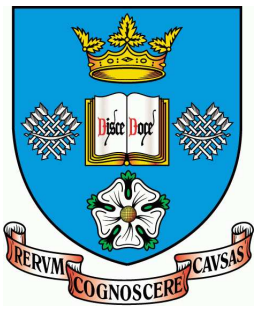
Robust Automatic Speech Recognition with Missing and Unreliable Data

Ljubomir Josifovski

Department of Computer Science
University of Sheffield, UK



August 2002

Dissertation submitted to the University of Sheffield
for the degree of Doctor of Philosophy

In memory of my loving father.

Abstract

Automatic speech recognition (ASR) systems have made dramatic performance leaps in the recent past. Yet, the notion that the key to making recognition more robust is to reduce the difference between training and test conditions is still commonly held. As ASR applications move from tightly controlled to more natural environments with a varying number of unpredictable sound sources, this assumption is becoming less and less viable. Decoding the speech source of interest while listening to several sound sources at the same time seems a more accurate description of the ASR process that suits these challenging environments. This thesis discusses the theoretical and practical issues which arise from this viewpoint. The aim is to explore the division of the problem of robust ASR into two subproblems: (a) identification/separation of the speech and noise using speech properties alone; and (b) recognition based on the resulting partial evidence. The basic assumption is that some regions of the speech time-frequency representation remain relatively unaffected by the noise, that they can be identified and that they alone are sufficient for ASR. In contrast to conventional techniques which require models of all sources in the auditory scene and their subsequent decoding even when only one of the sources is of interest, the techniques described in this thesis make no such requirement. However, they are flexible enough to use this information if it is available.

Two techniques are used to adapt a conventional Hidden Markov model (HMM) based ASR system to use partial evidence: (i) marginalisation of the state distributions, so that only the likelihood of the reliable regions is assessed; and (ii) imputation of the unreliable regions by replacing the unreliable features with a single point from the state conditional distributions. In both cases, the "counterevidence" - assessing which states are unlikely to have generated the speech underlying the unreliable regions dominated by noise - further constrains the decoding. The techniques are evaluated on the Aurora 2 connected digit recognition task, and seem to perform competitively. In the experiments, the reliable features are identified via local SNR estimates derived through stationary and adaptive on-line noise estimates. The potential of the techniques is indicated by using the clean speech to identify the reliable regions in the noisy speech, where the accuracy is maintained even at -5 dB. The simple all-or-nothing assumption (the feature is either reliable or unreliable) gives rise to a model linking the recognition and the separation as two interdependent sides of the search for the most likely explanation of the noisy data.

Acknowledgements

First I would like to thank by supervisor Phil Green who has made this work possible. His direct support through balanced encouragement, suggestions, criticism and freedom, as well as the indirect support as head of the Speech and Hearing Group (SPandH) in Sheffield which turned out such a terrific place to work and be during the course of my studies, has been invaluable.

I thank my adviser Martin Cooke who has been a source of inspiration throughout the PhD and an early pioneer of many of the ideas explored here. He also provided the most of the software for the early set of experiments.

I have benefited greatly from the interaction with other members of the Speech and Hearing Group in Sheffield. Without trying to mention each one individually, I just wish to thank Miguel Carreira-Perpiñán for lending himself available for long discussions, and Jon Barker for putting together the CASA toolkit (CTK), used in the latter set of experiments.

David Pearce from the Motorola UK Laboratories in Basingstoke has been a patient listener and supporter of our ideas, and I thank him for making possible a productive eight months internship in Basingstoke.

The work reported here would have been impossible without the support from Motorola and the University of Sheffield. The work was also supported by the Chevening scholarship of the British Council, provided by the Foreign and Commonwealth Office, and a PhD scholarship from the Ministry of Science of Republic of Macedonia.

Finally, I wish to thank my family. Thank you Petrula for the courage, patience and the unconditional support. Thank you Kalen and Vedar for bringing me such a joy and fun. Thank you mum and dad for the unwavering support in difficult circumstances.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Current Automatic Speech Recognition (ASR) systems perform acceptably in controlled environments (Baker et al., 1991; Kubala et al., 1991; Murveit et al., 1991). The performance is good enough to be deployed in commercial products. However, when used in "noisy conditions", their performance deteriorates rapidly to a point where they are unusable in practise (Agaiby et al., 1997). We refer to this as a problem of robustness of the ASR systems. Compared to the human performance in less then ideal conditions, ASR systems perform an order of magnitude worse, even when specially adapted to cope with that particular kind of degradation, as illustrated by Table 1.1 which summarises a recent comparison on various corpora and types of speech (Allen, 2002; Lippmann, 1996). Current systems seem usable on speech which is generated for machine recognition, rather than for listeners. Applications like recognition of spontaneous speech in spoken dialogue systems (especially over the phone), recording legal proceedings, taking minutes of meetings or recognition/transcription of broadcast news (Pallet et al., 1998) are still too hard. The performance figures cited in the literature vary from 2% word error rate (WER) for airplane travel system with medium vocabulary to 50% WER for large vocabulary dialog system (Comerford et al., 1997). Human WER on similar spontaneous speech is around 4%. It seems that the problem of robustness is one of the important obstacles on the way to wider deployment of speech enabled products (Sagayama and Kiyoami, 1997). In this thesis, we will use the term "robustness" to mean "robustness to added noise" or, more generally, "robustness to the presence of other sound sources".

1.2 Objectives of the thesis

It is well documented that humans can cope with unnatural and unseen degradations, seemingly without prior training or adaptation. They can ignore broad range of degradations and deletions in time and frequency domain, while still taking into account whatever information or cues are left for the recognition. Humans seem capable of utilising the partial information left in the degraded speech (Allen, 1994). This is exactly the capability that "missing data" approach researched in this work tries to utilise for improved accuracy of an realistic automatic speech recognition (ASR) system in less–then–ideal conditions. The main aim of the work reported here is to investigate whether a realistic system, working with realistic speech corpora and in realistic noisy conditions[1] is feasible, and whether it seems possible to deliver performance improvements now or in a foreseeable future. Other aims of the thesis include reviewing and consolidating previous work that might be of interest, looking for a realistic candidate techniques than could be used for speech identification, investigating the variants of the techniques for adapting the ASR system to handle partial speech, identifying the root causes for performance degradation in noisy conditions.

[1] but still disregarding the Lombard effect

Corpus	Size	Conditions	% Machine error	% Human error	Error ratio
Alphabetic	26	20 talkers, 8 listeners	5 (isol)	1.6 (cont)	3
RM	1000	null grammar	17	2	8
WSJ-NAB	5000	quiet (trained)	7.2	0.9	8
Switchboard	14000	spontaneous (tel. BW)	43	4	11
WSJ-NAB	5000	10 dB (trained)	12.8	1.1	12
WSJ-NAB	65000	close mic	6.6	0.4	16
WSJ-NAB	65000	omni mic	23.9	0.8	30
RM	1000	word–pair grammar	3.6	0.1	36
WSJ-NAB	5000	quiet (not trained)	42	0.9	47
WSJ-NAB	5000	22dB (not trained)	77.4	0.9	70
TIdigits	11	connected	0.72	0.009	80
word spotting	20 words	judgement errors	24	0.3	80

Table 1.1: Comparative summary of HSR and ASR performance (the results are from Lippmann (1996) summarised by Allen (2002))

The missing data work reported here is applied only to recognising speech in presence of *additive noise*. Other types of noises and/or degradations (convolutional noise, reverberation, etc.) are out of the scope of this thesis.

1.3 Contributions

Missing data ASR extends/adapts the monosource ASR model dominant today to a multisource auditory scene. The extension is "economical" in a sense that not all sources need to be decoded if only one is of interest. The idea has been applied to a full–blown ASR system for connected digit recognition, together with techniques for actual separation of the sound sources (however primitive they are). The main contribution of the thesis is that the complete missing data (MD) ASR chain has been tested with realistic noises, expected to be encountered in a typical application. Several techniques, like Token Dependent Noise Estimation (Section 6.4.6) and adaptive local SNR estimation (Section 6.4.2), have been developed during the course of the experiments for the purpose of obtaining a mask estimate. The SNR based separation has been tested completing the MD ASR chain in full. A new technique named bounded state based data imputation (Section 5.5.5, also pp. 86) has been developed that can be used not only to recognise, but also to reconstruct the full speech spectrum out of the partial speech. The probability of the mask estimate has also been integrated in the MD ASR system (Sections 5.3, 5.4.1) leading to improvements in the accuracy.

The missing data (MD) model separates the problem of robust ASR into two distinct parts: speech–noise separation/identification and speech modelling. This allows for use of "simulated data", so it is possible to judge which of those two underperforms in noisy conditions. The work present here points to poor speech identification, rather then poor speech modelling, as the root cause of performance degradation.

All the work reported here has been done in collaboration with other members of the Speech and Hearing Group (SPandH) in the Department of Computer Science at the Sheffield University, UK, during the course of several years. Appendix E attempts to acknowledge the contributions of the people involved in the work reported here. Any missattributions and/or lack of are solely my fault, and I would gladly correct the errors if/when notified.

1.4 Overview of the thesis

The thesis consists of seven chapters and several appendices.

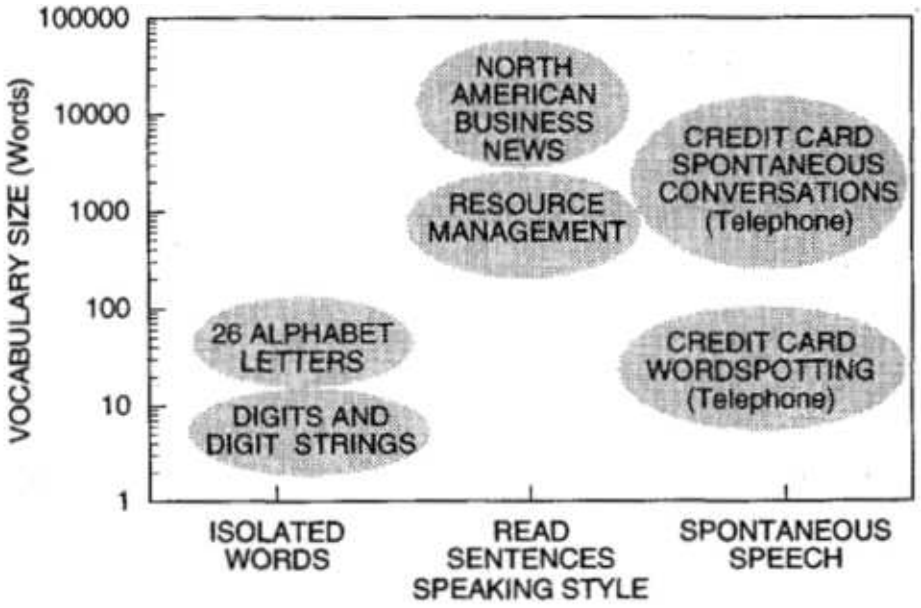


Figure 1.1: Six talker–independent speech recognition corpora used to compare humans and machines (reproduced from (Lippmann, 1996)).

Chapter 2 is a review of the other approaches to the problem of robustness in ASR taken so far. The chapter follows a fairly standard taxonomy of techniques arising from the view that lack of robustness is due to a mismatch between the conditions during training and testing of the ASR system.

In chapter 3 the original motivations and the structure of a ASR system operating under assumptions that the data is missing are lied down. It also reviews the previous MD work.

Chapter 4 discusses the problem of sound sources separation from the mixture. Section 4.2 shows how computational auditory scene analysis (CASA) may be used to tackle this problem. Section 4.4 outlines some of the techniques for local Signal–to–Noise Ratio (SNR) estimation that can be used as an (computationally cheap) alternative to CASA methods. Section 4.3 discusses the model used for general (not only for speech) blind source separation and its relation to MD.

Chapter 5 lays down the techniques used to implement the ideas underlying our current understanding of the MD process in a context of a standard HMMs–based recogniser. It introduces a model for MD ASR treating the mask as a random variable.

In the subsequent Chapter 6 the results of the experiments where a MD ASR system is applied to "standard" tests in robust ASR are discussed. Connected digits are the corpus of choice, and experiments are carried out using both some of the NOISEX noises and the Aurora 2 noisy digits databases.

The last chapter, Chapter 7, is a discussion about the relationship between the MD techniques and some of the "more established" as well as "emerging" techniques for improving the robustness of the ASR systems. It also attempts to answer some commonly asked questions about MD ASR, speculates about a possibility of a system that fully integrates the recognition and the separation parts of the system and draws the main conclusions from this work.

Chapter 2

Review of techniques for robust automatic speech recognition

2.1 Introduction

Todays ASR systems perform well enough to be deployed wide range of applications. However, moving the applications from tightly controlled to real world environments remains a challenge. The migration exposes problems rooted in our ignorance about how to model the sources of speech variability in changing acoustic environment, previously masked by the assumption of a quiet environment and single source auditory scene.

The aim of this chapter is to present and discuss various approaches that have been used to increase the robustness of the ASR systems. It briefly introduces some of the factors contributing to the speech variability first. It then moves on to concentrate on the variability due to noise. Subsequently, it introduces the notion of an *acoustic environment* to describe the interaction between the speech and the noise. Next it proceeds with classification of the techniques for improving the ASR systems robustness into three large groups. Number of techniques from each group are introduced in the following text. At places the division seems a bit artificial – but this is a common problem facing any attempt to bring under a common umbrella techniques that have evolved mainly independently and over some period of time. However, it is evident that the classification is coherent with the architecture of the todays prevalent statistical ASR systems. Towards the end of the chapter combinations of techniques as applied in real world tasks are discussed. The chapter concludes with a summary of the techniques. Appendix A summarises the performance of the techniques published in the reviewed literature. However, the value of the summary in assessing the relative merits of the techniques is limited. Until very recently, there was no de facto standard task (and corpus) to assess the techniques for robust ASR. The researchers have been using vastly different systems and speech corpora, leading to inability to compare the techniques across different research groups.

2.2 Origins of speech variability

Some common reasons for variability in speech are:

- contamination with noise (additive, convolutional, reverberation);
- speaking style (Lombard effect, speaking rate);
- inter speaker variations (voice quality, pitch, gender, dialect);
- task/context (dialogue, dictation, conversation).

