end of W_i at time t, the normalised score is computed as

$$S_i(t, seg, n) = \frac{S_e(t, seg, n) - S_b(t, seg, n)}{D(t, seg, n)} \tag{6.12}$$

The resulting score, $S_i(t, seg, n)$, represents the average rate of score increase of a token as it passes through the 'attended' identifier-word sequence. The target speaker is then identified as the one for which this score reaches the highest value when comparing across all time frames and all segregation hypotheses:

$$target = \underset{n}{\text{argmax}}[\max_{t,seg} S_i(t, seg, n)] \tag{6.13}$$

Fig. 6.13 shows an example of the token scores generated when applying to the ASR task.

Table 6.5: *Target speaker identification accuracy (%) based on token scores. Figures in brackets indicate the percentage of mixtures for which the target is misidentified as the masker.*

	TMR (dB)					
Condition	-9	-6	-3	0	3	6
Overall	85.7	90.3	94.7	96.3	96.2	99.0
ST	82.8	91.0	96.4	98.2	97.3	99.1
SG	85.5 (2.8)	86.6 (3.9)	92.7 (2.8)	95.0 (2.8)	93.3 (2.2)	98.3 (0.0)
DG	89.0 (1.5)	93.0 (1.5)	94.5 (4.5)	95.5 (3.0)	97.5 (2.0)	99.5 (0.0)

The dotted vertical line indicates where the identifier-word 'white' finishes. In Fig. 6.13c the solid line shows token scores, S_i, for the best segmentation produced by the target speaker model at each time. The dashed lines represent the token scores for the best segmentations generated by the remaining speakers. It can be seen that in the first 0.2 seconds no valid tokens reached the non-emitting node at the end of 'white'. When valid tokens started to arrive at the non-emitting node, initially, the scores were low for all speakers because the observations did not fit well to any model for 'white'. Around the time when 'white' finished (indicated by the dotted vertical line at 0.7 seconds) the target speaker received tokens with considerably higher scores, causing a significant jump, while the scores of the other speakers were still relatively low. This is the time when the fragments forming the word 'white' aligned perfectly with the target speaker model. After that, the scores of tokens arriving in the non-emitting node started to become lower.

Although in the current task the identifier-word is known to occur at a fixed position in the utterance, the speaker identification technique allows for more general situations. The grammar of the word segment prior to the identifier-word sequence, W_b, does not need to represent a fixed number of words. For example, W_b may be an arbitrarily lengthened sequence of words taken from a vocabulary that excludes the keyword, in which case the speaker identification algorithm would essentially be co-occurring with a general keyword spotting task.

Speaker Identification Results

Tab. 6.5 presents the target speaker identification results. Compared to results in Tab. 6.4, it is clear that at TMRs above -6 dB the new model offers significantly better results than the original SFD system. At 0 dB TMR the overall speaker identification accuracy was increased

from 76.7% to 96.3%. The masker speaker was very rarely selected. It should be noted that the success of this technique is largely due to the manner in which the speech mixture has been initially separated into spectro-temporal fragments dominated by individual sources. The token scores are generated by the speech fragment decoder based on the most likely fragment combination in each frame. Hence, the system is performing *fragment-based* speaker identification. If the system was constructed using a conventional Viterbi decoder, computing token scores based on full-band observation evidence, the token that passes through the models of the target speaker would not necessarily generate the peak score, S_i.

The model only performs worse than the previous SFD system when either at the extreme -9 dB TMR or when in the same talker (ST) condition. The previous system has the advantage in the ST condition because the attention-driven scheme is designed to reduce target/masker identification confusions which do not occur when both target and masker are the same speaker. In this case it is better to base speaker identification on the whole utterance than on the brief identifier-word 'white'.

Difficulties at -9 dB TMR are occurring probably because in this condition the word 'white' is occasionally fully masked. The fragments masking the word 'white' will be labelled as 'background' and for each speaker there will be a similarly scoring best path. When this happens the cue to the target identity is lost and all speakers become potential targets. In this case, listeners may be using an additional strategy that is not modelled here. For example, at very low TMRs even if the word 'white' is not heard, the colour spoken by the *masker* could be heard clearly, and hence the *masker speaker* can be identified. Listeners can then aim to report the letter and digit that appear *not* to have been spoken by the masker. This strategy would be most effective in the different gender condition.

Another strategy that listeners could be using is to infer that if they have not heard the word 'white' then the target is probably the quieter of the two speakers. Therefore they should focus attention on the quieter speaker when listening for the Grid reference. These strategies could be modelled by using a parallel invocation of the attention-driven speaker identification mechanism which would identify the masker speaker rather than the target. Depending on whether it was the target speaker or the masker speaker that was more reliably identified, the utterance could either be decoded with the target speaker models, or a parallel combination of all speaker models except the masker, respectively.

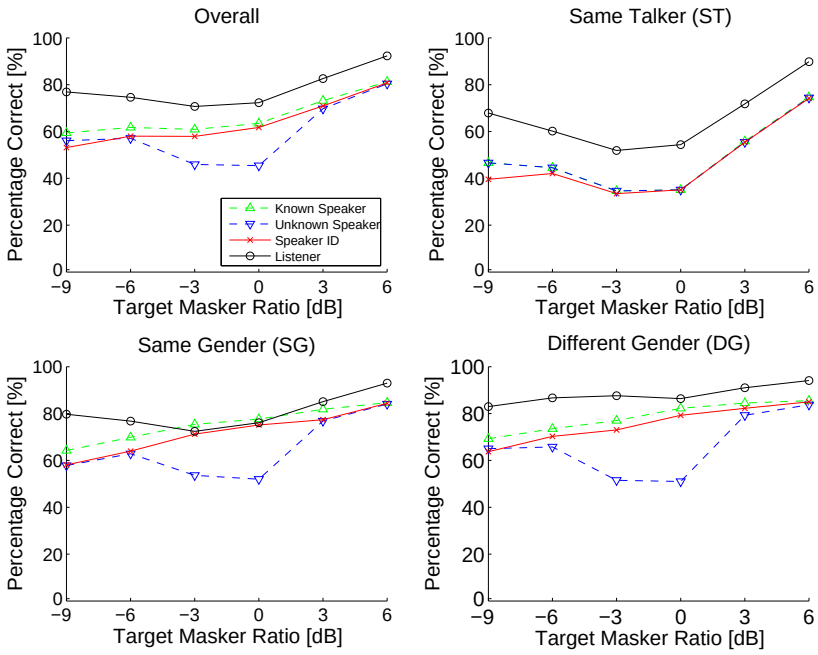

Figure 6.14: *Keyword recognition results for the SFD system incorporating the fragment-based speaker identification model compared against listener data and the known speaker and unknown speaker SFD configurations previously reported in Fig. 6.11 (the model-based results).*

6.8.2 Employing Speaker Identification in SFD

In the final set of recognition experiments, the speaker identification module was integrated into the SFD system. At the end of an utterance, instead of choosing the decoding path from the speaker that gave the highest overall utterance likelihood, the speaker identified by employing the speaker identification technique described above was chosen [4]. The best word sequence hypothesis through this speaker was taken as the recognition output. Fig. 6.14 shows the new recognition results (also in Tab. 6.6) plotted against those previously shown for the SFD system in both 'known speaker' and 'unknown speaker' configurations. Listener results taken from [37] are also plotted for comparison.

[4] This can also be accomplished using a two-pass process where the SFD system was evaluated in the 'known speaker' configuration using the models of the target speaker that has been identified by the speaker identification technique in the first pass.

Table 6.6: *Keyword recognition correct percentage (%) of SFD employing the fragment-based speaker identification technique.*

	TMR (dB)					
Condition	-9	-6	-3	0	3	6
Overall	53.17	58.00	57.92	61.75	70.92	80.83
ST	39.59	42.08	33.48	35.07	55.43	74.21
SG	58.10	63.97	71.23	75.14	77.37	84.36
DG	63.75	70.25	73.00	79.25	82.25	85.00

Discussions

The new SFD system produced significant improvements in recognition accuracy over the 'unknown speaker' results. The speaker identification module appeared to be not particularly sensitive to informational masking. At -3 and 0 dB TMRs where previously informational masking effects frequently led the system to incorrectly report the masker utterance, the improvements are especially large. In the same gender (SG) and the different gender (DG) conditions, the dip in recognition performance around the 0 dB TMR no longer exists, as a result of the eliminated target/masker confusions by employing the speaker identification technique. No improvement over the 'unknown speaker' results was observed in the same talker (ST) condition, as the target speaker identification accuracy in this condition was not better than the previous SFD system. It was more reliable to identify the target speaker based on the whole utterance.

More recognition errors were also reported at -9 dB TMR by the new system. This performance pattern was nearly in line with its target speaker identification accuracy. The recognition accuracy was just a little less than the 'known speaker' results. In fact, the new results can be almost precisely modelled by taking the 'known speaker' results in cases where the speaker identification has been correct, and taking chance level performance where speaker identification is incorrect (chance performance on this task is 7%).

Comparing the new SFD system with human listeners, the system has an overall word error rate that is almost twice as large at most TMRs. Despite the speaker-dependent HMMs being employed in the SFD, it is likely that listeners have more sophisticated acoustic models which encode many top-down cues, such as accents of speakers. The letter-digit keyword

recognition task needs access to subtle phonetic cues to distinguish many of the letters (e.g. 'm' and 'n'; 'p' and 'b'). At low TMRs the effect of energetic masking starts to become significant, and fragments of the target becomes rapidly smaller and less coherent. In these conditions the subtle phonetic cues required for speech recognition may not be available in the HMMs, and can be lost in the less coherent fragments. To achieve performance as good as listeners', more sophisticated acoustic models and better source separation algorithms may be needed.