Kajarekar et al. (1999)'s study on the origins of the speech variability found that the phone (with influence spreading beyond phoneme boundaries), the context and the speaker contribute most to the variability and are interdependent. A principal component analysis (PCA) of the sources of variability of the models trained on speaker–dependent speech pointed to the gender difference as the most significant factor (Kuhn et al., 1998).

Noise

Additive noise usually results from a microphone picking other sound sources in addition to the speech that is to be recognised. These are sounds generated by the office equipment, coming from the traffic on the street, etc. The human auditory system is so robust to this degradation that humans aren't aware of it most of the time. Additive noise is additive to the speech signal in the time domain and in the complex spectral domain. It can be also assumed additive on average in the power spectral domain. The noise can be *stationary* (constant) or changing with time (*non–stationary*). The short burst of noise are known as *impulsive* noise.

Convolutional noise or *linear filtering* refers to the way speech changes on its path from the source (mouth) until it is converted in digital form. The reasons are numerous – from interaction with the walls of the room to the imperfect transduction by the microphone, telephone, etc. Convolutional noise is multiplicative to the speech signal in frequency domain, hence the term linear filtering.

In a typical environment, a microphone mounted on a table in front of a speaker picks up not only the actual speech, but copies of the speech that bounced off the walls and arrived at the microphone with some latency and distorted. This is known as *reverberation*. So, the signal observed through the microphone is a sum of the original speech (coming through the direct path) and several delayed copies (although their amplitude diminishes quickly) of this speech convolved (linearly filtered) with the rooms impulse response. These secondary, tertiary... (and further) copies of the speech can not be simply considered as noise. The noise is assumed independent (and thus uncorrelated) with the speech, and exactly the opposite is true in the reverberation condition–the additive components coming from the reflections are obviously strongly dependent on the original speech. Usually algorithms based on multiple observations (arrays of microphones) are used to handle this type of degradation, but the results are far from perfect. This is one of the reasons why in most applications, users of the todays speech technology products use close–talking microphones. Humans seem to have a robust mechanism for suppression of the copies of the signals arriving up to 40ms after the initial signal, if they are not significantly louder (the precedence effect).

Other factors

The *Lombard effect* (Junqua, 1993; Junqua et al., 1998) refers to a change in the speaking style when the speaker is in a noisy environment. The speaker articulates her/his speech in such a way that it is more noise–robust for human perception. Therefore, this affects all information extracted from the speech signal (speech features) used by the present ASR systems at a great extent, hindering its performance. It is not simply the case of speaking loudly and/or slowly. Making more vocal effort changes articulation style in a complex way. The Lombard effect also makes data–gathering for robust ASR difficult: speech produced in a truly noisy environment will be different from speech produced in quiet, with noise added on later.

The speaking rates of the speakers can vary significantly too. It changes not only in response to the acoustic environment, but also because of a number of other factors.

The voice quality, gender, age, dialect are another source of variation in the speech that ASR systems have to take into account. Todays systems are usually trained on a large collection of speakers, making them speaker independent (SI). During the enrolment phase (if there is one), or using few initial sentences of the dialogues (when there isn't an enrolment phase), the system derives the speaker dependent (SD) models from the SI models by some form of adaptation of the SI models. However, for most of the present systems there usually exists a small category of

speakers for whom (due to different factors) the system exhibits exceptionally high WER – the so called "sheep & goats" phenomenon. This is another issue that hinders speech enabled products and solutions.

Trying to apply ASR systems trained on read speech to the task of ASR of spontaneous speech speech researchers have found that there is a big difference in the articulation of the speech, the pronunciation and the vocabulary of the speaker when (s)he has a task to accomplish (various dialogue systems with some aim like information retrieval, tickets reservation, etc). Continuous read speech (e.g. dictation), speech in a dialogue and conversational speech all make difference for the present recognisers. Therefore, the recogniser has to be tailored to the specific task.

The following work is mainly concerned with the variability due to additional noise. This can be considered as robustness in the stricter sense. The term "additional noise" refers to any sound in the auditory scene that is not the speech the systems attends and tries to recognise. It can be speech, as well. Techniques for handling convolutional noise will be reviewed as well, since it is also commonly encountered noise.

2.3 "Mismatch between the training and testing conditions" view to the problem of robustness

The "mainstream view" on the problem of robustness in ASR today is that performance degradation in ASR systems is due to the difference between the statistical properties of the speech they receive at their input when employed in a real-life application, and the speech used for training (estimation) of the parameters of their statistical models during system construction (Gong, 1995; Furui, 1997). This is commonly referred to as a "mismatch between training and testing conditions" view of the problem. Usually, the training conditions are: clean speech (although this need not be always the case) and speech gathered from different speakers, with different genders, speaking rates, dialects etc. (Paul and Baker, 1992; Muthasamy et al., 1992; Phillips et al., 1992; Cole et al., 1992).

Training the recogniser in matched noisy conditions

The "mismatch" description of the problem implies its solution: collecting training data in the same conditions as the testing data ("the same" in a statistical sense, that is, drawn from the same source/distribution). Therefore, no mismatch is going to occur and the ASR system is both designed and trained and used on the same type of speech. However, many factors influence variability in the real world. They are also interdependent. It is costly and difficult to put together enormous amounts of data to reflect all possible combinations of sources of mismatch.

A larger problem is that it is not certain that this approach can deliver the performance needed for applications in noise. There is a possibility that deriving models from such heterogeneous data may lead to flat models with poor discrimination performing badly in every particular, concrete condition (Lee, 1997). Recently, recognisers trained in a "multiconditional regime" on speech with added noise (four realistic noise types) at several SNRs (clean, 20dB, 15dB, 10dB, 5dB) featured average word error (WER) increase from 1.48% on clean speech to 38.29% on speech contaminated with noise at 0dB global SNR (Hirsch and Pearce, 2000). The task was connected digits recognition.[1] Even with a reasonable match between the training and the testing data, the performance remains unacceptable for application. Further, the recogniser doesn't show any further degradation in performance when tested on a data contaminated with four other (then the ones used for training) types of noise. This weakens the claim that the solution is to match the training and the testing conditions exactly. The existing techniques for noise suppression and robust ASR manage to decrease the WER significantly (Hariharan et al., 2000). But this

[1] Informal trials with small number of subjects on the same corpus indicate that the performance of human listeners just about starts to deteriorate at 0 dB SNR

particular task is fairly simple and it is difficult to say whether the widely practised techniques will scale up to a more difficult one.

Adaptive ASR

Another way to reduce the mismatch between the training and testing conditions is to make the recogniser adaptable. Ideally, the recogniser would be aware of the conditions it operates in. Upon detecting an particular acoustic condition, it should adapt correspondingly. This approach relies on (explicit or implicit) prior knowledge about the nature of the noise and how it differs from the speech it is mixed with. The usual assumptions are that the speech and the noise are independent and thus uncorrelated, and that the noise varies slowly. Unfortunately, many common environments feature reverberation and impulsive noises which fall out of this category.

Most of the recognisers today use a combination of both matched training and adaptation to improve the robustness. The training data taken represents as much of the variability as possible without sacrificing too much performance. During recognition, various adaptation techniques, that are both knowledge based (assuming some noise characteristics and how it combines with the speech in the mixture) and data driven (using speech data from the particular environment in which the recognition occurs) are usually deployed in the recogniser.

2.4 Modelling the acoustic environment

Speech models trained on clean speech are sufficient for ASR of clean speech, without additive or convolutional noise. However, in noisy conditions, the observations are result of a complex acoustic environment, with the speech source being only a part in it. The other two components are:

- the noise source(s)
- the way speech signal combines with the noise signal(s) to form the observations that are fed into the recogniser

The knowledge of the above factors, in addition to the knowledge of the speech, completely describes the auditory scene for the purposes of ASR.

Additive acoustic model

For example, if speech and noise are additive in some domain $\mathbf{x} = \mathbf{s} + \mathbf{n}$, then the p.d.f. of the noisy speech is the convolution of the p.d.f.s of the speech and the noise:

$$p_X(\mathbf{x}) = \int p_S(\mathbf{u}) p_N(\mathbf{x} - \mathbf{u}) d\mathbf{u} \tag{2.1}$$

Even techniques that don't assume explicitly the nature of the two factors above, implicitly use some knowledge about them. So they too can be accounted for in this view of the noisy environment.

Confining the modelling to speech mixed with convolutional and additive noise only, the observed noisy speech z can be expressed as a function of the speech s, the convolutional noise h and the additive noise n ($\otimes$ denotes the convolution operator):

$$x(t) = s(t) \otimes h(t) + n(t) \tag{2.2}$$

This model (in various guises[2]) has been used extensively by most of the researchers in the robust ASR in the past (Acero, 1990; Gales, 1995; Moreno, 1996; Stern et al., 1996). The power spectrum of the observed signal is:

$$|X(w)|^2 = |S(w)H(w)|^2 + |N(w)|^2 + 2|S(w)H(w)||N(w)| \cos(\theta) \tag{2.3}$$

[2]Matrouf and Gauvain (1997) use the equivalent $x(t) = h(t) \otimes (s(t) + n(t))$

where θ is the angle between the speech and the noise vectors. Assuming no prior knowledge about the θ (i.e. it is uniformly distributed in $[0, 2\pi]$), the expected value of the noisy power spectrum is:

$$\begin{aligned} \mathcal{E}\{|X(w)|^2\} &= \frac{1}{2\pi}\int_0^{2\pi} |S(w)H(w)|^2 + |N(w)|^2 + 2|S(w)H(w)||N(w)|\cos(\theta)d\theta \\ &= |S(w)H(w)|^2 + |N(w)|^2 \end{aligned} \tag{2.4}$$

Hence the frequent assumption that with enough smoothing the speech and noise power spectrums add up to the power spectrum of the noisy speech.

Often, additional assumptions are introduced in the environmental model:

- the additive noise n is independent of the speech s
- the convolutive noise h is constant over time and independent of the speech s

The first one is almost universally true. It is necessary for derivation of the p.d.f. of the noisy observations out of the speech and noisy p.d.f.s. The second assumption allows for the convolutive noise to be modelled as an additive constant in the log–spectral domain. There are models for speech and noise separation which do not make this assumption (Section 4.3). However, almost all ASR systems assume at least slowly varying gain in the log-spectral domain.

MAX model

Another common model of how the speech and the noise combine (in the feature domain) to produce the noisy speech features is the MAX model. It is applied on features in log–spectral (or log–filterbank) domain. It has been observed that in this domain increasing noise leads to gradual submergence of the less energetic speech features beneath the noisy ones (Figure 3.2). This can be modelled as if the noisy speech (feature) is:

$$\mathbf{x} = \max\{\mathbf{s}, \mathbf{n}\} \tag{2.5}$$

The principal reason for this is the logarithmic–like compression exhibited by the human hearing, and mimicked by the feature extraction process in all ASR systems. Assuming that the speech and the noise are additive in time and power spectrum domains (e.g. previous section), the compression of their sum can be approximated by a compression of the bigger of the two:

$$\log(\mathbf{x}) = \log(\mathbf{s} + \mathbf{n}) \approx \log(\max\{\mathbf{s}, \mathbf{n}\}) = \max\{\log(\mathbf{s}), \log(\mathbf{n})\} \tag{2.6}$$

Figure 2.1 depicts the MAX approximation, the correct value and the relative absolute error of the approximation for a typical range of speech and noise filterbank features. The relative error is greatest along the line $s = n$, but quickly decreases with the inverse of s (and n). In the extreme case at $(0, 0)$ the relative error reaches 100%. However, in real speech this rarely happens. Most of the time either speech or noise features have values well above zero and the approximation is useful. The effect of this environmental model on the p.d.f. of the noisy speech is discussed in Section 3.5.1.

2.5 Techniques for robust ASR

The mismatch between the training and testing conditions because of additive and convolutional noise can be reduced at several levels of the ASR system's speech processing chain. Commonly encountered approaches can be classified as (Gong, 1995):

- using inherently robust features
- speech signal "enhancement"

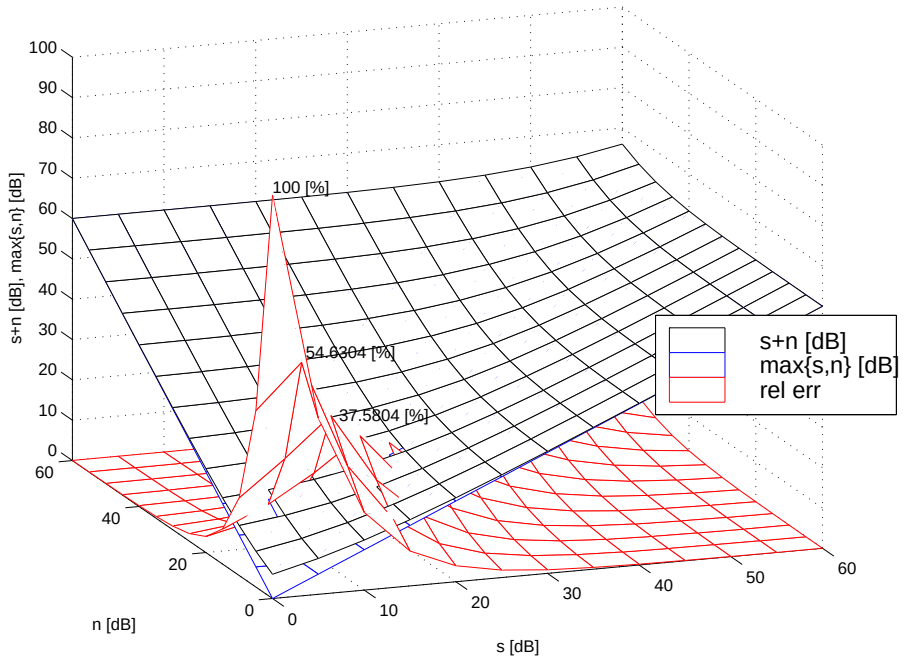


Figure 2.1: MAX approximation (in blue) compared to the correct value (in black) and the relative error (in red) for a typical range of speech (s) and noise (n) filterbank features

- speech model adaptation

Another view of the taxonomy of mismatch reduction is that the mismatch because of the noise can be considered to happen, and can be reduced in:

- the feature space - either by "enhancing" the speech features so that the noisy speech features are similar to the clean ones; or by robust feature extraction that gives similar features both for clean and noisy speech
- the model space - change the model parameters so that the changes in the features due to noise are compensated in the models (Furui, 1997).