These hypotheses can also be drawn from the observation that the SFD system performed relatively poorly in the same talker (ST) condition compared to listeners' results. In this unrealistic condition, although the identity of the target speaker is clear (i.e. the speaker identification module made few errors), the target and masker speakers are the same. Target identification alone is not sufficient to avoid foreground/background confusions. In this condition it is also more difficult to separate the simultaneous sources which have the same pitch range and vocal tract length. It can be seen in Fig. 6.9 that the generated fragments are less coherent than the other two conditions (SG and DG). It is likely that listeners employ subtle top-down cues and advanced primitive grouping processing to link glimpses of the identifier-word 'white' with those of the letter-digit keywords spoken by the same target speaker. They are potentially aware of long term inter-dependencies in speech that arise from variations in the manner of speaking even between pairs of utterances of the same speaker.

In the same gender (SG) condition the system achieved the same level of performance as humans at 0 dB and -3 dB TMRs. Tab. 6.4 shows that the system was able to reliably identify the target speaker using the identity-word fragments in this condition. Once the target speaker is correctly identified, the system is able to report the target speaker's keywords as reliably as listeners. However, in these subject listening experiments listeners, unlike the SFD system, did not have access to pre-trained models of the potential speakers [37]. As a result, it is possible that listeners experienced increased difficulty identifying and tracking the target speaker because. It would be instructive to repeat the listener experiments after having allowed the listeners to familiarise themselves with the speakers in the Grid corpus. This may result in listeners exhibiting significant reduction in informational masking effects in the SG condition.

6.9 Summary

This chapter has described a novel approach which exploits the tree-like structure in the correlogram to identify coherent fragments for automatic speech recognition in monaural acoustic mixtures. This technique is compared with a system that employs only the summary correlogram in a coherent measurement experiment. The use of the full correlogram leads to more reliable spectral separation and multipitch estimation, therefore producing highly coherent fragments.

These fragments are also employed by the speech fragment decoding system in a simultaneous speech recognition task. Recognition performance is significantly above that of a conventional HMM ASR system, and is relatively insensitive to the noise level over a broad range of TMR conditions. The system also exhibits a performance pattern similar to that of listeners, with characteristic dips in the TMR range of 0 dB to -3 dB. This pattern of results can be broadly explained as the combined effects of energetic masking and informational masking.

This performance dip at -3 dB and 0 dB TMRs also happened in the different gender conditions. It appears that the decoder is unable to use speaker differences alone to reliably follow the correct source. When there are words known to be spoken by the target speaker, listeners will pay close attention to them and then link the rest reference with the target words. There is no such a mechanism in SFD. An attention-driven speaker identification system is then introduced which is able to pay attention to a particular word – the target speaker is identified as the one for which fragment-based likelihood scores reach the highest when passing through the attended word. The fragment-based approach can be adopted to reliably identify the target speaker. Nearly all of the target/masker confusions that occurred in the original systems are eliminated. Results of the system around 0 dB are greatly improved, especially when considering mixtures of speakers of identical gender, where the SFD performance at 0 dB is not significantly different from that of listeners.

6.9.1 Comparison with Other Systems

A key strength claimed for the solution is that it does not need tailoring to the specific details of the additive noise environment. Considering the current implementation, notwithstanding the assumption made by the pitch estimator that there are at most two harmonic sources

with significant energy at any time instant, the fragment generation stages restrict themselves to using only very general properties of the way that sounds combine. They do not make assumptions about the specifics of either the foreground or background sources. Fragments of the target speech source will be grouped irrespective of the nature of the background. The top-down component only requires a statistical model for the target speech. There is no statistical model representing the background.

The fragment decoding approach contrasts with other techniques that rely on detailed models of both foreground and background. For example, in the Speech Separation Challenge a common approach is to train HMMs for both the target and the masker speaker and then to model how pairs of acoustic states combine. Combining detailed models of this type can lead to better results than those reported here for this problem [e.g. 110, 184]. However, there is an assumption here that one has access to models of the background speakers. In the Challenge data the masker utterance was chosen from the same closed-set of speakers that provided the target utterance, and for which training data was available. Hence, background models were readily available. If the target utterances had been mixed with maskers from a different set then speaker-dependent model composition strategies would not be available. An alternative strategy would be to combine the speaker-dependent foreground models with a speaker-independent background model. The lack of specificity of the background model would presumably lead to poorer results. If the separation task was further generalised so that the background contained one *or more* masker speakers, then the model combination approach would become even more difficult to apply. The SFD approach described here, however, could be applied with essentially no change, and providing good quality fragments could be located, it would be expected to produce a good quality recognition result.

It is also interesting to compare our system with other CASA-inspired systems in this Challenge. For example, Srinivasan et al. [175] use harmonicity to segregate the voiced portions of individual sources in each time frame, and the unvoiced portions are segmented based on an onset/offset analysis. However, their system requires estimating a missing-data mask for the target speaker, in order to perform missing-data speech recognition. Time/frequency segments are combined by searching all possible pairs of the 34 speaker models in the Grid corpus, for the pair that gives the highest speaker identification score. Therefore their system suffers the same limitations as model combination methods, i.e. it would not work if in

the test set there exist masking speakers that are not available in the training set. More importantly, in their system the source segregation hypothesis is fixed prior to the source recognition stage, which may be wrong and cannot be recovered when the speech recognition models are available. The SFD system produced much better results than their system.

6.9.2 Further Development

The current system only exploits dendritic structures in a single correlogram for spectral integration. When spectral components exhibit equal similarities to different dendritic structures representing different sound sources, they are arbitrarily assigned to one without considering information in adjacent correlograms. The assignment of these confusing components may become more obvious if the correlation between correlograms through time is examined.

For simultaneous speech with the same fundamental frequency, although more difficult, listeners are able to perform separation with an accuracy significantly greater than the chance level [167, 4]. The current technique is not able to deal with this situation. Simultaneous voices may be perceived with different timbres even when having the same $F0$ [21]. The timbre is represented in the correlogram as energy distribution across frequency. Therefore, although two concurrent voices with the same $F0$ will exhibit dendritic structures in correlogram with the same stem location, the structure will show inconsistent branches. This could provide possible cues for separation of voices with the same $F0$. Other grouping cues, such as common onsets/offsets and spatial cues, should also work interactively with the pitch cue in forming an overall segregation [46].

The data set used in this study is artificially mixed simultaneous speech. Using the artificially mixed data sets enables us to conduct controlled experiments. Although experiments [37] show that this task is challenging even for human listeners, it lacks some realistic factors such as reverberation and the Lombard effect. Work to investigate the robustness of the fragment generation technique to reverberation is underway. For example, Christensen et al. [29] conducted localisation experiments employing the fragment generation technique discussed in this chapter. Their experiments used binaural speech recorded in a real, reverberant environment. The fragment generation technique was employed to identify local spectro-temporal fragments in which the SNR is high. A fragment-level location estimate was then

constructed by integrating binaural cues of each pixel within a fragment. The fragment based processing is shown to provide significant improvements over their baseline approach in which the binaural location cues were applied directly to the reverberant speech.

Future work will also aim to develop a statistical model of primitive sequential grouping that will weight segmentation hypotheses according to continuity of primitive properties across fragments and through time. For example, the system fails to make use of the pitch continuity across fragments. Two fragments with incompatible pitch tracks cannot belong to the same speaker. This is a strong sequential grouping cue, especially in the condition where the two competing speakers are of different genders.

Chapter 7
Conclusions and Future Development

7.1 Summary of the Thesis

This thesis has investigated a fragment-based approach to robust ASR which works by coupling the problems of source segregation and recognition. Whereas most robust ASR techniques have problems with non-stationary noise, the SFD system mimics listeners in that it is able to take advantage of unmasked glimpses of speech [31] in such conditions. Recognition performance is significantly above that of conventional HMM-based ASR systems on small vocabulary tasks (e.g. the Aurora 2 and the Grid task), and is relatively insensitive to the noise level over a broad range of SNR and noise conditions.

The speech fragment decoding (SFD) implementation described by Barker et al. [15] provides a general framework for CASA-inspired ASR, but in this thesis several possibilities to improve the framework were investigated. When only glimpses of the target speech are available, the decoder often produces word matches with unrealistic durations due to the weak duration constraints in HMM-based ASR systems. Chapter 4 investigated the effect of duration constraints at both state-level and word-level. Evaluated on a connected-digit recognition task, Chapter 4 showed that it is more effective to model duration constraints at word-level. Explicit word duration modelling is able to offer significantly lower word error rates (WERs) in various noisy conditions by favouring word matches with proper durations in the decoding process. Modelling the prepausal lengthening effect – the property that before a speech pause the preceding speech unit tends to lengthen – is also proved beneficial in reducing WERs.

A key strength claimed for the solution is that it does not need tailoring to the specific details of the noise environment. For example, the same SFD technique has been applied in the past with both speech [16] and non-speech backgrounds [15]. There is no statistical model representing the background. However, recognition experiments using oracle fragments suggest that the top-down information in speech models is often insufficient to recruit correct speech evidence. Although the minimum assumption about the background is the strength of SFD, constraints that can distinguish speech from noise will benefit the decoding process. Chapter 5 introduced a 'speechiness' measure into the SFD framework – a degree of confidence that the fragment is part of the speech foreground. The measure is employed to bias the decoder towards selecting fragments that are more likely to be part of the speech source. A modulation filtering technique which emphasises the characteristic low-frequency modulation energy of speech is shown to be an effective speechiness measure for the various types of noise employed. Recognition experiments show that the speechiness measure can help the decoder employ more reliable speech evidence and therefore produce significant improvements in accuracy.

The quality of fragments is fundamental to the performance of SFD. If fragments contain too much energy that belongs to different sources in the first place, the accuracy of fragment-based recognition is likely to be poor. The number of fragments also determines the computational load of SFD. Having less fragments means there are less segregation hypotheses to be considered. Chapter 6 described a novel approach which exploits the tree-like structure in the correlogram to identify coherent fragments for automatic speech recognition. The improved fragment generation technique is able to produce reliable multipitch estimates and more coherent fragments from simultaneous speech.

Evaluated on a simultaneous speech recognition task the system exhibits performance curves similar to those of listeners, with characteristic dips in the SNR range of 0 dB to -3 dB. This pattern of results can be broadly explained as the combined effects of energetic masking and informational masking. This performance dip also occurs in the conditions where the two competing speakers are of different genders. It appears that the decoder is unable to use speaker differences alone to reliably follow the correct source. A fragment-based speaker identification approach is adopted to reliably identify the target speaker. This approach allows the decoder to actively attend to a word sequence which is known to be spoken by

the target speaker. By exploiting this information nearly all of the target/masker confusions that occurred in the original systems are eliminated.

All the techniques presented in this thesis for improving SFD can be combined in a single system. The HMM-unrolling technique for word duration modelling in Chapter 4 works in the model space, and do not require changes to the existing fragment decoder architecture. The multistack decoding technique requires modifications to the Viterbi decoder used in SFD, and therefore is not flexible in designing new systems. The speechiness measuring technique presented in Chapter 5 was applied to oracle fragments, but the same technique can be applied to fragments generated using techniques presented in Chapter 6. This work is currently being investigated.

7.2 Novelty of the Work

The work discussed in this thesis is based on a novel framework for CASA-inspired ASR, which considers source separation and recognition as being coupled problems. The previous implementation of SFD has been extended by introducing various constraints to improve speech decoding in a multisource environment.

While explicit duration constraints have less impact on ASR in quiet conditions which match the training condition, the work has demonstrated that they help the decoder in adverse conditions. The duration constraints can be incorporated by using a simplified stack decoder, or by using unrolled HMMs which does not require modification to the existing decoder. The system has also investigated the lengthening property that speech units have before a speech pause. Modern speech recognition systems do not normally represent this, which can lead to errors.

Although the speechiness measure is a crude approximation to the segregation model in SFD, it provides an efficient way to exploit extra top-down constraints when available. These extra constraints are essential for SFD to follow the correct source in a multisource environment. The modulation spectrogram, which is normally used as a reliable speech recognition feature, is employed by the system in a different way – fragments high in the 4-Hz modulation energy are considered as having more 'speechiness'. The measure is effective for many types of noise

as it focuses on properties that are generally unique to speech. The technique is a valid extension to SFD in the way that they both do not require statistical noise models.

The fragment generation techniques reported are based on the auditory scene analysis accounts of sound organisation. They do not make assumptions about the specifics of either the foreground or background sources. Fragments of the target speech source will be grouped irrespective of the nature of the background. The techniques are novel because instead of the summary correlogram employed in most CASA-based systems, the full correlogram is analysed which provides more informative source separation cues. This leads to more reliable multipitch estimation and therefore highly coherent fragments.

7.3 Limitations

The acoustic mixtures used throughout the study were prepared by artificially mixing noise signals with speech recorded in quiet conditions. This unrealistic scenario poses limitations on the experiments reported. Speakers normally modify the way they speak in noisy conditions to make their communication easier, i.e. the Lombard effect on speech [106]. This may have effects on speech such as increased loudness and slower speaking rate. Reverberation will also cause problems. Therefore the assumption that the unmasked glimpses of speech in noise will match clean speech models is somewhat unrealistic. This is a limitation of many ASR systems evaluated using artificial data.

7.3.1 Duration Modelling

The unrealistic situation is also a problem for duration modelling. The duration statistics were obtained using speech recorded in a quiet environment. Therefore the duration modelling techniques proposed assume that word durations remain constant in various noise conditions. In a realistic situation, duration models should be dynamically adapted based on feedback from a speaking rate detector.

The current techniques are also limited to small vocabulary tasks where word-level HMMs can be used. Large vocabulary speech recognition tasks typically employ phone-level HMMs. Although similar techniques can be applied to model phone durations instead of word dura-

tions, it is certain that more complex duration models need to be considered in order to get any improvement in accuracy.

7.3.2 Speechiness Measures

The speechiness measures based on modulation filtering are not temporally very precise. Emphasising the 4-Hz syllabic rate requires segment duration of at least 250 ms and therefore they may not be very informative in short term, e.g. a 10 ms frame commonly employed in speech processing. Although the current system overcomes this by averaging modulation energy over all the T/F components included in a fragment, these measures are less reliable for small fragments.

The speechiness measuring technique will be less effective if the background noise has a rhythm similar to the syllabic rate. Essentially it assumes that the target source is the only speech source and therefore will not apply to conditions such as simultaneous speech. Therefore the technique needs to be generalised to represent unique properties of the target source.

The current method determined each fragment in a segregation hypothesis to be part of the foreground independently. This assumption provides an efficient implementation for the segregation model, but the independence may not always be the case. For example, a fragment dominated by energy of an unvoiced consonant may be very like noise on its own. However, if it is followed by a fragment in which energy matches that of a vowel, they may together form a complete syllable. A better implementation would consider applying speechiness measures to each segregation hypothesis which may include multiple fragments.

7.3.3 Fragment Generation

The current system only exploits dendritic structures in a single correlogram for spectral integration. When spectral components exhibit equal similarities to different dendritic structures representing different sound sources, they are arbitrarily assigned to one without considering information in adjacent correlograms. The assignment of these confusing components may become more obvious if the correlation between correlograms through time is examined.

For simultaneous vowels ('double vowels') with the same fundamental frequency, listeners are able to perform identification with an accuracy significantly greater than the chance level [167, 4], although with more difficulties than if fundamental frequency is different. The current technique is not able to deal with this situation. Simultaneous voices may be perceived with different timbres even when having the same $F0$ [21]. The timbre is represented in correlograms as energy distribution across frequency. Therefore, although two concurrent voices with the same $F0$ will exhibit dendritic structures in correlograms with the same stem location, the structure will show inconsistent branches. This could provide possible cues for separation of voices with the same $F0$. Other grouping cues, such as common onsets/offsets and spatial cues, should work interactively with the pitch cue in forming an overall segregation [46].

7.4 Future Development

7.4.1 Statistical Segmentation Models

Auditory segmentation includes simultaneous grouping which organises sound components across frequency, and sequential grouping which links segments to form continuous temporal streams. Nearly all CASA studies have focused on the deterministic signal-driven processing which describes how sound components may be grouped across time/frequency according to the correlations of their characteristics. Signal-driven processing by itself is generally less robust due to the great variability of speech properties over time and their deterministic nature.

Novel approaches to statistically modelling the foreground/background segmentation will be investigated. Signal-driven processing can be employed to suggest initial local groupings based on signal properties such as harmonicity, spectral regularity and sound location, which have been proved effective in organising spectral components. Potential ways of combining multiple grouping cues will be investigated in order to produce more reliable spectral groupings. The next stage will be to investigate statistical sequential grouping algorithms that act on these signal properties. Instead of deterministic signal-driven processing, an acoustic generative model can be built to accounts for the temporal variability of the signal properties. The statistical segmentation model will ultimately allow the learnt patterns of speech encoded in

the recognition models, which are normally ignored in most CASA models, to be effectively combined with signal-driven grouping cues to sequentially organise sound scenes.

7.4.2 Fragment-Based Model Combination

The fragment decoding approach contrasts with other techniques that rely on detailed models of both foreground and background. These techniques assume that one has access to models of the background. However, if knowledge of the background does happen to be available, combining detailed background models can lead to much better results. The fragment generation front-end employed in SFD lessens the need for a strong background model but it does not make such a model obsolete. If an HMM for the background source were available then fragments could still be assigned to either foreground or background. In this case likelihoods should be maximised jointly over both fragment assignment and state sequences for both the foreground and background models. This is similar to HMM decomposition [182] but with constraints from fragments. The probability calculations would require some modification and the state-space would be larger. Therefore pruning will be essential during the search process. Further work is needed to examine ways in which background knowledge can be smoothly integrated into the fragment-based architecture.

7.4.3 Fragment-Based Model Adaptation

Future work will also include investigating how the fragment constraints can be exploited to adapt and elaborate existing models of both the foreground and background. In its deployment the system will generally require adaptation strategies that operate in an online mode where models are adapted in real-time. This is important for applications such as a personal assistant. General speaker-independent models may be needed at the beginning. A solution to more robust performance lies in the on-line adaptation of speaker-independent models toward the target speaker.

Techniques need to be developed that exploit the partial segmentation produced by the fragment generation process to allow robust adaptation of speech models from noisy speech data in a multisource environment. For example, partial tracebacks generated by the decoder can be used to suggest an ongoing segmentation hypothesis. The source model parameters

can then be adapted toward the partially observed acoustics of each source. Adaptation techniques will need to be designed to be compatible with the fact that most sources will have regions labelled as 'unreliable' where the segmentation dictates that they are masked by competing sources.