Figure 2.2 depicts the techniques for robust ASR that are going to be reviewed in this chapter.

This classification mainly arises from the architecture of the present statistical ASR systems, where the recognition happens in two independent stages.

In the first stage, the information content of the speech signal is reduced by some transformation in such a way (intended) to preserve information considered to be "important" for the recognition, and discard the rest of it. Most often this is the "gross shape" of the spectrum. The result of the transformation is a feature vector.

In the second stage, a discrete state space is searched for the most probable path of quasi–stationary articulatory configurations that might have resulted in the observed sequence. In the speech recognition systems today this is commonly expressed as looking for a "word" W_0 out of dictionary of all "words" the recogniser can recognise with the property:[3]

$$W^* = \underset{W}{argmax} P(W|O) \tag{2.7}$$

where O are the observed features extracted in the first stage. This can be rewritten as:

$$W^* = \underset{W}{argmax} P(O|W)P(W) \tag{2.8}$$

[3]assuming isolated words recogniser

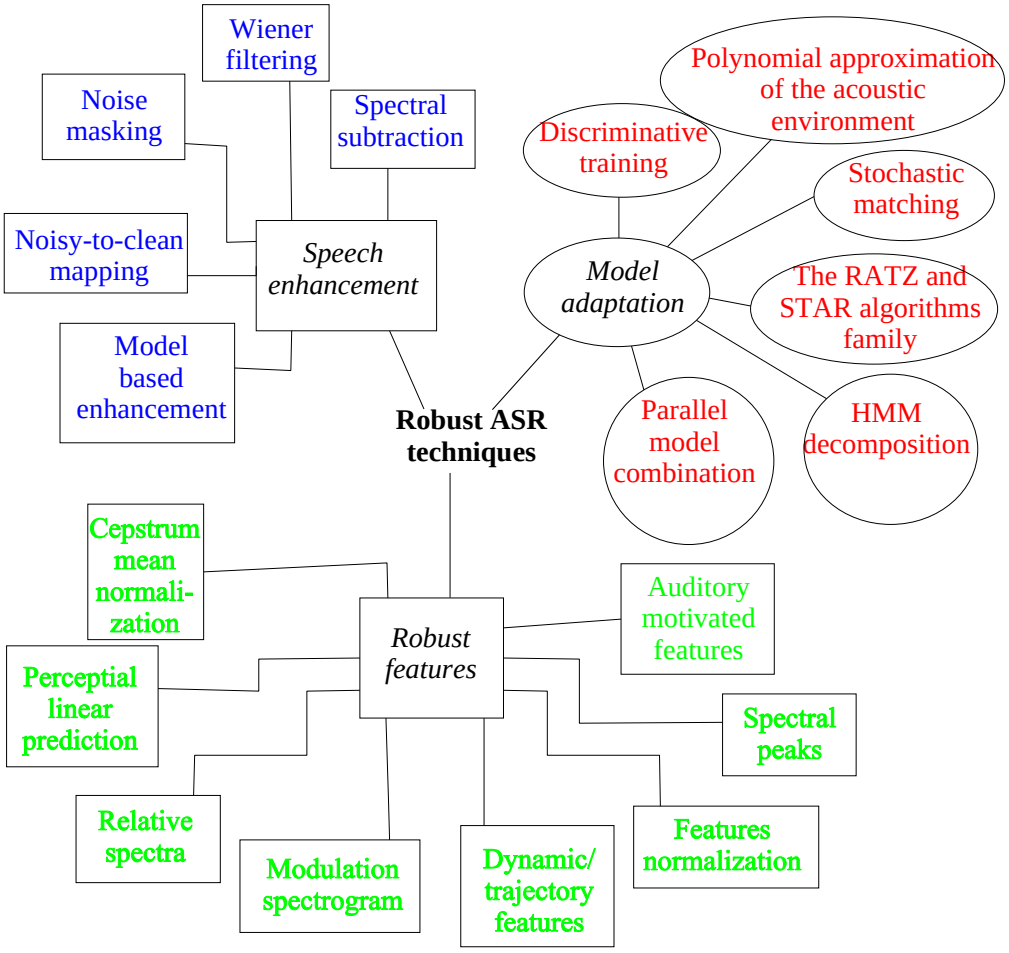


Figure 2.2: Scheme of the techniques for robust ASR

The first factor $P(O|W)$ is the *acoustical model*. It expresses the constraints about the way the articulators could have moved to give rise to a particular sequence of feature vectors O in the realization of the "word" W. The constraints are derived from hundreds of hours of speech during the system training (construction). The second factor $P(W)$ is the *language model*. It expresses the probability of the word W before any acoustic evidence is seen. It is usually derived from a huge collection of written texts.

The recogniser can generate all possible hypothesis W, compute the probability of each one given the evidence, and choose the most probable hypothesis W^* as an answer. For an isolated words recogniser this is a simple task. In a connected words recognition task, the hypothesis space may grow larger, but it can be searched efficiently and exhaustively with the Viterbi algorithm. Further, only on a small vocabulary task the recogniser can use separate model for each word. The large vocabulary tasks require modelling of the subword units (typically context sensitive phone[4]), and stringing them together according to a pronunciation dictionary for the words in the lexicon. Typically the hypothesis space grows too large for an exhaustive search. Various hypothesis pruning techniques that remove the unlikely hypothesis from the search space early are essential.

2.6 Speech enhancement

Techniques from this class aim to reduce the statistical difference between clean training and noisy testing features using some apriori knowledge about the speech, the noise and/or the way they combine. All systems today use some of the techniques from this class. Most of them originate from attempts to improve speech intelligibility. They all introduce a new problem to the robust ASR as well: the "enhanced" or "cleaned" (because they mostly deal with noise attenuation) speech maybe more intelligible to the humans, but not to the ASR systems. The degradations occurring due to "cleaning" are particularly harmful as they are non–linear, unnatural and their extent is hard to quantify in advance. On the other side, the present statistical ASR systems are quite sensitive to degradations producing data unseen during the training (the problem of statistical "outliers").

One way to limit the damage is to train the ASR system on "cleaned" clean speech so that the statistical models train to handle the degradation. Another possibility is to modify the enhancement process in such a way to balance the enhancement and the degradation of the speech.

Speech enhancement is particularly attractive when other ASR system components can not be changed.

2.6.1 Spectral subtraction

Spectral subtraction (SS) is the dominant technique today for *cleaning* the speech from the additive noise. The prerequisite is that the average noise spectrum can be estimated. This is usually done by detecting the most recent non–speech region and estimating the noise spectrum there. Various noise–estimation methods are discussed in section 4.4. The underlying assumption is that the noise will not change abruptly and will be close to its mean during this short period of time. The assumed model of the environment is Eq. (2.2) with $h(t) = 1$:

$$x(t) = s(t) + n(t) \tag{2.9}$$

Spectral subtraction takes place in the spectral domain, after a short term windowed Fourier transform is applied to the signals (Boll, 1979):

$$X(jw) = S(jw) + N(jw) \tag{2.10}$$

where $X(jw) = \sum_{k=1}^{L} x(k)e^{-jwk}$. The magnitude of the estimated clean speech $\hat{s}$ (the result of SS) equals the magnitude of the noisy speech less the average noise magnitude, while the phase is

[4] loosely defined as phoneme–like unit

the same as the phase of the noisy speech:

$$\hat{S}(jw) = [|X(jw)| - E\{|N(jw)|\}]e^{j\theta_x(e^{jw})} \tag{2.11}$$

The noise magnitude is averaged only across the sufficiently silent frames considered to be pure noise because of absence of speech activity (gaps between the words, syllables, pauses, etc). Boll (1979) proceeds further with:

- half wave–rectification of $\hat{s}$ (setting the negative values to zero, while retaining the positive ones)
- residual noise reduction by selecting the smallest magnitude value of the three adjacent (across time) values in the same frequency bin where the current amplitude is smaller then the maximal noise residual during the non–speech period
- additional noise suppression in the periods without speech in the resulting cleaned speech $\hat{s}$; the period is classified as such if $\hat{s}$ is lower then -12 dB

The original method was developed for speech enhancement, and after cleaning it proceeds further with speech reconstruction (using the phase of the noisy speech).

Although the original SS operates in magnitude domain, to justify it in a statistical sense it should be applied to the power spectral domain. As in Eq. (2.3), under the assumptions that speech and noise are independent and that the short term averages of the noise and cleaned speech describe their true values sufficiently well, the following holds (Kermorvant, 1999):

$$|\hat{S}(jw)|^2 = |X(jw)|^2 - E\{|N(jw)|^2\} \tag{2.12}$$

The enhanced speech itself suffers from unnatural coloration known as "musical noise", further speech distortion and some of the noise that was not removed. An extension known as "non–linear spectral subtraction" (NSS) introduces several parameters to balance these adverse effects (Berouti et al., 1979; Lockwood and Boudy, 1991). NSS computes $\hat{s}$ as:

$$\begin{aligned} D(w) &= G[|X(jw)|^{2\gamma} - \alpha E\{|N(w)|^{2\gamma}\}] \\ |\hat{S}(jw)|^2 &= \begin{cases} D^{1/\gamma}\ , & \text{if } D^{1/\gamma} > \beta E\{|N(jw)|^2\} \\ \beta E\{|N(jw)|^2\}\ , & \text{otherwise} \end{cases} \end{aligned} \tag{2.13}$$

The parameters α, β (overestimation factor), γ (the power spectrum exponent) and the gain G (normalisation gain introduced to compensate for the distortions caused by γ) are all hand tuned on small databases to achieve satisfactory results. In the context of an ASR system, the clean speech is itself "cleaned" with NSS to achieve further robustness to the distortions caused by the cleaning.

The NSS was further extended to include an estimate of the noise variance (in addition to the mean) and to operate in the log–spectral domain which is more suited to an ASR system (Xie and Campernolle, 1993). The minimum mean square error (MMSE) estimator of the cleaned speech given the noisy speech uses the assumed probability density functions (p.d.f.) of the speech and the noise to estimate directly the cleaned speech in the log–spectrum.

Variants of SS and combinations with other techniques

Numerous improvements to the original SS have been proposed over time. Median smoothing of the signal after the subtraction was found to reduce the "musical noise" as good as more complicated schemes without the need for manual tuning of the parameters (Linhard and Klemm, 1997). Improvements in an intelligibility test have been reported when adapting the gain $\frac{\hat{S}(jw)}{X(jw)}$ recursively (Linhard and Haulick, 1998). Singh and Srdiharan (1998) found that a critical band SS, where the noise spectrum is considered constant in all frequency bins within a critical band,

can improve the quality of the cleaned speech. Further, Virag (1995) incorporated masking across critical bands (a known property of the human auditory system) into the SS scheme, claiming improvements.

Spectral subtraction has been successfully integrated into the Parallel Model Combination (PMC) approach to model compensation (Flores and Young, 1993). Both the means and the variances of the HMM system ware compensated for the additive noise as well as for the degradation due to SS. Schless and Class (1998) used similar but simpler scheme where the musical noise was balanced with SNR dependent α and β. The SS with masking has been used in conduction with PMC (Drygajlo et al., 1995) as well. Section 2.8.1 discusses the PMC scheme for model compensation in detail.

In almost all cases, it is very hard to asses if the improvements reported would generalise to an ASR system in a particular setup. However, all real–world systems use some variant of spectral subtraction. It is important to use exactly the same SS scheme both during the training (even on clean speech) and testing, so that the models can "learn" the distortions introduced by the SS.

2.6.2 Wiener filtering

Wiener filtering is commonly used as alternative or complementary technique to spectral subtraction for removing additive noise. The filter is designed to minimise the minimum mean square error (MMSE) in time domain and is a maximal likelihood filter if the distributions of the speech and the noise are Gaussian. This assumption that the signals are quasi-stationary and distributed Normally is common in speech processing (McAulay and Malpass, 1980). Vaseghi and Milner (1993, 1997) used a Wiener filter in power spectral domain:

$$H(w) = \frac{E\{|S(jw)|^2\}}{E\{|S(jw)|^2\} + E\{|N(jw)|^2\}} \tag{2.14}$$

Then, the magnitude of the cleaned speech is:

$$|\hat{S}(jw)| = H(w)|X(jw)| \tag{2.15}$$

If the spectrum was computed from a infinite time series, the Wiener filter would be a particular case of spectral subtraction. However, the mean of the clean speech (in addition to the mean of the noise) is rarely available (except in mock experiments where the clean speech is available). Two possible ways around this are:

- assuming piecewise stationarity of the speech and independence of the speech and noise, use the mean noisy signal instead of the clean one (Stahl et al., 2000; Agarwal and Cheng, 1999)
- perform model adaptation instead of speech filtering (Beattie and Young, 1992; Vaseghi and Milner, 1993, 1997; Downey, 1996). This is particularly attractive for HMM based systems. The means of the clean speech are readily available as the means of the state p.d.f.s. The distributions are Gaussian (or mixtures of). The quasi-stationarity of the speech is ensured at state level.

Vaseghi and Milner (1997) compared Wiener filtering to SS and model adaptation on several NOISEX noises and found that Wiener filtering outperformed SS and was close to model adaptation. Downey (1996) reported on isolated digits recognition in car noise where the Wiener filter outperformed masking and SS and was as good as PMC.[5].

2.6.3 Noise masking

Noise masking is another commonly used technique for speech enhancement. It's inspired by a known effect in the auditory system where stronger signals mask the weaker ones in the sense

[5]parallel model combination, Section 2.8.1

that the weaker signal is not perceived. This can happen across the neighbouring frequency bands, across time frames, etc. The masking property is exhibited at various levels of the human audition chain as well (e.g. in the firing patterns of the neuron's response (Moore, 1982)). The end result is that the masked signal is not perceived. Examples of the effects of noise on the outputs of a standard spectral log–filterbank representation are shown on Figure 3.2. This may be one of the methods for noise suppression that human auditory system uses to enhance the local SNR. Features incorporating forward and backward noise masking in time by essentially high–pass filtering the cepstral trajectories were found to perform better both in clean and noisy speech then the cepstral features alone (Aikawa et al., 1996).

One way to simulate the masking property is to detect the frequency bands where the energy is below a certain threshold (and thus is believed to belong to the noise), and replace this value by the value of the mask for the subsequent processing (Klatt, 1976). Therefore the variance because of the noise is decreased. This was originally implemented in the context of a dynamic–time warping (DTW) recogniser. In the original scheme if the bin in the template or the observation are below the noise threshold they are replaced with it in the further calculation.

Bridle et al. (1984) introduced further refinements in the "noise marking" scheme. Depending on the relation between the observation, the noise mask level and the template mean (or the Gaussian state p.d.f. in the experiments by Varga and Ponting (1989)), the acoustic match score is computed differently for each case.

Holmes and Sedgwick (1986) used a probabilistic interpretation and extension to a HMM–based recogniser of the masking property. Assuming an environmental MAX model, the energy of the speech must be below the energy of the noise when it is masked. Therefore, the noisy bins that have masked the speech can still be used to weight against the models which have significant energy in these bins. Further, a measure of the probability that a state generated the masked speech was introduced: the area below the p.d.f. up to the noisy value. The usage of MAX model is motivated by the observation that only in small number of bins the speech and the noise will have comparable energy. In most of the bins, either the speech or the noise will dominate.

A systematic comparison of the three methods that utilise masking to achieve robustness found Klatt's method advantageous (Varga and Ponting, 1989). The following comparative summary depicts the way the state emission probability is calculated with the proposed methods that utilise masking for robustness:

Condition	Klatt (1976)	Bridle et al. (1984)	Holmes and Sedgwick (1986)
$N < O < \mu$	$\mathcal{N}(O;\mu,\sigma)$	$\mathcal{N}(O;\mu,\sigma)$	$\mathcal{N}(O;\mu,\sigma)$
$O < N < \mu$	$\mathcal{N}(N;\mu,\sigma)$	$min\{\mathcal{N}(O;\mu,\sigma),\mathcal{N}(d;\mu,\sigma)\}$	$\mathcal{C}(N;\mu,\sigma)$
$O < \mu < N$	$\mathcal{N}(N;N,\sigma)$	$\mathcal{N}(d;O,\sigma)$	$\mathcal{C}(N;\mu,\sigma)$
$N < \mu < O$	$\mathcal{N}(O;\mu,\sigma)$	$\mathcal{N}(O;\mu,\sigma)$	$\mathcal{N}(O;\mu,\sigma)$
$\mu < N < O$	$\mathcal{N}(O;N,\sigma)$	$\mathcal{N}(O;\mu,\sigma)$	$\mathcal{N}(O;\mu,\sigma)$
$\mu < O < N$	$\mathcal{N}(N;N,\sigma)$	$\mathcal{N}(d;O,\sigma)$	$\mathcal{C}(N;\mu,\sigma)$

Table 2.1: Comparative summary of the state emission probability calculation when utilising masking (after Varga and Ponting (1989))

In the Table 2.1, $\mathcal{N}(x;\mu;\sigma)$ is the state emission probability distribution function – a Gaussian with mean μ and variance σ^2, $\mathcal{C}(x;\mu,\sigma)$ is its cumulative distribution, O is the observed noisy value and d is empirically chosen constant.

Noise masking has also been successfully used together with PMC as an alternative to SS (Drygajlo et al., 1995). Mellor and Varga (1993) applied noise masking with MFCC features by masking in the spectral domain and subsequently transforming the features. Noise masking was found to perform as good as PMC down to 3dB, but worse for lower SNRs, on an isolated digits task.

2.6.4 Noisy–to–clean mapping

Another idea for achieving robustness is to find some kind of *mapping* between the noisy and clean speech features. Applying this transformation to the noisy features would yield the clean ones. In order to estimate the mapping, both clean and noisy versions of the signal are necessary. Next, the form of the functional mapping has to be determined. The choice varies from simple parametric types like linear regression (Mokbel et al., 1992) to non–parametric non–linear estimators such as multi–layer perceptrons (MLP) (Mokbel et al., 1992; Gao and Haton, 1993; Trompf et al., 1993). In order to optimise the mapping, a function measuring the similarity between the cleaned and clean features has to be selected. Typically this is the mean square error (MSE) function. Then the optimisation problem of searching for the parameters of the mapping to minimise the error criterion can be solved with standard optimisation techniques. The choice of techniques is not limited – for example, Kobayashi et al. (1993) employed an iterative procedure (using Wiener filter) for maximisation of posterior probability (MAP) of an all–pole model.

The performance of this technique is limited by the assumptions it is based on. It depends on how the chosen mapping function and error criterion suit a particular noise. So, while on the same type of noise the mapping will yield good results, for different noise types the result can be unpredictable.

2.6.5 Model based enhancement

The model based techniques explicitly assume a certain model of the clean speech and derive its parameters from data. During the enhancement process, the noisy speech, constrained by the model, is modified in such a way as to fit the model better. It is therefore considered enhanced. In a similar way to the noisy–to-clean mapping in the previous section, the nature of the model and the nature of enhancement determine the process and have to be decided upon beforehand. The main difference from noisy–to-clean mapping is that an explicit model of clean speech is assumed and the estimated from the data.

Short term spectral amplitude MMSE estimation

Ephraim and Malah (1984) derived an MMSE estimator of the short–time spectral amplitude (STSA) to enhance speech contaminated with stationary additive noise. The STSA was assumed to have normal probability density with mutually independent Fourier coefficients. The estimator was further expanded to encompass the *signal presence uncertainty*. It was found that the STSA estimate can be improved if it is conditioned on the probability of the signal being present or absent. This effectively amounts to switching between two estimators. When the signal is absent (noisy operating condition) the noise fills in the silence. The probability of presence/absence of each spectral component was assumed independent of the others. The optimal estimator of the phase under the same statistical model was derived as well. It was found to have nonunity modulus. So, when combined with the STSA estimator, the resulting estimator is no longer optimal. It was also discovered that the best phase estimator constrained to have modulus of one, is the phase of the noisy signal. This is the reason why the phase of the noisy signal is usually used when the enhanced speech is reconstructed.

Similar MMSE estimators but for log–spectral (Compernolle, 1989b) and for log–filterbank (Erell and Weintraub, 1993a) domain respectively have also been derived. They either require a noise model (Compernolle, 1989b), or estimate of the conditional distribution of the clean and the noisy speech (Erell and Weintraub, 1993a), in addition to the clean speech model. The assumed combination of speech and noise was the additive model in power spectral domain. The latter method was tested in the context of an HMM system (Erell and Weintraub, 1993b) on the RM task. MMSE of log–filterbank features outperformed the one of STSA, and conditioning on the energy gave big win for both techniques. All above mentioned estimators were reported to give increased ASR accuracy over the respective baseline both when training on clean and noisy data.

Using a–priori speech constraints

A feature enhancement scheme utilising the morphological constraints was used to compensate the cepstral features of the recogniser for the adverse influence of additive noise, stress and Lombard effect simultaneously (Hansen, 1994). Both the parameter enhancement and the stress compensation were conditioned on estimated noise mean and variance. The Lombard effect on the feature vectors fed to the recogniser was modelled as an additive bias. It was conditioned on the so called "stress class", and was itself modelled as a random Gaussian variable. The limited amount of true speech with Lombard effect was handled by parameter smoothing techniques. With all noises and all SNRs significant improvements in performance over the baseline were obtained. Hansen and Arslan (1995) used an iterative method that assumes an all pole model of speech for enhancement. All constraints were derived from the fact that human articulators are a slowly moving physical system:

- the all–pole model has to be stable
- the poles have to be at certain positions
- poles can not move too quickly from frame to frame.

The linear prediction (LP) model is one of the most commonly used speech models, since there is a prior knowledge of what the all–pole model of the speech should look like,

Yegnanarayana et al. (1999) introduced the idea of identifying the high SNR regions in the time–frequency plane and amplifying those regions correspondingly (instead of attenuation of the low SNR regions). The identification of the regions is based on the measure of the flatness of the LP spectrum. The measure draws from the ratio of energies of the linear predictors residual signal of the clean and noisy speech. The same idea was also applied to enhancement of reverberant speech (Yegnanarayana et al., 1999). Instead of SNR, a measure called Signal to Reverberant component Ratio (SRR) was optimised. The LP residual is changed depending on the estimated SRR to enhance the high SRR regions. In both cases the SNR rather then ASR accuracy was measured and was reported to improve in additive noise.

Using speech HMM for enhancement

The speech model need not necessarily be a simple one. Couvreur and Hamme (2000) used an HMM to model both the speech and the noise. The forward–backward algorithm was used to get the posterior probability of all joint $(speech, noise)$ states to have generated the noisy speech. Alternatively, it is possible to find only the most probable sequence with a Viterbi search, taking a hard decision for each frame. In either case, Parallel Model Combination (PMC – see Section 2.8.1) can be used to obtain the parameters of the composite model. Once the state sequence or posterior probabilities of the states given the noisy data are known, either a maximum likelihood (ML), or maximum posteriori (MAP) estimate of the clean speech can be generated using state conditioned Wiener filters. This complex speech model yielded significant improvements both in the quality of the enhanced speech and the accuracy of its recognition.

An autoregressive HMM (AR–HMM) system was used in the same manner for speech enhancement (Logan and Robinson, 1998). Because of the AR–HMM, the additive property of the speech and noise holds in the parameter domain too. So it is easier to derive the parameters of the composite $(speech, noise)$ states. Only Viterbi search (with hard state–to–frame alignment) was used to obtain an estimate of the cleaned speech. On a small vocabulary, speaker dependent task, the compensated system trained on clean speech approached the performance of the system trained on noisy speech.

2.7 Robust features

The term *robust features* refers to applying a transformation in the first stage of processing of the speech signal (feature extraction) that will (hopefully) result in similar, if not the same,

feature vectors both with speech used for training and speech that is to be recognised (Picone, 1993). Ideally, this should be possible regardless of the source of the variability. Because the feature vectors won't differ greatly, the subsequent processing will be the same in both cases. *Robust distance measures* may be also employed, as the feature vectors will be "similar", but not the "same". The distance measure is usually more important in the systems based on template matching via dynamic time warping (DTW), then for the HMM based ones.

Hernando and Nadeu (1991) found that using autocorrelation of the signal instead of the signal itself to fit an all pole model gives significant performance gain. This was applied together with a distance measure operating in the cepstral domain. In a similarly motivated development, in addition to autocorrelation–based features, the autocorrelation feature trajectory was filtered with a high pass filter to suppress the slowly varying components prior to computing the cepstral coefficients through a discrete cosine transform (Yuo and Wang, 1999).

Paliwal (1998) introduced spectral subbands centroid (SCC) features, and supplement cepstral with SCC features to improve the robustness. The features are related to the formants of the speech, but can be extracted easily and reliably from the power spectrum of the speech signal. The spectrum is divided in small number of sections (3 to 4), and a number of subbands fall into one section. The SCC feature is computed as:

$$C_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^\gamma(f) df}{\int_{l_m}^{h_m} w_m(f) P^\gamma(f) df}$$

where l_m and h_m are the lower and the higher edges of the m-th section of the spectrum, $w_m(f)$ is the shape of the filter, $P(f)$ is the power spectrum and γ is a constant controlling the dynamic range of the power spectrum.

In another attempt to derive a robust feature extractor, a two–sided linear predictor followed by singular value decomposition (SVD) was used to improve the resistance to additive noise (Wong et al., 1993).

All authors reported improvements over the baseline systems. However, it is not clear whether integrating different schemes results in further performance improvements.

2.7.1 Cepstral mean normalisation

Removing the slow variations out of cepstral features can be accomplished via *cepstrum mean normalisation (CMN)* technique. This simply means calculating the mean of the cepstral features over a word or sentence of speech, and removing (subtracting) the value out of the features. An alternate strategy is to employ a speech/noise detector and calculate the mean only over the speech parts. Subtraction in the cepstral domain removes the effect of convolutional noise on the signal (the same is true for log-spectrum, too). Typically this is the impulse response of the microphone. The technique is easy to implement, and effective (Stern et al., 1997). Extension of the technique, termed *segmental cepstrum mean normalisation (SCMN)* incorporates estimation of the variance (in addition to the mean) of each feature, and subsequent feature normalisation using both the mean and the variance (Vikki and Laurila, 1997). Because it is very cheap and yet effective, some variant of cepstral normalisation is part of almost every practical ASR system.

2.7.2 Perceptual linear prediction

A special form of linear predictive (LP) analysis (all pole modelling of the power spectrum) known as *Perceptual LP (PLP)* (Hermansky, 1990) appears to be more effective in obtaining noise resistant features then the ordinary LP. The central idea is to fit the poles to a warped, Mel–scale spectrum, rather then the linear one. This is in line with with our knowledge about human audition that not all frequencies are equally important (i.e. carry the same information content) for ASR. The emphasis is on a better fit at the lower frequencies, to the expense of the fit at the higher frequencies. Further, the Mel–scale introduces smoothing at the lower frequencies

reducing the need for the all–pole model to fit the fine structure of the speech (as pitch harmonics) that are unrelated to the vocal tract shape.

The technique incorporates two other properties of human hearing in further processing stages: the critical band analysis is followed by equal loudness pre–emphasis (according to the equal loudness curve) and intensity–to–loudness conversion (by taking the cubic root) before the LP coefficients are computed. A discrete cosine transform is applied to the LP coefficients to compute the final PLP features.

PLP of lower order seems to perform same or better then "ordinary" LP of higher order. PLP is one of the two (the other being Mel–frequency cepstra) feature extraction frontends in wide use in the ASR systems today (Hunt, 1999).

Kryze et al. (1999) replaced the Mel–scale filters with a hierarchical unbalanced tree of low– and high–pass filters implementing a discrete wavelet transform to improve noise robustness. The resulting transform has adaptive time–frequency resolution. Significant absolute improvement in performance was reported on clean TIMIT data mixed with car noise.

2.7.3 Relative spectra

Hermansky and Morgan (1994) devised a *representative relative spectra (RASTA)* for handling slowly varying additive and convolutional noise. The idea is to suppress any components in the speech that change more slowly or quickly then the "typical" range of speech change. As PLP, it is loosely inspired by human audition. Human perception tends to respond to a changes of the value of the input in addition to the absolute value of the input itself.