Appendix A
The Gammatone Filter

A.1 Definition

The gammatone filter [101] is widely used in models of the auditory system and defined by its impulse response as:

$$g(t) = at^{n-1}cos(2\pi ft + \phi)e^{-2\pi bt} \quad (t > 0) \tag{A.1}$$

where a is the amplitude; n is the order of the filter which largely determines the slope of the filter's skirts; f is the filter centre frequency; ϕ is the phase; b is the bandwidth of the filter and largely determines the duration of the impulse response. With an order of four Patterson et al. [148] showed that the impulse response of the gammatone function provides an excellent fit to the human auditory filter shapes derived by Patterson and Moore [147]. The bandwidth of the gammatone filter is usually set according to the equivalent rectangular bandwidth (ERB), which Glasberg and Moore [75] summarised based on human auditory data with the function:

$$ERB(f) = 24.7(4.37 \cdot 10^{-3}f + 1) \tag{A.2}$$

For fourth-order filters the bandwidth b is given as $1.019ERB(f)$.

A.2 An Efficient Implementation

Efficient digital implementations of the gammatone filter were proposed in [91, 30, 172]. The implementation employed in this thesis was based on the implementation by Cooke [30] using the impulse invariant transformation. For each sample, t, the signal was first multiplied by a complex exponential $e^{-j2\pi ft}$ at the desired centre frequency f, then filtered with a base-band gammatone filter, and finally shifted back to the centre frequency region by multiplying the signal by $e^{j2\pi ft}$.

The cost of computing the two complex exponentials for every data sample is significant. Experiments showed that it takes up to 70% of the total computational load. To reduce the computation cost, the exponential computation was transformed into multiplication by rearranging the two complex exponential functions, e.g.,

$$e^{-j2\pi ft} = e^{-j2\pi f} e^{-j2\pi f(t-1)} \tag{A.3}$$

The term $e^{-j2\pi f(t-1)}$ is the exponential calculated for the previous sample at time $t-1$. Therefore only one complex exponential $e^{-j2\pi f}$ for the first sample needs to be computed and the rest can be obtained recursively by multiplication. In practice, the two complex exponential functions are computed using trigonometric functions, i.e.,

$$e^{-j2\pi ft} = \cos(2\pi ft) - j\sin(2\pi ft) \tag{A.4}$$

and

$$e^{j2\pi ft} = \cos(2\pi ft) + j\sin(2\pi ft) \tag{A.5}$$

Experiments showed that this implementation is 4 times faster than the implementation proposed by Cooke [30].

Appendix B
Prepausal Duration Examples in Switchboard I

The prepausal lengthening effect (see Section 4.3.1 for details) is very strong in conversational speech such as the SVitchboard corpus [108] (a subset of the SwitchBoard I corpus). Although there is an intro/inter-speaker difference in the speaking rate, the duration of words (mainly vowels) is heavily influenced by the following pause. Fig. B.1 illustrates this effect. Two sentences both containing the word *know* are used here. In sentence (a) *know* occurs before another word and its duration lasts 141 ms. In sentence (b) where *know* precedes a speech pause, its duration is significantly longer (436 ms).

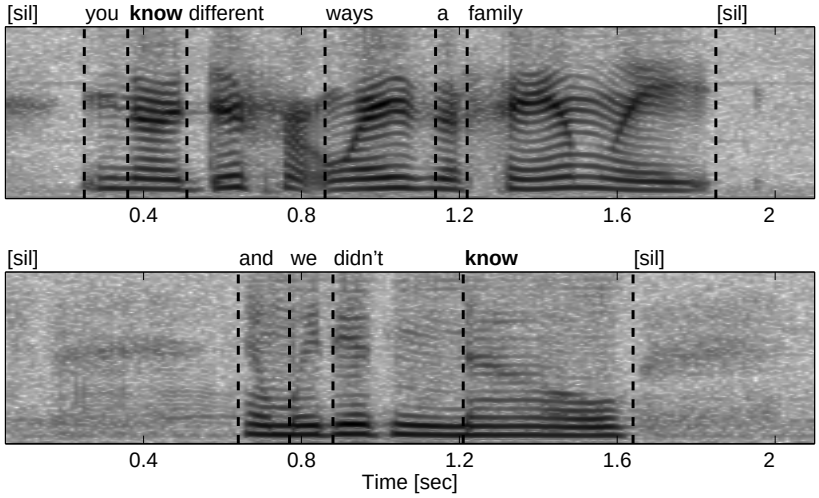

Figure B.1: *An example from the SVitchboard corpus to illustrate the prepausal lengthening effect. The transcription is shown at the top of the spectrogram of each audio signal with segmentation indicated by dashed lines. The word know lasts 141 ms in (a) and 436 ms in (b) where it precedes a speech pause ([sil]).*

More listening examples are available at http://www.dcs.shef.ac.uk/~ning/research/prepausal/. The following tables show word duration statistics in the SVitchboard corpus.

Table B.1: *Mean durations (Mn.) and standard deviations (s.d.), in milliseconds, of the most frequent 20 words in SVitchboard I. N = number of cases; Mn. Inc. = mean duration increase.*

	All examples		Non-prepausal			Prepausal				
word	Mn.	s.d.	N	Mn.	s.d.	N	Mn.	s.d.	Mn. Inc.	Inc.%
1: i	141	88	7633	129	74	748	260	127	131	101%
2: and	296	167	4348	272	157	1055	398	170	126	46%
3: you	141	76	4456	126	57	589	251	105	124	98%
4: oh	335	219	2915	249	146	1305	527	234	278	112%
5: that	231	105	2935	209	95	1153	287	108	78	37%
6: right	399	133	677	366	159	2987	407	126	41	11%
7: it	150	75	2428	137	69	907	186	81	49	36%
8: know	213	112	2065	172	86	1091	290	114	118	68%
9: to	148	107	2500	124	79	412	298	131	174	140%
10: that's	265	84	2488	258	76	198	354	124	96	37%
11: well	271	138	1740	225	119	892	362	127	137	61%
12: the	152	115	2042	119	81	546	277	135	158	133%
13: a	107	95	2145	82	67	431	229	118	147	178%
14: so	331	156	1431	263	139	1087	421	130	158	60%
15: but	249	119	1718	240	121	756	269	112	29	12%
16: of	121	82	1764	107	63	224	231	117	124	116%
17: it's	235	111	1676	216	94	305	336	138	120	56%
18: do	189	113	1411	160	87	269	343	111	183	115%
19: think	257	95	1353	243	84	241	330	114	87	36%
20: they	172	92	1293	155	75	209	279	115	123	79%
							Min		29	11%
							Max		278	178%
							Mean		124	77%
							s.d.		56	46%

Table B.2: *Duration statistics of 10 words in SVitchboard with the most insertion errors produced by a graphical models based ASR system.*

		Non-prepausal			Prepausal				
word	# Errors	N	Mn.	s.d.	N	Mn.	s.d.	Mn. Inc.	Inc.%
i	12	7633	129	74	748	260	127	131	101%
that	11	2935	209	95	1153	287	108	78	37%
oh	10	2915	249	146	1305	527	234	278	112%
you	10	4456	126	57	589	251	105	124	98%
and	6	4348	272	157	1055	398	170	126	46%
it	6	2428	137	69	907	186	81	49	36%
know	6	2065	172	86	1091	290	114	118	68%
well	6	1740	225	119	892	362	127	137	61%
right	5	677	366	159	2987	407	126	41	11%
it's	4	1676	216	94	305	336	138	120	56%
							Min	41	11%
							Max	278	112%
							Mean	120	63%
							s.d.	65	33%

Table B.3: *Duration statistics of 10 words with the most deletion errors in SVitchboard.*

		Non-prepausal			Prepausal				
word	# Errors	N	Mn.	s.d.	N	Mn.	s.d.	Mn. Inc.	Inc.%
i	16	7633	129	74	748	260	127	131	101%
that	11	2935	209	95	1153	287	108	78	37%
oh	10	2915	249	146	1305	527	234	278	112%
and	9	4348	272	157	1055	398	170	126	46%
is	9	1066	189	105	276	368	152	178	94%
it	8	2428	137	69	907	186	81	49	36%
the	8	2042	119	81	546	277	135	158	133%
do	7	1411	160	87	269	343	111	183	115%
so	6	1431	263	139	1087	421	130	158	60%
well	6	1740	225	119	892	362	127	137	61%
							Min	49	36%
							Max	278	133%
							Mean	148	80%
							s.d.	62	36%

Table B.4: *Duration statistics of 10 words with the most substitution errors in SVitchboard.*

		Non-prepausal			Prepausal				
word	# Errors	N	Mn.	s.d.	N	Mn.	s.d.	Mn. Inc.	Inc.%
it	23	2428	137	69	907	186	81	49	36%
i	22	7633	129	74	748	260	127	131	101%
that	16	2935	209	95	1153	287	108	78	37%
to	16	2500	124	79	412	298	131	174	140%
you	16	4456	126	57	589	251	105	124	98%
is	14	1066	189	105	276	368	152	178	94%
a	12	2145	82	67	431	229	118	147	178%
oh	12	2915	249	146	1305	527	234	278	112%
know	10	2065	172	86	1091	290	114	118	68%
the	10	2042	119	81	546	277	135	158	133%
							Min	49	36%
							Max	278	178%
							Mean	143	100%
							s.d.	62	45%

Appendix C
Noise Material Employed in the Speechiness Study

C.1 The Noise Material

Six types of noise with various characteristics were selected for the speechiness study reported in Chapter 5 and summarised as below:

1. *Violins*: Vivaldi Spring mvt 1 Allegro, harmonic source, clearly visible harmonics with significant amount energy occurring in high frequency bands (above 3 kHz)
2. *Piano*: Chopin Nocturne op 9 no 2, harmonic source, most energy in low frequency bands (below 2 kHz)
3. *Singing voice*: female vocal solo with piano accompaniment, harmonic source, mostly overlapping speech energy
4. *Drums*: taken from [14], fast rhythms with clear energy onsets synchronised across frequency, mostly overlapping speech energy
5. *Speech babble*: taken from the NOISEX–92 corpus [183], non-stationary, most energy inharmonic and overlapping speech energy
6. *Factory noise*: from NOISEX–92, a stationary background with highly unpredictable components such as hammer blows etc, inharmonic, full band noise

All noise signals were resampled to 25 kHz and normalised to have target rms of 0.05. Each noise signal is around 30 seconds long.