The technique can be used in conjunction with PLP (Hermansky et al., 1991). The spectral components that are obtained through the filter bank are compressed and filtered (the trajectory of the filter bank output over time is itself filtered) to suppress constant factors in each of them. The last step is all pole model estimation as with PLP. The basic technique, so called lin-log RASTA, operates in log-spectral domain. Filtering the constant/slowly varying components in this domain effectively subtracts from the signal the noise convolutional in time domain. The original IIR filter used was:

$$H(z) = \frac{0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4}}{z^{-4}(1 - 0.98z^{-1})} \tag{2.16}$$

The frequency response of the filter features a sharp zero at 0Hz, suppressing the DC component (convolutional noise in the log–spectral domain). The other two zeros are at 28.9 and 50Hz. The pole at $z = 0.98$ was latter replaced with pole at $z = 0.94$.

However, filtering in the log-spectral domain does not compensate for additive (in time domain) noise. After Hirsch et al. (1991) demonstrated that high pass filtering the envelope of the bands can be effective for additive noise removal, RASTA was applied to a linear–like domain for small spectral values and a logarithmic–like domain for large spectral values in order to compensate for both types of noise. This variant is known as J-RASTA. The effect is easily achieved by adding a small constant to the output of the filterbank before the log compression. This amounts to noise masking. Hunt (1999) argued that this is the same as using root–compression nonlinearity (for example used in PLP). Both ultimately achieve robustness by training the models with small amount of added noise.

It is notable that the numerator of the filter, $0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4}$, is essentially the transfer function of the delta–features calculation (Furui, 1986). Compared to RASTA, these features are much more selective in their frequency response. RASTA's passband is much broader. Openshaw and Mason (1996) performed similar to RASTA filtering in spectral domain instead of log–spectral domain.

Both RASTA and RASTA–PLP features appear to be effective with wide range of noises and both with additive and convolutional noise. They have also produced improved performance with reverberant speech.

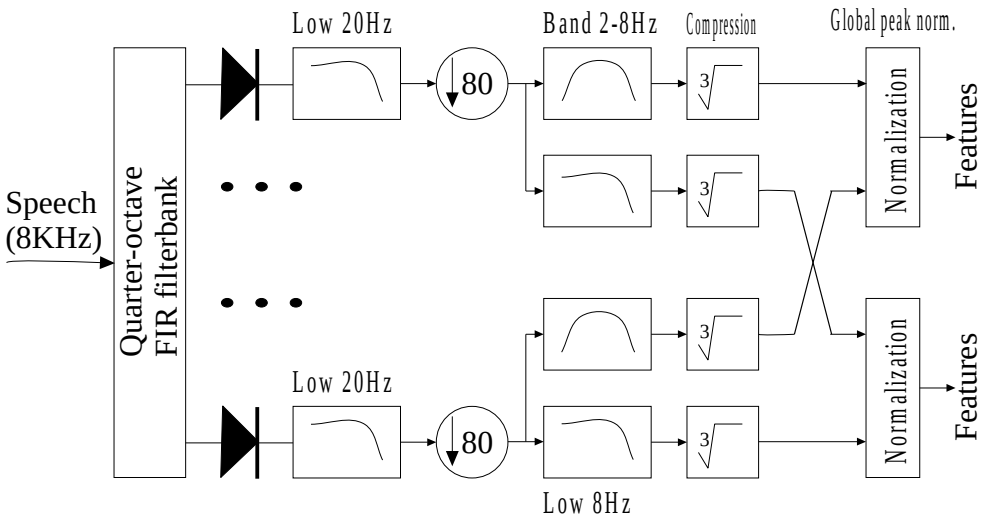


Figure 2.3: Extraction of modulation spectrogram features (after (Wu et al., 1998b))

2.7.4 Modulation spectrogram

The *modulation spectrogram* (Kingsbury et al., 1998) is another technique in the class of temporal filtering of the time trajectories of the log-filterbank outputs (performed by RASTA or dynamic features calculation, too).

Experiments on the relative importance of the modulation spectrum have indicated that components in the range of 1 to 16Hz are the primary carriers of the information required for ASR, with a dominant component around 4Hz (Kanedera et al., 1997, 1998)[6]. It seems that modulation frequencies below 2 and above 10Hz become less important in noisy speech, while those below 1Hz significantly degrade the accuracy in noisy environment. Modulation spectrogram captures information distributed over intervals of syllabic duration (100–250ms). Experiments with a hybrid HMM/MLP system demonstrated that incorporating the syllable length information either by using modulation spectrogram features and/or using wider context input window (185ms instead of the usual 105ms) significantly decreased the word error rate (WER) both for clean and reverberant speech Wu et al. (1998b).

The modulation spectrogram features are computed by analysing speech with a critical-band FIR filterbank first. Greenberg and Kingsbury (1997) half–wave rectified each output of the filterbank next, low-pass filtered it with 28Hz cut-off frequency, downsampled 100-fold, normalised by its long term average, bandpass filtered so that only the modulation frequencies between 0 and 8kHz pass through, limited with dynamic range limiter of peak 30dB and smoothed with bilinear transform at the end. Wu et al. (1998b) used a variant where the envelopes of the quarter–octave FIR filterbank outputs where low–pass filtered, downsampled, then low–pass and band–pass filtered, compressed with cubic–root compression and normalised with their global peak (Figure 2.3).

Improvements were demonstrated on reverberant speech. On a large vocabulary ASR task modulation spectrogram features have been found to carry complementary (to the PLP features) information enhancing the performance (Robinson et al., 2000).

[6] in the latter paper DFT instead of filterbank is used

2.7.5 Other dynamic and trajectory filtering features

Although not always introduced specifically with the aim of improving robustness, various features that emphasise the dynamic nature of the speech seem to increase the robustness of the ASR systems. Derivatives can remove slowly changing convolutive noise when applied in log-spectral or cepstral domain; the same is true for additive noise with the derivatives applied in spectral domain. The derivatives of the "static" features are calculated either via simple difference, or via regression (Furui, 1986). Those regression features are part of almost all ASR systems today:

$$\Delta x(t) = \frac{\sum_{n=-N}^{N} n x(t-n)}{\sum_{n=-N}^{N} n^2} \tag{2.17}$$

The difference and the regression dynamic features can be of the first, second or higher orders. Regression derivatives of 1st, 2nd and 3rd order have been found to increase the robustness of the models trained on cleaned speech and tested on Lombard speech (Hanson and Applebaum, 1990). Interestingly, inclusion of the static features neither improved nor hindered the performance. Since each additional set of derivatives significantly increases the feature vector dimensionality, and some of them are correlated, PCA can be used to truncate the feature vector by removing the redundant features (Trompf et al., 1993).

Hirsch et al. (1991) found that high pass filtering of the trajectories of the log–filterbank features features increased performance both in clean and noisy conditions. A simple IIR filter was used:

$$y(n) = x(n) - x(n-1) + 0.7y(n-1) \tag{2.18}$$

In a more general approach, features extracted from a full two dimensional Mel–cepstrum (TDMC) were used to increase the performance in clean and in noisy speech (Kitamura et al., 1992; Milner, 1996). TDMC is defined as a two-dimensional Fourier transform of Mel–scaled log spectra in the frequency and time domains. The ways of selecting an appropriate subset of TDMC features is also discussed in these papers.

More data–driven approaches have been applied recently to the task of filtering the time trajectories of the spectral parameters (Nadeu et al., 1997; Avendano and Hermansky, 1997). Nadeu et al. (1997) designed optimal filters for the time trajectories suited to a particular task/speech database. The derived filters tend to support the claim of the importance of the modulation frequencies around the syllable rate (3Hz on the database that was used). Avendano and Hermansky (1997) designed filters for speech enhancement. The criterion for optimal mapping was MMSE, and clean and noisy speech at various SNRs from the TIMIT database were used to derive the filters. It was found that:

- the filters for high SNRs were quite flat
- the filters for mid SNRs were band–pass, enhancing the modulation frequency of around 5Hz
- at low SNRs, the filters were low gain, low cut–off frequency and low–pass

2.7.6 Feature normalisation

Tibrewala and Hermansky (1998); Hakkinen et al. (1999) reported on a simple technique of on–line normalisation of feature mean and variance, effective with a wide range of noises. The statistical models of the today's ASR systems on average perform better if fed with features with roughly the same means (preferably 0) and variances (preferably 1). It is computationally cheap to normalise them with a first order recursion:

$$\begin{aligned} \mu(t) &= \alpha\mu(t-1) + (1-\alpha)x(t) \\ s(t) &= \alpha s(t-1) + (1-\alpha)x^2(t) \\ \sigma^2(t) &= s(t) - \mu^2(t) \\ \bar{x}(t) &= \frac{x(t)-\mu(t)}{\sigma(t)} \end{aligned} \tag{2.19}$$

where $x(t)$ is the feature that is used for recognition (filterbank energy, cepstral coefficient, LP coefficient) and $\bar{x}$ is the "normalised" feature fed in the recogniser. A typical value for α is $\alpha = 0.995$. The transformation is applied to each feature independently.

Tibrewala and Hermansky (1998) reported 75% decrease in word error rate on the task of recognising isolated digits with a wide range of noises. At low SNRs it was found that both the mean and the variance normalisation contribute equally to the improvement. At high SNRs normalising the mean alone was enough to achieve the improved performance.

2.7.7 Spectral peaks

While studying highly distorted sine–wave speech (SWS), (Barker and Cooke, 1997; Barker, 1998) used *spectral peaks* for robust speech recognition. SWS is speech produced by time varying sinusoids mimicking the amplitude and frequency variation of the first three formants. Tests on the Resource Management (RM) corpus showed improvement when peaks were used in recognition, regardless whether they were used during training.

Decrease of WER on a discrete–word recognition task was also reported when position and motion of the dominant spectral peaks were incorporated into a conventional Hidden Markov Model (HMM) based system (Strope and Alwan, 1998). The system detects peaks on the outputs of auditory filters with automatic gain control (AGC), groups them together into threads and smoothes the trajectory by fitting it into a second order polynomial. Again, peaks (and their derivatives) were used together with other features (cepstral coefficients and their derivatives).

It is regularly observed that the spectral peaks are less affected by noise then the "valleys" that fill with noise. Figure 2.4 shows smoothed spectrogram–like features on the left panels in clean condition (a), 20dB factory noise (b) and 0dB (c). On the middle panels (d), (e) and (f) are the corresponding spectral peaks features. It is notable that they change much less then the whole spectrum. However, many spurious peaks arise as the SNR decreases. This is due to the definition of a spectral peak employed here: in each frame, channels with energy higher then their neighbouring channels are marked to contain a spectral peak (Barker, 1998). The right panels (g), (h) and (i) show three exemplary spectral slices of the clean speech (the black line), the 20dB (green line) and the 0dB noisy speech (the blue line). As the SNR decreases, the spectral peaks are the last to be covered with noise.

2.7.8 Auditory motivated robust features

Since human audition is so robust to noise, many researchers have tried to replicate the better known parts of the human auditory chain in hope of achieving robustness.

The Ensemble Interval Histogram (EIH) (Ghitza, 1986) features were derived from a computational model of the auditory nerve–fibre firing pattern. There are 85 cochlear filters followed by level crossing counters over a finite time interval. The representation preserves fine spectral detail in low–frequency regions and fast time response in the high–frequency regions. It was compared to, and found to be more robust than an FFT derived frontend.

Hunke et al. (1998) used an auditory frontend of 120 FIR filters with frequency response derived from the solution of a 3-D cochlear hydrodynamic model. Similarly to the modulation spectrogram, the model (among other things) encompasses an automatic gain control (AGC), saturation of the magnitude at 30dB and downsampling to the rate of 100Hz so that it can be used as a plug-in replacement for a standard ASR frontend. The robustness was compared with MFCC, RASTA and J-RASTA and was significantly better (especially at lower SNRs) with the majority of the noises.

Tian et al. (1998) took a similar approach with a frontend that consisted of: FFT, intensity to loudness conversion and equal loudness correction, Mel–scaling and loudness–to–firing rate conversion. The model of Dobrin et al. (1995) went further by feeding the firing rate into a model of the central auditory system for recognition of isolated words. Gao et al. (1992) incorporated a feedback block to simulate the efferent–induced depression of the basilar membrane motion in addition to the (more or less) standard model of auditory periphery. Patterson et al. (1994)'s

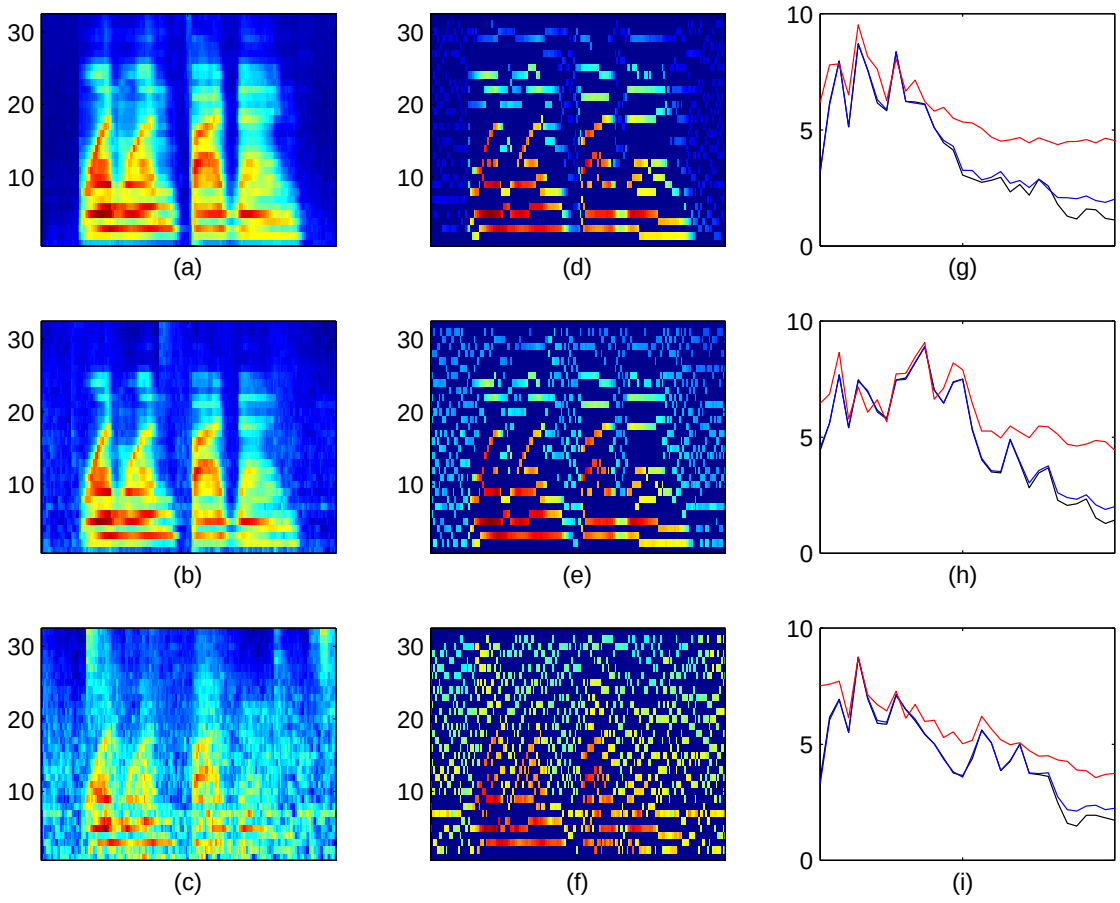


Figure 2.4: The left panels depict smoothed 32–channel spectrogram–like features of clean speech (a) and speech mixed with factory noise at 20dB (b) and 0dB (c) global SNR; the middle panels show the corresponding spectral peaks of the clean speech (d) and the noisy speech at 20dB (e) and 0dB (f); the right panels (g), (h), (i) show three spectral slices of the clean (black), 20dB (blue) and 0dB noisy speech (red).

Auditory Image Model (AIM) is binaural and mimics both the frequency analysis of the cochlea and the lateral analysis in the midbrain. The frequency and the laterality are the two dimensions of the representation. In addition, temporal integration takes place in each point in the plane, effectively adding a half dimension. The neural activity is buffered and when large peak is detected, it is added pointwise to the previous pattern stored in a static buffer. The integration stabilises periodic patterns.

Perdigao and Sa (1998) compared several auditory models with the commonly used features (RASTA, J-RASTA, MFCC, LPC) on a common task of isolated digits recognition. Almost all of the proposed auditory motivated features perform better then the "conventional" MFCC frontend[7], and especially in noise.

However, it seems that the consensus between the speech technologists is that the drawbacks of the auditory inspired models outweigh their gains and none are widely used in the practical ASR systems. For example, the AIM model generates 10000 features per frame – 3 orders of magnitude more then the MFCC frontend. A more general problem with the auditory features is that they are all fairly redundant and mutually dependent. The present statistical ASR systems are better suited to compact and independent feature representations. Smaller number of as independent

[7]DFT followed by triangular filters spaced on the Mel scale, squashing non–linearity and DCT transform

features as possible reduces the number of parameters of the system thus reducing the amount of training data needed. Further, although all are derived from the knowledge gathered while exploring the human auditory system, application of the auditory models usually requires tuning of a significant number of parameters.

2.8 Model adaptation

Techniques from this group aim to compensate the mismatch between the training and the testing conditions by suitably modifying the parameters of the models. Usually, researchers strive to make the new model parameters same or similar enough to the parameters that would have been estimated if all training data was spoken in the particular noisy condition that the recogniser is decoding at the moment. With rare exceptions, almost all techniques try to compensate for additive and/or convolutional noise, disregarding the Lombard effect. The speech is assumed to be independent of the noise. Therefore, the speech and the noise models can be inferred separately, making the techniques from this group very attractive. When the noise source changes, only the new noise model needs to be trained. This approach fits well the HMM based systems where the physical meaning of the parameters is well understood.

There are two unknown factors in the process: the statistical distribution of the noise, and how the speech and the noise combine to give the noisy observation.

The noise can be either known in advance, or it can be estimated on–line from the noisy speech. If it is known in advance, it can be modelled just as the speech is. For a fairly stationary noise, a single state model would suffice. For more complicated noises, a two state model has to be estimated. Noises requiring more then two states are rarely utilised in laboratory tests. On–line noise estimation is more attractive as in theory it adapts the recogniser to the noise at recognition time, accommodating for noises unknown at training time. However, it is much harder and less successful then off–line noise estimation. It is typically achieved with some sort of speech activity detector. Noise is estimated during the speech pauses. It is implicitly assumed that the noise will be fairly stationary and will not change significantly until the next pause. Usually only simple (mono–state) noises are estimated this way. Section 4.4 reviews various on–line noise estimation techniques in more detail.

Rose et al. (1994) treated the problems in combining the known speech and noise models to obtain a noisy model with very general acoustic environment functions (governing how the speech and noise combine together to produce the noisy speech) in detail. Specific examples, with functions like additivity in the spectral domain, additivity in log–spectral domain and the maximum in log–spectral domain were considered.

The model of the acoustic environment that is most commonly used is Eq. (2.2). It naturally arises from the physics of the sound. However, it is not the only one. For the purposes of spectral subtraction in magnitude domain additivity in the spectral magnitude domain has been assumed (Section 2.6.1). It has also been noticed that in the log–spectral domain, for the case of additive noise, the MAX approximation (observed speech being maximum of the speech and the noise) holds pretty well and simplifies the speech and noise combination (Nadas et al., 1989; Rose et al., 1994; Varga and Moore, 1990; Holmes and Sedgwick, 1986; Gales, 1997). This model and its implications will be discussed in more detail in Chapter 3.

In many approaches there is not a strict divide between model and feature compensation. Since the parameters of the used models (notably HMMs with Gaussian functions for state p.d.f.) have straightforward interpretation in relation to the features (i.e. they are the means and the variances of the features), the computed compensation factors can be applied in the feature, as well as in the parameter domain (Moreno, 1996; Beattie and Young, 1992; Vaseghi and Milner, 1993; Downey, 1996).

2.8.1 Parallel model combination

Parallel model combination (PMC) (Gales, 1995) is a popular technique for compensation of the effects both of additive and convolutional noise. The noise has to be known in advance (this is more often the case), or estimated on–line. The means alone, or the means and the variances of the speech models can be compensated. The technique allows for compensation of the mixture weights (in the iterative version) as well as the number of mixtures (but this is rarely used).

The assumed acoustic environment model is Eq. (2.2). Since the equations involved don't have a closed form solution, there are several PMC variants depending on the assumptions made in order to obtain an approximate solution. We will consider the case of PMC in additive noise. The convolutive noise in spectral domain amounts to additive noise in the log–spectral (and cepstral domain). If both the speech and the noise are distributed normally, their sum is going to be distributed normally, too, and the compensation is straightforward. The only inconvenience is when both are modelled with Gaussian mixtures. Then the number of noisy mixtures for the noisy speech is going to be a product of the numbers of mixtures of the speech and the noise models. Fortunately, a single Gaussian is usually sufficient to model the noise.

The simplest case of PMC with additive noise is when both the speech and the noise are single Gaussians. The *non–iterative PMC* assumes that the distribution of the corrupted speech is going to be Gaussian, too[8], and that the maximal likelihood state alignment (in clean speech) obtainable via Viterbi search is not going to change because of the noise.

If the feature parameters are cepstral, then models parameters (means and variances) have to be mapped back to log–linear domain (superscript *cep* denotes cepstral, *log* log–spectral and *lin* spectral domain, and subscript *pmc* denotes the compensated speech parameters, s clean speech parameters and n noise parameters):

$$\begin{aligned} \mu^{log} &= C^{-1}\mu^{cep} \\ \Sigma^{log} &= C^{-1}\Sigma^{cep}(C^{-1})^T \end{aligned} \tag{2.20}$$

where C is the discrete cosine transform (DCT) matrix with $C_{i,j} = \cos(i(j-0.5)\pi/B)$ (B is the number of filterbank channels). Since usually less then B cepstral coefficients are retained, there are not enough to correctly reconstruct the log–spectral parameters. Zeros are appended instead, with small loss in accuracy (the higher cepstral coefficients are truncated because they carry little information useful for ASR).

One possibility for evaluation of the moments of the noisy speech distribution is via numerical integration. The compensated mean in the log domain can be estimated as:

$$\mu^{log}_{pmc} = \mu^{log} + \mathcal{E}\{\log(\exp(s^{log}) + \exp(n^{log}))\} \tag{2.21}$$

One efficient approximation is shown in (Gales, 1995).

Another possibility is to assume that the sum of two log–normally distributed random variables is a random variable log–normally distributed itself. In that case (Gales, 1995):

$$\begin{aligned} \mu^{lin}_{pmc} &= \mu^{lin}_{s} + \mu^{lin}_{n} \\ \Sigma^{lin}_{pmc} &= \Sigma^{lin}_{s} + \Sigma^{lin}_{n} \end{aligned} \tag{2.22}$$

Then the compensated parameters in the log–spectral domain are:

$$\begin{aligned} \mu^{log}_{pmc,i} &= \log(\mu^{lin}_{pmc,i}) - 0.5\log(\frac{\Sigma^{lin}_{pmc,ii}}{(\mu^{lin}_{pmc,i})^2} + 1) \\ \Sigma^{log}_{pmc,ij} &= \log(\frac{\Sigma^{lin}_{pmc,ij}}{\mu^{lin}_{pmc,i}\mu^{lin}_{pmc,j}} + 1) \end{aligned} \tag{2.23}$$

The third possibility is to ignore the variance of the speech as well as the noise and straightforwardly compensate the mean:

$$\mu^{log}_{pmc} = \log(\exp(\mu^{log}_{s}) + \exp(\mu^{log}_{n})) \tag{2.24}$$

[8]This is a poor assumption as it is most often bimodal (Gales, 1995; Moreno, 1996)

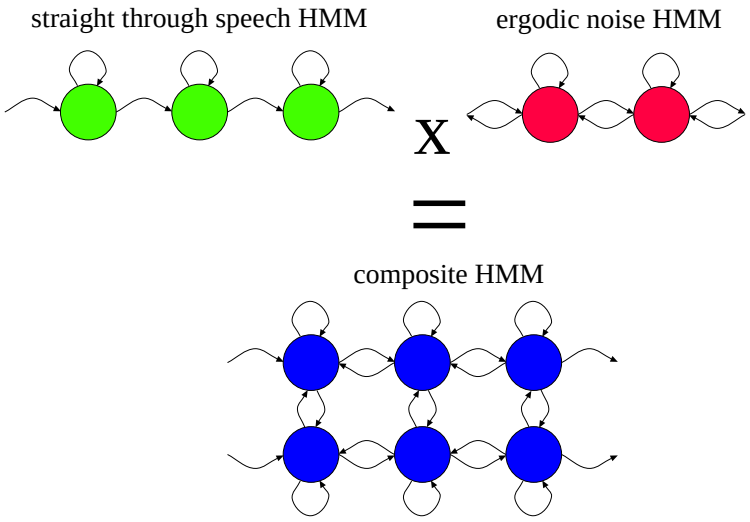


Figure 2.5: Composition of models

Once compensated, log–spectral parameters are mapped into the cepstral domain with:

$$\begin{aligned} \mu^{cep}_{pmc} &= C\mu^{log}_{pmc} \\ \Sigma^{cep}_{pmc} &= C\Sigma^{log}_{pmc}C^T \end{aligned} \tag{2.25}$$

Those parameters are used for recognition of the noisy speech.

The non–iterative PMC is fast and efficient. It has been extended to compensate multiple Gaussian mixtures per state (Yang and Haavisto, 1995) and simple (Gales, 1995) or more complex (Yang et al., 1995; Crafa et al., 1998) dynamic parameters (differences or regression parameters) of the first and second order. For the cases where a multistate noise model is required, the speech and the noise models can be concatenated into a composite speech–noise model (Fig. 2.5) (Martin, 1993).

The *iterative PMC* version (Gales, 1995) doesn't assume a single maximum likelihood alignment that is not going to change. Instead, much like in the Baum–Welch HMM training, each feature vector can be generated by every state with a certain probability. Thus, the assumption about one–to–one mapping between the clean and the noisy speech Gaussians in the mixture is relaxed, and a maximum likelihood fit for the Gaussian mixture given the noisy data can be sought[9]. Therefore the modelling capability is greatly improved. The problem of having no closed form solution for the compensation equations is exaggerated here since the numerical solution can not be used – it would involve computationally costly multidimensional integration. Data–driven PMC (DPMC) (Gales, 1995) circumvents the problem by drawing (generating) sufficient number of clean speech and noise samples from their respective distributions, combining these two sets with the assumed model of acoustic environment, and then using the noisy samples for a standard ML estimate of the parameters of the state p.d.f. Gaussian mixture. Crafa et al. (1998) introduced a Bayesian variant of DPMC which relies on a prior general noisy speech model. It uses a weighted combination of the means (and variances) of the apriori general noisy speech model (trained on speech with added noise) and the means (and variances) estimated via DPMC. The advantage is that less samples need to be drawn via DPMC in this case (for the same accuracy in noisy distribution estimation).

There have been number of extensions and applications of the basic technique to different ASR systems setups. Docio-Frnandez and Garcia-Mateo (1998) addressed the problem of lack of enough

[9]usually the noisy speech distribution retains the same number of mixtures as the clean speech for practical reasons

data for robust estimate of the noise model. A library of noises in the form of Gaussian mixtures was trained off–line. In the two variants of the method, either the on–line noise data was used to select a few of the Gaussians from the library with the highest posterior probabilities, or, the library "model" (in the form of single mixtures–single states with ergodic "grammar") was incorporated into the Viterbi search like a model to be matched in the beginning of each sentence. Flores and Young (1993) compensated the distortion of the speech caused by non–linear SS within the PMC framework. Plain SS has been used in conjunction with PMC (Schless and Class, 1998), as well as SS with parameters adapted according to the auditory noise masking thresholds (Drygajlo et al., 1995). Selective PMC compensation, where only the unreliable features with local SNR (estimated vie spectral subtraction) below a certain threshold were compensated was shown to perform better then plain PMC in a noisy digit recognition task (Renevey and Drygajlo, 1999).