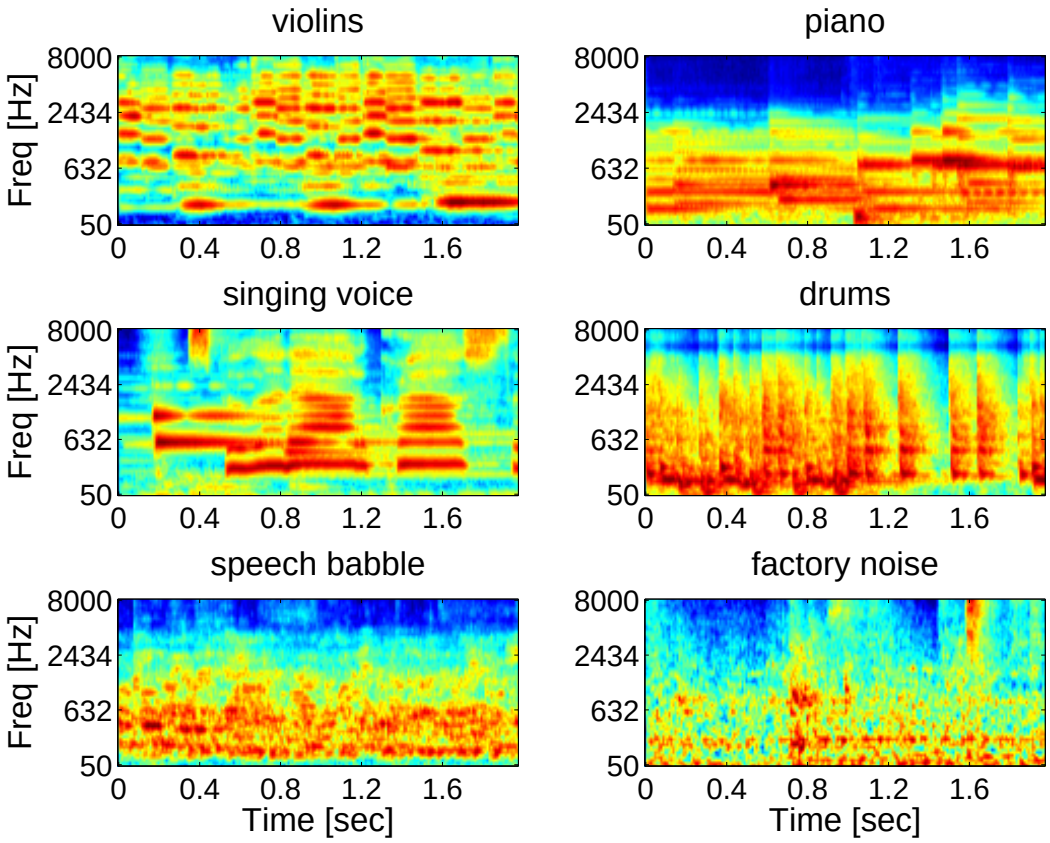

Figure C.1: *Spectrograms of the 6 types of noise used in the 'speechiness' study.*

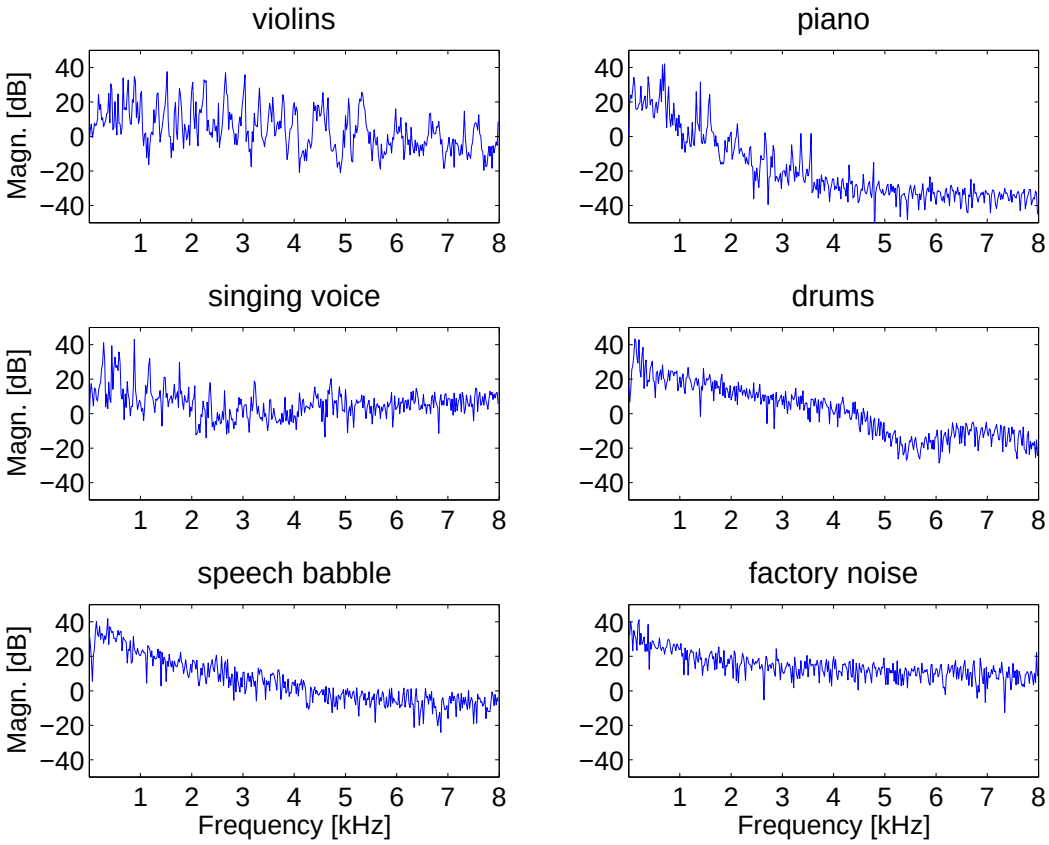

Figure C.2: *Long-term spectrum of the 6 types of noise used in the 'speechiness' study.*