2.8.2 HMM decomposition

HMM decomposition (Varga and Moore, 1990) is a general technique for simultaneous recognition of any number of signal sources modelled with a discrete state space model. It is a Viterbi search through the expanded joint N–dimensional state space of all N models. Figure 2.6 illustrates the search for the case of two sources. The "recogniser" Eq. (2.8) can in this case be expanded as:

$$(W1_0, W2_0) = \underset{(W1,W2)}{argmax} P(W1, W2|O) = \underset{(W1,W2)}{argmax} P(O|W1, W2)P(W1, W2) \tag{2.26}$$

In each joint state in the expanded space the probability of a particular observation[10] (observed in that frame) has to be computed. The computation depends on the function of the acoustic environment, that governs how the sources combine together to produce the single observation. For example, if the assumed environment is Eq. (2.2), the joint distribution of combined state can be obtained via PMC. For a general "mixing operator" $\otimes$ the probability of observation O is (Varga and Moore, 1991) (the subscript denotes the source number):

$$P_1 \otimes P_2(O) = \oint_{\mathcal{C}} P(O_1, O_2) \tag{2.27}$$

where $\mathcal{C}$ is the contour of all couples (O_1, O_2) such that $O = O_1 \otimes O_2$.

It has already been noted that in the case of a log–filterbank features and additive noise model in the spectral domain the max operator is a good approximation (Holmes and Sedgwick, 1986; Nadas et al., 1989). Because of the compression the log function exhibits, it can be approximated with: $\log(O_1 + O_2) \approx \log(\max\{O_1, O_2\})$. Therefore for model decomposition (Varga and Moore, 1991):

$$\begin{aligned} P(O) &= P(\max\{O_1, O_2\}) = P(O_1 < O_2)P(O_2) + P(O_2 < O_1)P(O_1) \\ &= \mathcal{C}(O_2, \mu_1, \sigma_1)\mathcal{N}(O_2, \mu_2, \sigma_2) + \mathcal{C}(O_1, \mu_2, \sigma_2)\mathcal{N}(O_1, \mu_1, \sigma_1) \end{aligned} \tag{2.28}$$

where $\mathcal{N}(O, \mu, \sigma)$ is the Normal distribution and $\mathcal{C}(O, \mu, \sigma)$ is the cumulative probability function of the normal distribution $\int_{-\infty}^{O} \mathcal{N}(x, \mu, \sigma)dx$.

Kadirkamanathan (1992) tested the max operator and two alternatives: a three piece linear approximation and non–iterative PMC with log–normal approximation (Eqs. (2.22) and (2.23)) (Gales and Young, 1992). They were compared on an isolated digits recognition task with artificially added noise. It was concluded that the three piece approximation performed somewhat better then the max operator, and that the PMC adaptation offered no advantage, especially at lower SNRs. The model (de)composition technique is a general search for the most likely explanation when multiple concurrent independent processes influence the observation. Takiguchi et al. (2000) used the technique to model different acoustic transfer functions arising when the speaker is in various positions relative to the microphone. Three positions were considered, and were assigned

[10] can be naturally expanded to handle multiple observations, as well

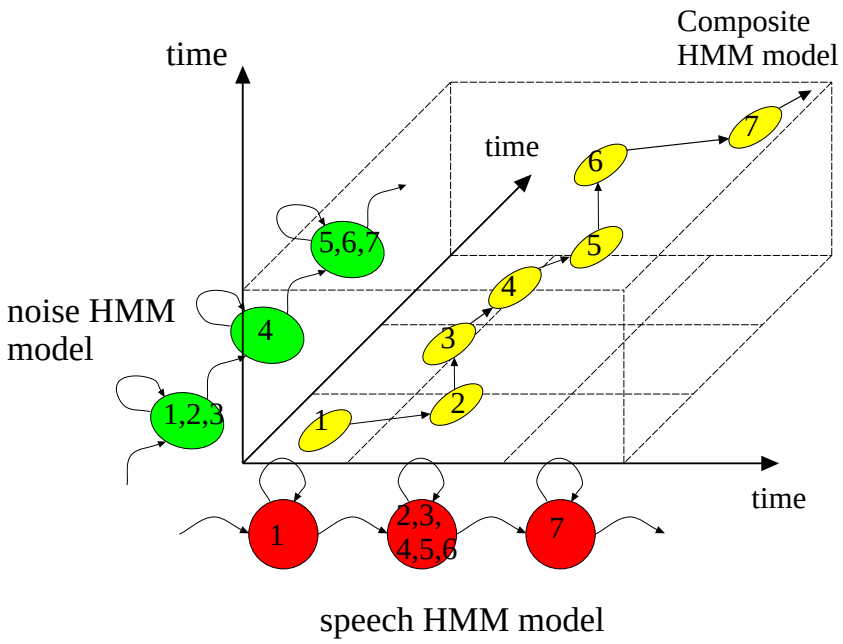


Figure 2.6: Decomposing an observed sequence

to the states of an ergodic HMM. The recognition was carried out with model decomposition, explaining the observed acoustics in terms of the words uttered and the position of the speaker.

Model decomposition can be used to infer both the speech and the noise model from speech contaminated with noise. The only hurdle is that the training is much slower because of the combined models space. The ML formulae for the forward–backward algorithm have been derived repeatedly by Kadirkamanathan and Varga (1991) and Graciarena (2000). Both used the MAX environmental model. Kadirkamanathan and Varga (1991) experimented with multistate noise models and found that much of the data is explained by the noise, making the speech models sharper. Graciarena (2000) experimented with a single state noise model and noted an improvement, over using a clean speech model together with noise model derived from manually selected "noise only" portions of the database.

Although it has not been attempted (because the combined models space expands exponentially with the number of models involved, and so does the computational cost), it is possible to infer models that explain the observations in multiple dimensions like speech (the words spoken), gender, speaking rate, distance and direction to the microphone, etc. Any number of sources of variability can be thrown in. The mixing function on a state level must be chosen appropriately to explain how they combine to produce the observed data. The sources of variability could be "peeled off" from the observations this way, instead of bundled together. This may result not only in sharper models, but in extraction of models that can be combined independently to match the conditions during the recognition (e.g. gender, speaking rate, etc.).

2.8.3 The RATZ and STAR family of algorithms

The Multivariate–Gaussian–Based Cepstral Normalisation (RATZ) Statistical Reestimation (STAR) (Moreno, 1996; Stern et al., 1996) algorithms refer to the same idea applied in features space (RATZ) or in the models space (STAR). The assumed acoustic environment model is Eq. (2.2). It is rewritten as:

$$\mathbf{x} = \mathbf{s} + \mathbf{g}(\mathbf{s}, \mathbf{h}, \mathbf{n}) \tag{2.29}$$

where

$$\mathbf{g}(\mathbf{s},\mathbf{h},\mathbf{n}) = \mathbf{h} + 10\log_{10}(1 + 10^{\frac{\mathbf{n}-\mathbf{s}-\mathbf{h}}{10}}) \tag{2.30}$$

is the additive environmental bias, and x, s, h, n are the noisy speech, the clean speech, the convolutive noise and the additive noise in the log–power domain. It is assumed that the distribution of the noisy speech is Gaussian, and the mean and the variance of its distribution can be expressed as (Moreno, 1996):

$$\begin{aligned} \mu_x &= \mu_s + \mathcal{E}\{\mathbf{g}(\mathbf{s},\mathbf{h},\mathbf{n})\} \\ \Sigma_x &= \Sigma_s + \mu_s\mu_s^T + \mathcal{E}\{\mathbf{g}(\mathbf{s},\mathbf{h},\mathbf{n})\mathbf{g}(\mathbf{s},\mathbf{h},\mathbf{n})^{\mathbf{T}}\} + \mathcal{E}\{\mathbf{s}\,\mathbf{g}(\mathbf{s},\mathbf{h},\mathbf{n})^{\mathbf{T}}\} - \mu_x\mu_x^T \end{aligned} \tag{2.31}$$

Both compensated parameters can be expressed as a sum of the parameters of the clean speech distribution plus correction factors. Since there are no closed form solutions, the distributions of the noises h and n are assumed to be Gaussian and independent too. Then the integrals are approximately computed by drawing sufficient number of samples from the speech and the noise(s) distributions (i.e. using Monte–Carlo methods, the same way as DPMC in Section 2.8.1).

The correction factors are computed for each Gaussian in an HMM system with Gaussian mixtures as state p.d.f.s. The number of the Gaussians and their apriori probabilities (the mixture weights) stay the same. If a stereo (paired clean and noisy) data is available, the correction factors can be computed directly by using the state/frame aligned data. Without stereo data, when it is not known which state generated which frame, an iterative Expectation Maximisation algorithm is derived and utilised.

After the correction factors are found, the hypothesised distribution of the noisy speech is fully known. STAR then proceeds with the recognition of the noisy speech. RATZ compensates the feature vectors, therefore, it has derive a single estimate $\hat{\mathbf{s}}$ of the clean speech $\mathbf{s}$ when the noisy one $\mathbf{x}$ is observed. It was chosen to obtain the MMSE estimate $\hat{\mathbf{s}}_{\mathbf{MMSE}} = \mathcal{E}\{\mathbf{s}|\mathbf{x}\}$.

In addition to blind and stereo variants of RATZ and STAR, there are also several other variants like interpolated, SNR dependent, etc, differing in the amount of the assumed prior knowledge about the acoustic environment.

2.8.4 Polynomial approximation of the acoustic environment function

The intractability of the non–linear mismatch function (Eq. (2.30)) between the clean and the noisy speech in the "usual" model of the acoustic environment (Eq. (2.2)) was the reason for employing various approximations and numerical simulations to calculate the distribution of the degraded speech. Another way to tackle the evaluation problem is to use a polynomial series expansion of the mismatch function (Eq. (2.30)) close to the points of interest, thus easing the computation of the noisy speech distribution while retaining reasonable accuracy of approximation (Moreno, 1996; Kim et al., 1998; Raj et al., 1997).[11]

Moreno (1996) carried out a Vector Taylor series compensation (VTS) in the log–power domain using first and second order expansion. The vector Taylor series approximation of the mismatch function in the neighbourhood of $\mathbf{s_0}$ is (Moreno, 1996):

$$\mathbf{g}(\mathbf{s},\mathbf{h},\mathbf{n}) \approx \mathbf{g}(\mathbf{s_0},\mathbf{h},\mathbf{n}) + \mathbf{g}'(\mathbf{s_0},\mathbf{h},\mathbf{n})(\mathbf{s}-\mathbf{s_0}) + \frac{\mathbf{1}}{\mathbf{2}}\mathbf{g}''(\mathbf{s_0},\mathbf{h},\mathbf{n})(\mathbf{s}-\mathbf{s_0})^{\mathbf{2}} + \ldots \tag{2.32}$$

where the first derivative of the vector function $\mathbf{g}(\mathbf{s},\mathbf{h},\mathbf{n})$ with respect to the vector $\mathbf{s}$ evaluated at $\mathbf{s_0}$ is a diagonal matrix, the second derivative is a diagonal 3-D *tensor*, etc... In the "blind" case the adaptation has to be iterated in EM fashion since the state/frame alignment is not known in advance.

Raj et al. (1997) estimated the moments of the noisy speech similarly to VTS, and then these moments are used to fit a straight line approximation of the environment function in a method called Vector Polynomial Approximations (VPS). In addition to model compensation, it was also used for speech enhancement – the MMSE estimate of the clean speech given the noisy speech

[11] Couvreur and Hamme (2000) used the same technique in the context of model–base speech enhancement

and its (derived by compensation) distribution was calculated. Kim et al. (1998) applied the technique in cepstral domain. First order blind VTS with assumed constant convolutional noise **h** and normally distributed additive noise **n** was tested on isolated and connected word recognition task with artificially added white car noise at various SNRs. It was found to perform better then PMC with log–normal approximation, especially at lower SNRs.

2.8.5 Stochastic matching based methods

It has already been mentioned in the Sections 2.8.1, 2.8.3 and 2.8.4 that in the "blind" case, when no state/frame aligned data is available, one can resort to iterative model/state alignment followed by parameter reestimation. This was termed stochastic matching by Shankar and Lee (1995). It is a general technique applicable with any (assumed) model of the acoustic environment. This model need not have a physical meaning (like Eq. (2.2)), but merely a convenient form to allow for derivation of the update formulas for the unknown parameters. One of most popular forms (due to its simplicity) for a mismatch function is the linear regression function (Shankar and Lee, 1995; Siohan et al., 1995). Other functions that have been used include projection–based likelihood measure (Chien et al., 1998) as well as various non–linear functions (Wong and Shi, 1998).

This method is widely used for *speaker adaptation*, i.e. modification of the models to model the current speaker better. Usually, a small amount of speaker dependent (SD) data is available and the task is to adapt the speaker independent (SI) models using the SD data. In addition to maximising the likelihood (ML criterion), other criteria can be optimised like the maximum posteriori probability (MAP) or the minimal classification error (MCE).

In general, the methods from this class tend to be used in addition to other noise robustness measures, rather then on their own.

2.8.6 Discriminative training

It is possible to improve the robustness of the recogniser by improving the discrimination between the units/models of the recogniser. This is typically achieved by *discriminative training* (Mizuta and Nakajima, 1992). Since in mismatched conditions the observations are "outliers" to (not drawn from) the original speech distribution, one way to limit the damage is to make the borders between units/models as wide as possible to decrease the possibility of misrecognition. The discriminative training objective function not only increases the likelihood of the correct classification, but also decreases the probability of misclassification. The optimisation method is usually some form of Generalised Probabilistic Descent (GPD) on the error function. Chu and Zhao (1998) paired discriminative training with a SNR dependent weighting of the dynamic features. Merhav and Lee (1993) explicitly assessed the sensitivity of the models to the conditions mismatch through a generalised likelihood ratio test. This test asymptotically achieves exponentially decaying probability of error for the worst–case mismatch condition.

The discriminative training aims to achieve maximal a–priori (before any noise is seen) robustness to mismatched conditions and it can be easily complemented with other techniques for improving the robustness to noise.

2.9 Combinations of techniques in real systems

In real–world applications different techniques are usually combined in the same system to achieve maximal effect.