C.2 Examples of Oracle Fragments

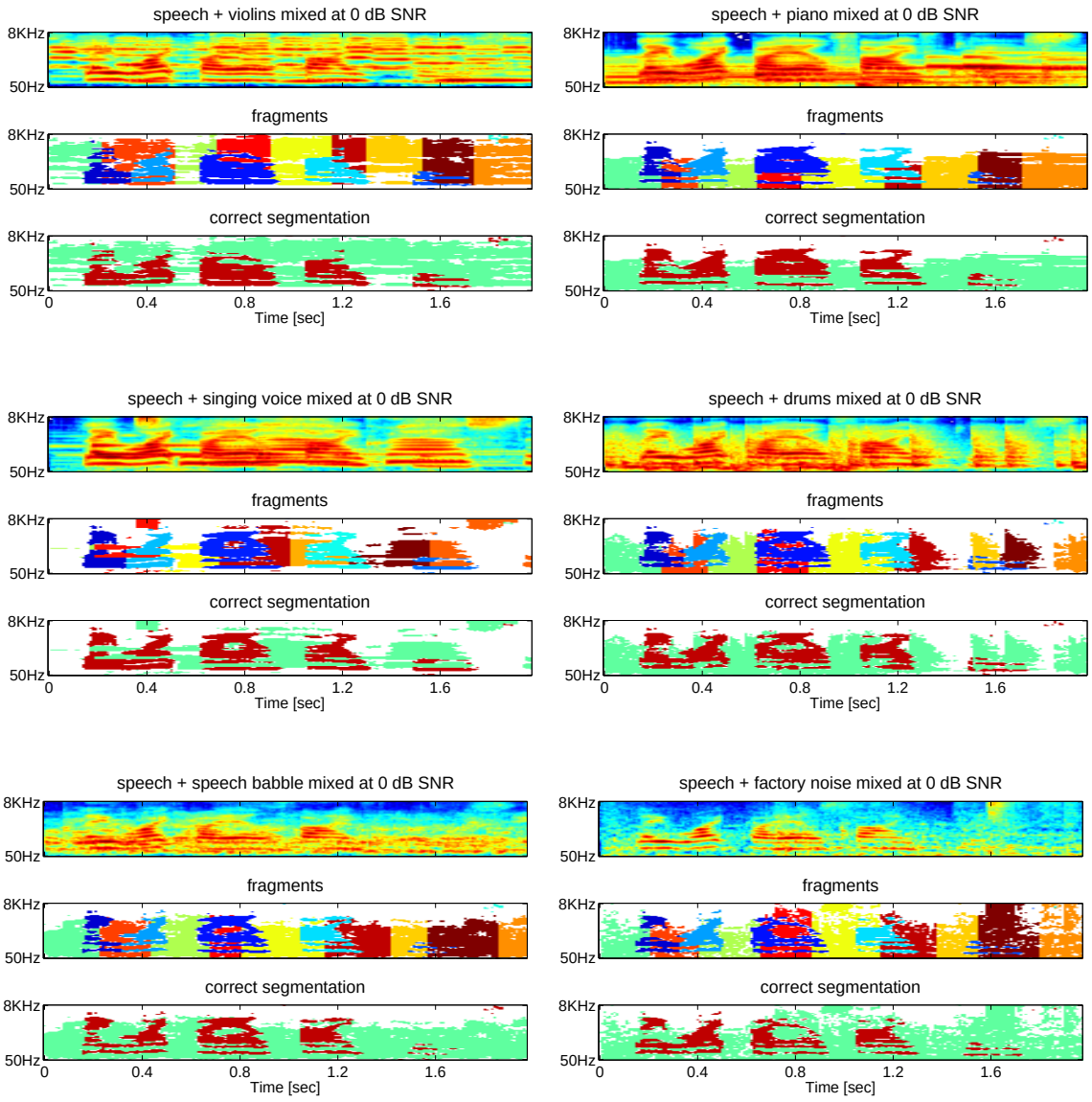

Figure C.3: *Examples of the oracle fragments for various speech/noise mixtures, SNR = 0 dB.*

Bibliography

[1] S. Ahmed and V. Tresp. Some solutions to the missing feature problem in vision. In S. Hanson, J. Cowan, and C. Giles, editors, *Advanced in Neural Information Processing Systems*, volume 5, pages 393–400. Morgan Kaufmann, San Mateo, CA, 1993.

[2] J. Allen. How do humans process and recognize speech? *IEEE T. Speech. Audi. P.*, 2 (4):567–577, 1994.

[3] J. A. Arrowood and M. A. Clements. Using observation uncertainty in HMM decoding. In *Proc. ICSLP'02*, pages 1561–1564, 2002.

[4] P. Assmann and Q. Summerfield. Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. *J. Acoust. Soc. Am.*, 85(1):327–338, 1989.

[5] P. Assmann and Q. Summerfield. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 88(2):680–697, 1990.

[6] L. Bahl, F. Jelinek, and R. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE T. Patt. Anal. Mach. Intell.*, 5:179–190, 1983.

[7] L. Bahl, P. Gopalakrishnan, and R. Mercer. Search issues in large vocabulary speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition*, Snowbird, UT, 1993.

[8] J. K. Baker. The Dragon system – An overview. *IEEE T. Acoust. Speech.*, 23(1):24–29, 1975.

[9] J. Barker. *The relationship between auditory organisation and speech perception: Studies with spectrally reduced speech.* PhD thesis, University of Sheffield, UK, 1998.

[10] J. Barker, M. Cooke, and D. Ellis. Decoding speech in the presence of other sound sources. In *Proc. ICSLP'00*, Beijing, 2000.

[11] J. Barker, L. Josifovski, M. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proc. ICSLP'00*, pages 373–376, Beijing, 2000.

[12] J. Barker, M. Cooke, and P. Green. Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise. In *Proc. EUROSPEECH'01*, pages 213–216, Aalborg, Denmark, 2001.

[13] J. Barker, M. Cooke, and P. Green. Linking auditory scene analysis and robust ASR by missing data techniques. In *Proc. WISP'01*, pages 295–307, 2001.

[14] J. Barker, M. Cooke, and D. Ellis. Temporal integration as a consequence of multi-source decoding. In *ISCA Workshop on the Temporal Integration in the Perception of Speech*, 2002.

[15] J. Barker, M. Cooke, and D. Ellis. Decoding speech in the presence of other sources. *Speech Commun.*, 45(1):5–25, 2005.

[16] J. Barker, A. Coy, N. Ma, and M. Cooke. Recent advances in speech fragment decoding techniques. In *PROC. INTERSPEECH'06*, pages 85–88, Pittsburgh, PA, 2006.

[17] J. Barker, N. Ma, A. Coy, and M. Cooke. Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Comput. Speech. Lang.*, in press 2008.

[18] L. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic function of a markov process. *Inequalities*, 3:1–8, 1972.

[19] H. Bourlard. Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 1–10, Tampere, Finland, 1999.

[20] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proc. ICSLP'96*, pages 422–425, Philadelphia, 1996.

[21] A. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge MA, 1990.

[22] D. Broadbent and P. Ladefoged. On the fusion of sounds reaching different sense organs. *J. Acoust. Soc. Am.*, 29(6):708–710, 1957.

[23] G. Brown and M. Cooke. Computational auditory scene analysis. *Comput. Speech. Lang.*, 8(4):297–336, 1994.

[24] G. Brown and D. Wang. Separation of speech by computational auditory scene analysis. In J. Benesty, S. Makino, and J. Chen, editors, *Speech enhancement : What's new?*, pages 371–402. Springer, New York, 2005.

[25] G. Brown, J. Barker, and D. Wang. A neural oscillator sound separator for missing data speech recognition. In *Proc. Int. Joint. Conf. on Neural Networks*, pages 2907–2912, 2001.

[26] D. Burshtein. Robust parametric modeling of durations in hidden Markov models. In *Proc. IEEE ICASSP'95*, pages 548–551, 1995.

[27] R. Carlyon and T. Shackleton. Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms? *J. Acoust. Soc. Am.*, 95: 3541–3554, 1994.

[28] E. Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, 1953.

[29] H. Christensen, N. Ma, S. Wrigley, and J. Barker. Integrating pitch and localisation cues at a speech fragment level. In *PROC. INTERSPEECH'07*, pages 2769–2772, Antwerp, 2007.

[30] M. Cooke. *Modelling auditory processing and organisation*. Cambridge University Press, Cambridge, MA, 1993.

[31] M. Cooke. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 119 (3):1562–1573, 2006.

[32] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Commun.*, 35(3–4):141–177, 2001.

[33] M. Cooke, P. Green, and M. Crawford. Handling missing data in speech recognition. In *Proc. ICSLP'94*, pages 1555–1558, Yokohama, Japan, 1994.

[34] M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. In *Proc. IEEE ICASSP'97*, pages 25–28, Munich, 1997.

[35] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and uncertain acoustic data. *Speech Commun.*, 34(3):267–285, 2001.

[36] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.*, 120:2421–2424, 2006.

[37] M. Cooke, M. Garcia Lecumberri, and J. Barker. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. *J. Acoust. Soc. Am.*, 123:414–427, 2008.

[38] A. Coy and J. Barker. A multipitch tracker for monaural speech segmentation. In *PROC. INTERSPEECH'06*, pages 1678–1681, Pittsburgh, PA, 2006.

[39] A. Coy and J. Barker. An automatic speech recognition system based on the scene analysis account of auditory perception. *Speech Commun.*, 49(5):384–401, 2007.

[40] T. Crystal and A. House. Segmental durations in connected-speech signals: Current results. *J. Acoust. Soc. Am.*, 83(4):1553–1573, 1988.

[41] T. Crystal and A. House. Segmental durations in connected-speech signals: Syllabic stress. *J. Acoust. Soc. Am.*, 83(4):1574–1585, 1988.

[42] X. Cui and A. Alwan. Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR. *IEEE T. Speech. Audi. P.*, 13(6): 1161–1172, 2005.

[43] J. Culling and C. Darwin. Perceptual separation of simulation vowels: Within and across formant grouping by F0. *J. Acoust. Soc. Am.*, 93(3):3454–3467, 1993.

[44] J. Culling and Q. Summerfield. Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am.*, 98(2):785–797, 1995.

[45] S. Cunningham and M. Cooke. The role of evidence and counter-evidence in speech perception. In *Proc. ICPhS'99*, pages 215–218, 1999.

[46] C. Darwin. Auditory grouping. *Trends in Cognitive Sciences*, 1(9):327–333, 1997.

[47] C. Darwin and R. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *J. Acoust. Soc. Am.*, 102:2316–2324, 1997.

[48] C. Darwin and R. Hukin. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am.*, 107(2):970–977, 2000.

[49] C. Darwin and R. Hukin. Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *J. Acoust. Soc. Am.*, 108(1):335–342, 2000.

[50] C. Darwin, R. Hukin, and B. Al-Khatib. Grouping in pitch perception: Evidence for sequential constraints. *J. Acoust. Soc. Am.*, 98(2):880–885, 1995.

[51] A. de Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *J. Acoust. Soc. Am.*, 93(6):3271–3290, 1993.

[52] A. de la Torre, A. M. Peinado, J. C. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE T. Speech. Audi. P.*, 13(3):355–366, 2005.

[53] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[54] L. Deng and X. Huang. Challenges in adopting speech recognition. *Communications of the ACM*, 47(1):69–75, 2004.

[55] L. Deng, A. Acero, M. Plumpe, and X. D. Huang. Large-vocabulary speech recognition under adverse acoustic environments. In *Proc. ICSLP'00*, pages 806–809, 2000.

[56] L. Deng, J. Droppo, and A. Acero. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE T. Speech. Audi. P.*, 13(3):412–421, 2005.

[57] W. Dowling. The perception of interleaved melodies. *Cognitive Psychol.*, 5:322–337, 1973.

[58] J. Droppo, L. Deng, and A. Acero. Evaluation of the SPLICE algorithm on the Aurora 2 database. In *Proc. EUROSPEECH'01*, pages 217–220, Aalberg, Denmark, 2001.

[59] J. Droppo, A. Acero, and L. Deng. Uncertainty decoding with SPLICE for noise robust speech recognition. In *Proc. IEEE ICASSP'02*, Orlando, Florida, 2002.

[60] R. Drullman, J. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064, 1994.

[61] S. Dupont. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In *Proc. ICSLP'98*, Sydney, 1998.

[62] N. Durlach, C. Mason, G. Kidd, Jr., T. Arbogast, H. Colburn, and B. Shinn-Cunningham. Note on informational masking. *J. Acoust. Soc. Am.*, 113(6):2984–2987, 2003.

[63] D. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT, Cambridge MA, 1996.

[64] D. Ellis. Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Commun.*, 27(3–4):281–298, 1999.

[65] J. Ferguson. Variable duration models for speech. In *Proc. Symp. On the Application of Hidden Markov Models to Text and Speech*, pages 143–179, New Jersey, 1980. Princeton.

[66] H. Fletcher. The nature of speech and its interpretation. *J. Franklin Instit.*, 193(6): 729–747, 1922.

[67] H. Fletcher. *Speech and hearing in communication*. Van Nostrand Co., New York, 1953.

[68] V. Gadde. Modeling word durations. In *Proc. ICSLP'00*, pages 601–604, Beijing, 2000.

[69] V. Gadde. Modelling word duration for better speech recognition. In *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[70] M. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Comput. Speech. Lang.*, 12:75–98, 1998.

[71] M. Gales and S. Young. Robust continuous speech recognition using parallel model combination. *IEEE T. Speech. Audi. P.*, 4(5):352–359, 1996.

[72] M. Gales and S. Young. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2007.

[73] Z. Ghahramani and M. Jordan. Supervised learning from incomplete data via an EM approach. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advanced in Neural Information Processing Systems*, volume 6, pages 120–129. Morgan Kaufmann, San Mateo, CA, 1994.

[74] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. IEEE ICASSP'89*, pages 532–535, 1989.

[75] B. Glasberg and B. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Res.*, 47:103–138, 1990.

[76] J. Godfrey, E. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. IEEE ICASSP'92*, volume 1, pages 517–520, San Francisco, CA, 1992.

[77] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Commun.*, 16 (3):261–291, 1995.

[78] R. Gonzales, R. Woods, and S. Eddins. *Digital Image Processing using MATLAB*. Prentice Hall, 2004.

[79] R. C. Gonzalez and P. Wintz. *Digital Image Processing*. Addison-Wesley, Reading, MA, 1987.

[80] Y. Gotoh, M. Hochberg, and H. Silverman. Using MAP estimated parameters to improve HMM speech recognition performance. In *Proc. IEEE ICASSP'94*, pages 229–232, Adelaide, 1994.

[81] P. Green, M. Cooke, and M. Crawford. Auditory scene analysis and HMM – Recognition of speech in noise. In *Proc. IEEE ICASSP'95*, pages 401–404, Detroit, 1995.

[82] S. Greenberg and B. Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. In *Proc. IEEE ICASSP'97*, pages 1647–1650, Munich, 1997.

[83] A. Hagen, A. Morris, and H. Bourlard. Different weighting schemes in the full combination sub-bands approach for noise robust ASR. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 199–202, Tampere, Finland, 1999.

[84] S. Harding, J. Barker, and G. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE T. Audio. Speech.*, 14(1):58–67, 2006.

[85] H. Hermansky. Perceptual linear predictive (PLP) analysis for speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990.

[86] H. Hermansky. Should recognizers have ears? *Speech Communication*, 25:3–27, 1998.

[87] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE T. Speech. Audi. P.*, 2(4):578–589, 1994.

[88] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proc. EUROSPEECH'91*, pages 1367–1370, Genova, Italy, 1991.

[89] H. Hermansky, S. Tibrewela, and M. Pavel. Towards ASR on partially corrupted speech. In *Proc. ICSLP'96*, pages 462–465, Philadelphia, 1996.

[90] M. Hochberg and H. Silverman. Constraining model duration variance in hmm-based connected-speech recognition. In *Proc. EUROSPEECH'93*, pages 323–326, Berlin, 1993.

[91] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice. Implementing a gammatone filter bank. Technical report, MRC Applied Psychology Unit, Cambridge, 1988.

[92] J. Holmes and W. Holmes. *Speech synthesis and recognition.* Taylor & Francis, Inc., Bristol, PA, USA, 2nd edition, 2002.

[93] T. Houtgast and H. Steeneken. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28:66–73, 1973.

[94] T. Houtgast and H. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, 77(3): 1069–1077, 1985.

[95] G. Hu. *Monaural speech organization and segregation.* PhD thesis, The Ohio State University, Biophysics program, 2006.

[96] G. Hu and D. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE T. Neural. Networ.*, 15:1135–1150, 2004.

[97] G. Hu and D. Wang. Separation of fricatives and affricates. In *Proc. IEEE ICASSP'05*, pages 749–752, 2005.

[98] G. Hu and D. Wang. Auditory segmentation based on onset and offset analysis. *IEEE T. Audio. Speech.*, 15:396–405, 2007.

[99] F. Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64: 532–556, 1976.

[100] F. Jelinek. *Statistical methods for speech recognition.* MIT Press, Cambridge, MA, 1998.

[101] I. Johannesma, P. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Proc. Symposium on Hearing Theory*, pages 58–69, IPO, Eindhoven, Netherlands, 1972.

[102] M. Jordan. Graphical models. *Statistical Science (Special Issue on Bayesian Statistics)*, 19:140–155, 2004.

[103] L. Josifovski, M. Cooke, P. Green, and A. Vizinho. State based imputation of missing data for robust speech recognition and speech enhancement. In *Proc. EUROSPEECH'99*, pages 2837–2840, 1999.

[104] B. Juang. Speech recognition in adverse environments. *Comput. Speech. Lang.*, 5: 275–294, 1991.

[105] B. Juang, L. Rabiner, S. Levinson, and M. Sondhi. Recent developments in the application of hidden Markov models to speaker independent isolated word recognition. In *Proc. IEEE ICASSP'85*, pages 9–12, 1985.

[106] J. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am.*, 93(1):510–524, 1993.

[107] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[108] S. King, C. Bartels, and J. Bilmes. Svitchboard 1: Small vocabulary tasks from Switchboard 1. In *PROC. INTERSPEECH'05*, pages 3385–3388, Lisbon, Portugal, 2005.

[109] B. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Commun.*, 25:117–132, 1998.

[110] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath. Super-human multi-talker speech recognition: The IBM 2006 Speech Separation Challenge system. In *Proc. INTERSPEECH'06*, Pittsburgh, 2006.

[111] G. Lathoud, M. Magimai-Doss, B. Mesot, and H. Bourlard. Unsupervised spectral subtraction for noise-robust ASR. In *Proc. ASRU'05*, pages 189–194, 2005.

[112] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech. Lang.*, 9:171–185, 1995.

[113] R. Leonard. A database for speaker independent digit recognition. In *Proc. IEEE ICASSP'84*, pages 328–331, 1984.

[114] S. Levinson. Continuously variable duration hidden Markov models for speech analysis. In *Proc. ICASSP'86*, pages 1241–1244, 1986.

[115] H. Liao and M. Gales. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Commun.*, 50(4):265–277, 2008.

[116] H. Liao and M. J. F. Gales. Joint uncertainty decoding for noise robust speech recognition. In *Proc. INTERSPEECH'05*, pages 3129–3132, Lisbon, Portugal, 2005.

[117] J. Licklider. A duplex theory of pitch perception. *Experientia*, 7:128–134, 1951.

[118] J. Lim and A. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586–1604, 1979.

[119] R. Lippmann. Speech recognition by machines and humans. *Speech Commun.*, 22(1): 1–16, 1997.

[120] F. H. Liu, R. M. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. In *Proc. ARPA Speech and Natural Language Workshop*, pages 69–74, 1993.

[121] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech Commun.*, 11:215–228, 1992.

[122] N. Ma and P. Green. Context-dependent word duration modelling for robust speech recognition. In *PROC. INTERSPEECH'05*, pages 2609–2612, Lisbon, 2005.

[123] N. Ma and P. Green. A 'speechiness' measure to improve speech decoding in the presence of other sound sources. In *PROC. INTERSPEECH'08*, Brisbane, Australia, 2008.

[124] N. Ma, P. Green, and A. Coy. Exploiting dendritic autocorrelogram structure to identify spectro-temporal regions dominated by a single sound source. In *PROC. INTERSPEECH'06*, pages 669–672, Pittsburgh, PA, 2006.

[125] N. Ma, J. Barker, and P. Green. Applying duration constraints by using unrolled hmms. In *PROC. INTERSPEECH'07*, pages 1066–1069, Antwerp, 2007.

[126] N. Ma, P. Green, J. Barker, and A. Coy. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Commun.*, 49(12):874–891, 2007.

[127] D. Marr. *Vision*. W.H. Freeman, San Francisco, CA, 1982.

[128] S. McAdams and J. Bertoncini. Organization and discrimination of repeating sound sequences by newborn infants. *J. Acoust. Soc. Am.*, 102(5):2945–2953, 1997.

[129] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE T. Acoust. Speech.*, 34(4):744–754, 1986.

[130] I. McCowan, A. Morris, and H. Bourlard. Improving speech recognition performance of small microphone arrays using missing data techniques. In *Proc. ICSLP'02*, pages 2181–2184, Denver, 2002.

[131] R. Meddis and M. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J. Acoust. Soc. Am.*, 89(6):2866–2882, 1991.

[132] R. Meddis and M. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 91(1):233–245, 1992.

[133] G. Miller and J. Licklider. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.*, 22:167–173, 1950.

[134] J. Ming and F. Smith. A probabilistic union model for sub-band based robust speech recognition. In *Proc. IEEE ICASSP'00*, pages 1787–1790, Istanbul, 2000.

[135] J. Ming and F. Smith. Union: A new approach for combining sub-band observations for noisy speech recognition. *Speech Commun.*, 34:41–55, 2001.

[136] C. Mitchell, M. Harper, and L. Jamieson. On the complexity of explicit duration HMMs. *IEEE T. Speech. Audi. P.*, 3(3):213–217, 1995.

[137] A. Morris, M. Cooke, and P. Green. Some solutions to the missing feature problem in data classification, with application to noise robust ASR. In *Proc. IEEE ICASSP'98*, pages 737–740, 1998.

[138] A. Morris, J. Barker, and H. Bourlard. From missing data to maybe useful data: soft data modelling for noise robust ASR. In *Proc. WISP'01*, pages 153–164, Stratford-upon-Avon, 2001.

[139] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination for noise robust asr. *Speech Commun.*, 34:25–40, 2001.

[140] H. Ney. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE T. Acoust. Speech.*, 32(2):263–271, 1984.

[141] A. Noll and H. Ney. Training of phoneme models in a sentence recognition system. In *Proc. IEEE ICASSP'87*, pages 1277–1280, 1987.

[142] M. Omologo, P. Svaizer, and M. Matassoni. Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Commun.*, 25:75–95, 1998.

[143] K. Palomäki, G. Brown, and J. Barker. Techniques for handling convolutional distortion with 'missing data' automatic speech recognition. *Speech Commun.*, 43:123–142, 2004.

[144] K. Palomäki, G. Brown, and D. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Commun.*, 43(4):361–378, 2004.

[145] N. Parihar and J. Picone. Analysis of the Aurora large vocabulary evaluations. In *Proc. EUROSPEECH'03*, pages 337–340, 2003.

[146] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE T. Speech. Audi. P.*, 8(3):320–327, 2000.

[147] R. Patterson and B. Moore. Auditory filters and excitation patterns as representations of frequency resolution. In B. Moore, editor, *Frequency Selectivity in Hearing*, pages 123–177. Academic Press Ltd., London, 1986.

[148] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and A. M. Complex sounds and auditory images. In Y. Cazals, L. Demany, and K. Horner, editors, *Auditory Physiology and Perception*, pages 123–177. Pergamon, Oxford, 1992.

[149] D. Paul. An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model. In *Proc. IEEE ICASSP'93*, volume 1, pages 25–28, San Francisco, 1993.

[150] D. Pearce and H.-G. Hirsch. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP'00*, volume 4, pages 29–32, Beijing, 2000.

[151] K. Power. Durational modelling for improved connected digit recognition. In *Proc. ICSLP'96*, pages 885–888, Philadelphia, 1996.

[152] P. Price, W. Fisher, J. Bernstern, and D. Pallet. The DARPA 1000–word Resource Management database for continuous speech recognition. In *Proc. ICASSP'88*, pages 651–654, New York, 1988.

[153] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286, 1989.

[154] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, N.J., 1993.

[155] L. Rabiner, J. Wilpon, and F. Soong. High performance connected digit recognition using hidden Markov models. *IEEE T. Acoust. Speech.*, 37(8):1214–1225, 1989.

[156] B. Raj, R. Singh, and R. Stern. Inference of missing spectrographic features for robust speech recognition. In *Proc. ICSLP'98*, Sydney, 1998.

[157] B. Raj, M. Seltzer, and R. Stern. Reconstruction of damaged spectrographic features for robust speech recognition. In *Proc. ICSLP'00*, pages 357–360, Beijing, 2000.

[158] B. Raj, M. Seltzer, and R. Stern. Reconstruction of missing features for robust speech recognition. *Speech Commun.*, 43(4):275–296, 2004.

[159] R. Remez, P. Rubin, S. Berns, J. Pardo, and J. Lang. On the perceptual organization of speech. *Psychol. Rev.*, 101:129–156, 1994.

[160] S. Renals and M. Hochberg. Efficient evaluation of the LVCSR search space using the NOWAY decoder. In *Proc. IEEE ICASSP'96*, pages 149–152, Atlanta, 1996.

[161] P. Renevey and A. Drygajlo. Statistical estimation of unreliable features for robust speech recognition. In *Proc. ICASSP'00*, pages 1731–1734, Istanbul, 2000.

[162] P. Renevey and A. Drygajlo. Introduction of a reliability measure in missing data approach for robust speech. In *Proc. EUSIPCO'00.*, 2000.

[163] N. Roman, D. Wang, and G. Brown. Speech segregation based on sound localization. *J. Acoust. Soc. Am.*, 114(4):2236–2252, 2003.

[164] S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proc. EUROSPEECH'03*, pages 1009–1012, 2003.

[165] M. Russell and A. Cook. Experimental evaluation of duration modelling techniques for automatic speech recognition. In *Proc. IEEE ICASSP'87*, pages 2376–2379, 1987.

[166] M. Russell and R. Moore. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *Proc. IEEE ICASSP'85*, pages 5–8, 1985.

[167] M. Scheffers. *Sifting vowels: Auditory pitch analysis and sound segregation*. PhD thesis, Groningen University, The Netherlands, 1983.

[168] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. IEEE ICASSP'97*, pages 1331–1334, Munich, 1997.

[169] M. Seltzer, B. Raj, and R. Stern. A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Commun.*, 43(4):379–393, 2004.

[170] S. Shamma. Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. *J. Acoust. Soc. Am.*, 78:1613–1621, 1985.

[171] Y. Shao and D. Wang. Robust speaker recognition using binary time-frequency masks. In *Proc. IEEE ICASSP'06*, pages 645–648, 2006.

[172] M. Slaney. An efficient implementation of the Patterson–Holdsworth auditory filter bank. Technical report, #35, Apple Computer Co., 1993.

[173] M. Slaney and R. Lyon. A perceptual pitch detector. In *Proc. IEEE ICASSP'90*, pages 357–360, Albequerque, 1990.

[174] S. Srinivasan and D. Wang. Transforming binary uncertainties for robust speech recognition. *IEEE T. Audio. Speech.*, 15(7):2130–2140, 2007.

[175] S. Srinivasan, Y. Shao, Z. Jin, and D. Wang. A computational auditory scene analysis system for robust speech recognition. In *Proc. INTERSPEECH'06*, 2006.

[176] R. M. Stern. Signal separation motivated by human auditory perception: Applications to automatic speech recognition. In P. Divenyi, editor, *Speech Separation by Humans and Machines*. Springer-Verlag, 2004.

[177] R. Stubbs and Q. Summerfield. Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 84:1236–1249, 1988.

[178] Q. Summerfield, A. Lea, and D. Marshall. Modelling auditory scene analysis: strategies for source segregation using autocorrelograms. In *Proc. Institute of Acoustics*, volume 12, pages 507–514, 1990.

[179] H. van Hamme. PROSPECT features and their application to missing data techniques for robust speech recognition. In *Proc. ICSLP'04*, pages 101–104, Jeju Island, Korea, 2004.

[180] L. van Noorden. *Temporal coherence in the perception of tone sequences*. PhD thesis, Eindhoven Univeristy of Technology, 1975.

[181] M. van Segbroeck and H. van Hamme. Vector-quantization based mask estimation for missing data automatic speech recognition. In *PROC. INTERSPEECH'07*, pages 910–913, Antwerp, 2007.

[182] A. Varga and R. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE ICASSP'90*, pages 845–848, 1990.

[183] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones. The NOISEX–92 study on the effect of additive noise on automatic speech recognition. Technical report, DERA Speech Research Unit, 1992.

[184] T. Virtanen. Speech recognition using factorial hidden Markov models for separation in the feature space. In *Proc. INTERSPEECH'06*, Pittsburgh, 2006.

[185] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE T. Inform. Theory*, 13:260–67, 1967.

[186] D. Wang and G. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE T. Neural. Networ.*, 10(3):684–697, 1999.

[187] D. Wang and G. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley–Interscience, 2006.

[188] R. M. Warren, K. R. Riener, J. A. Bashford, and B. S. Brubaker. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Percept. Psychophys.*, 57(2):175–182, 1995.

[189] R. M. Warren, J. A. Bashford, and P. W. Lenz. Intelligibility of bandpass speech. *J. Acoust. Soc. Am.*, 108(9):1264–1268, 2000.

[190] M. Weintraub. *A theory and computational model of auditory monaural sound separation*. PhD thesis, Department of Electrical Engineering, Stanford University, 1985.

[191] Z. Zhang and S. Furui. Piecewise-linear transformation-based HMM adaptation for noisy speech. *Speech Commun.*, 42:43–58, 2004.