Mokbel et al. (1997) used several techniques at different levels of the recognition process in an ASR system for telephone speech (mobile and fixed). In the feature space:

- cepstral normalisation was used to remove the long term channel effects
- cepstral feature trajectories were high pass filtered

- adaptive filtering implementing blind equalisation in the frequency domain was employed to minimise the MSE between the long term spectrum of the speech coming to the recogniser and the "typical" long term spectrum of the speech
- spectral subtraction was carried out as a next step in the chain to remove the additive noise

An on–line noise estimate was (re)estimated during the speech pauses. For this purpose, a five state model (silence, speech presumption, speech, plosive or silence, possible speech continuation) with transitions conditioned on the ratio between the short– and long–term frame energy was used. An end–point detector was also employed to ease the dialog management. The HMM model parameters were adapted as well. The optimisation criterion was MAP, alleviating the problem of insufficient adaptation data. The mapping from the original to the compensated models space was linear regression. The lack of adaptation data was further handled by clustering the Gaussians of the distributions to use the same data. At the higher level, several garbage models were trained on non–speech and out–of–vocabulary speech data. Significant performance gains were reported, making the system operate on both land and cellular phone lines speech.

Similarly, Compernolle (1989a) used spectral subtraction and automatic gain control. The noise estimator used histograms based (see Section 4.4 in Chapter 4) speech/noise discriminator. A small amount of noise was also used to mask the environment–dependent residuals. Spectral subtraction and thresholding (effectively masking), noise robust acoustic representation and noise–robust spectral distortion measures were used in (Bateman et al., 1992). RASTA filtered cepstral–time matrices and garbage models improved the performance of a telephone based system for town names recognition (Azzopardi et al., 1998). Spectral subtraction, features based on autocorrelation and spectral shaping, multiple microphones for spectrum equalisation and multiple models were used in a ASR system for car environment (Nakamura et al., 1993).

In all cases, the error rates have dropped, as expected. However, no end–user studies have been reported on how the techniques affect the usability of the automated systems in challenging conditions, and whether the systems actually achieve a satisfactory level of performance.

2.10 Summary

A variety of techniques aimed at increasing the robustness of the ASR systems in noisy conditions were reviewed in this chapter. In all cases the techniques employed alleviate the corruptive influence of the environment on the system performance, but rarely manage to bring it to the level of performance achieved in good conditions. Table A.1 in Appendix A is an incomplete list of improvements in the accuracy of various ASR systems with proposed techniques for robust ASR.

We are far from a comprehensive framework that would encompass and account for all the sources of speech variability in the real–life applications. The present methods also suffer from a variety of problems:

- model based schemes rarely compensate all parameters due to lack of data and computational costs;
- incorporating the noise variance in the models increases their variance and decreases the discriminability;
- feature compensation techniques lack the benefit of piecewise stationarity imposed by the models;
- the "inherently robust" features are robust to some noises and much less robust to others;
- the prevailing cepstral features tend to smear the noise (which in most cases is quite localised in the time–frequency plane) over all features in the feature vector.

It seems that although the robustness of the recogniser does improve when one or more of the above reviewed techniques are incorporated in it, the level of performance is not good enough for

many (envisaged) applications, and at present it is at least two orders of magnitude worse then what humans achieve in the same conditions (Lippmann, 1996).

Chapter 3

Missing data in speech processing

3.1 Introduction

In this chapter the idea that parts of the speech spectrum may be obscured by sounds from other sources will be discussed. Arguments supporting the idea from various sources will be put forward. It will be argued that the recognition should be performed in two distinct steps: (a) grouping of the evidence coming from the speech source of interest; and (b) adaptation of the recogniser to handle the partial speech. Techniques for pattern classification that enable the adaptation of the statistical systems will be reviewed, and their merits for this purpose assessed. Toward the end, previous studies of using the missing data approach to robust speech and speaker recognition will be reviewed, and relations to similar techniques highlighted.

3.2 Motivation

The claim that parts of the speech can be obscured and not observable is fairly unintuitive. A similar claim for vision is much more natural because most visual objects are opaque. Objects in the near field of vision regularly hide far field objects from our view. Yet we have no problem observing their existence when they are only partially visible. However, when it comes to speech, our intuition suggests that the effect of the sound scene must be additive. After all, the physics of sound make it so. Techniques like blind source separation (Section 4.3) that rely on this assumption have demonstrated sound separation when their preconditions are satisfied. However, it is known that in the human auditory system the louder sounds obscure the quieter ones. This happens at several levels, effectively removing the quieter sounds from the subsequent processing stages. Thus, for the purpose of hearing they are lost. We support this claim by several arguments:

(a) The masking occurring at various levels in the auditory chain makes the masked speech effectively lost for subsequent processing and thus missing. Forward and backward masking in the time–frequency (T–F) plane, and masking in the frequency bands surrounded by more energetic neighbouring bands is routinely used in speech coding and compression (ex: ISO/IEC 11172-3 (1993); ISO/IEC 13818-3 (1995)). Another example is the "capture effect" exhibited in the neural code. The most intense component dominates the neural response both in terms of the firing rate and the temporal response pattern (Moore, 1982). Lateral suppression and inhibition (discharge rate reduction in the presence of additional signal) at the level of a single fibre in the inner ear is also documented. (Greenberg, 1997).

(b) In natural auditory scenes, the number of sources is almost never one, and most of the time it is probably at least three. In addition, the role the speaker's *attention* plays in attending one source while pushing the rest in the "acoustic background", may make missing data/masking much more ubiquitous in real life than in controlled artificial environments with a single acoustic source.

(c) Communication via restricted bandwidth channels occurs on a daily basis between human listeners. Telephone speech is one example. Interfering band–limited noises are frequent in the natural acoustic environment. Yet humans don't have any problems when large parts of the spectrum are missing.

(d) Humans handle severe alterations of the signal in the time–frequency domain gracefully. Low and high pass filtered speech (Allen, 1994) and speech filtered through narrow and steep bandpass filters (Warren et al., 1995; Steeneken, 1992) retains very high intelligibility suggesting that:

- speech is redundant and can be intelligible even if only small parts of the spectrum are available
- human audition can adapt to partial, scattered and sparse evidence in the time–frequency plane

(e) There is enough evidence by now suggesting that the human auditory system organises the concurrent signals into perceptual streams (Bregman, 1990), regardless if it is confronted with complex signals like speech or much simpler signals. The organisation seems to be influenced both by bottom–up primitive rules and top–down schemas. The computational models of this problem of "binding" the multiple sensory information to a particular source in the auditory scene typically assign each observation to a single source only. All the criteria for organisation can rely only on the prior knowledge that physically distinct sources are independent. A typical organisation then looks for a subset of sufficiently (statistically) dependent signals which are sufficiently (statistically) independent from the other signals.

(f) From purely signal processing point of view: the human hearing periphery exhibits a strong compression transfer function, often loosely mimicked by log or cubic root compression following the spectral analysis in the todays ASR systems feature extraction module. The compression makes the approximation of a *sum* by a *max* operator much more viable. The downside of the compression non–linearity is that the signal must have large enough dynamic range so that the information transfer remains possible. The speech signal has a dynamic range of more then 12 orders of magnitude. In addition to this, speech (and other sources) too exhibit quite sparse representation in the time–frequency plane.

The bottom two panels of Figure 3.1 show a T–F representation (after the compression) of both the clean speech and noisy speech at global SNR of 0dB. Both are "seen through" the mask which selects the points where the speech is more energetic then the noise. The clean and the noisy speech seem indistinguishable when seen through the mask.

3.3 The missing data approach to robust speech recognition

The missing data approach to the problem of robust ASR assumes the following:

- local patches in any time–frequency representation of the speech spectrum remain mostly unaffected by the other sounds even at very poor global SNRs;
- they can be identified with a certain probability;
- there is sufficient quantity of (partial) information in the patches for recognition of the speech they originated from.

If these assumptions hold, and it is possible to engineer techniques that would produce results (under these assumptions), one would expect the recogniser to show graceful performance degradation in the cases of occlusion occurring when more then one source is present in the auditory scene. In that case, the two subproblems involving the application of missing data techniques to robust ASR are:

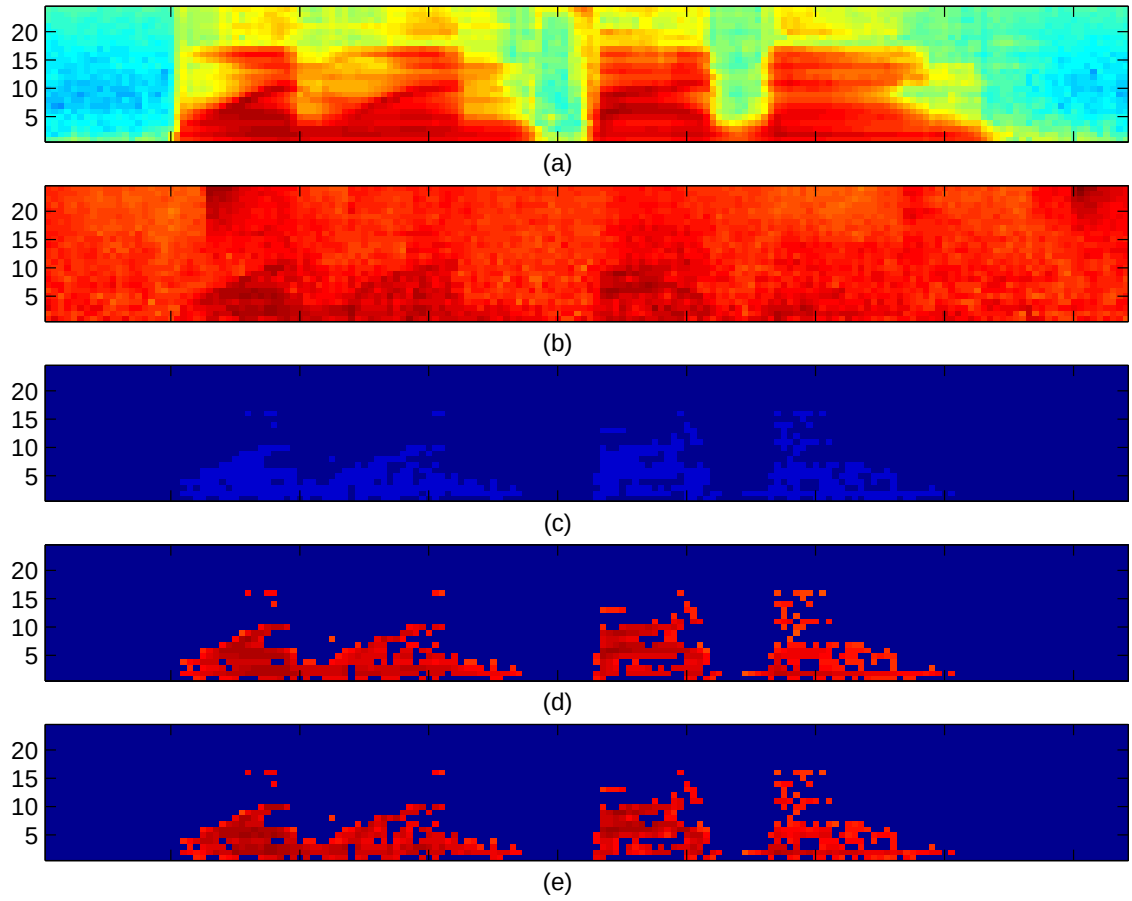


Figure 3.1: (a) Clean speech (spoken digits "1159") (b) Noisy speech at 0dB (c) The present data mask (the light blue area indicates the present data) (d) Clean speech as seen through the mask (e) Noisy speech as seen through the mask

- identification of the reliable parts of the spectrum
- recognition using only the reliable parts of the spectrum

3.3.1 Identification of the reliable parts of the speech spectrum

The identification of the reliable parts of the speech spectrum needs to use only the properties of the speech signal and the physics of sound. We are going to assume some form of time–frequency representation of the speech features from now on. But any representation preserving the local properties of the sources would suffice. It has already been noted by several researchers that features that preserve the localness of the time–frequency plane tend to be more amenable to noise compensation in unmatched conditions (Holmes and Sedgwick, 1986; Barker, 1998; de Veth et al., 1999). The separation process may be carried out:

- using only bottom–up constraints, like CASA (Section 4.2) does by using "low–level" features emerging from the physics of sound; no specific source models are required, but rather a more general knowledge about what features are bound together because they are likely to be coming from the same source
- using only top–down constraints in the form of whole source models (e.g. noise model), maybe additionally paired with some prior knowledge about the nature of the "match" between the models and observations (acoustic back–off and the UNION model, Section 3.5.3)

However, this is an artificial division. Both parts of the process contribute valuable constraints to the search for the best explanation of the attended source in multi–source environment. There is no reason not to use a constraint if one is available. Equally, there is no reason to rely on all constraints being available to do the search.

In the following text, it will be assumed that the above processes will divide the feature vector $\mathbf{x}$ into a reliable (present) and unreliable (missing) component $\mathbf{x} = (\mathbf{x}_p, \mathbf{x}_m)$. The distinction is from the viewpoint of the attended source. If there are two sources, then the present/missing division for the second source is going to be complementary/opposite to the first one. The process can be visualised as if a "mask" $\mathbf{x}_m$ has been placed over the feature vector $\mathbf{x}$ allowing us to see only the present components $\mathbf{x}_p$. The mask $\mathbf{m}$ determines the separation completely.

Figure 3.1 shows an example of what a mask might look like. The mask is derived for a mixture of the clean speech (a) (digits string utterance from TIdigits database (Leonard, 1984)) with noise (factory noise from NOISEX database (Varga et al., 1992)) at global SNR of 0dB (b). The representation is a standard Mel–spectrum filterbank (Young and Woodland, 1993). Panel (c) depicts the mask. The light blue dots indicate reliable regions[1]. Panels (d) and (e) show the clean and the noisy speech respectively as "seen through" the mask. They are almost indistinguishable.

It would be advantageous if the separation process delivers assessment of how probable is the proposed mask. It would be even better if it is possible to assess the probability of every possible mask. That means that there will be a probability $P(\mathbf{m})$ associated with each mask $\mathbf{m}$. Choosing a single mask (with a probability of unity) is one extreme case of this. The other extreme is failing to introduce any new constraints in the search and giving equal probability to every possible mask.

3.3.2 Recognition using the reliable parts of the spectrum only

The speech models need to have their probability evaluated with only parts of the observation vector coming from the source they are modelling. Since there is no prior knowledge which features are going to be missing, it is reasonable to assume that the models inferred during the training will have complete feature vectors, and will be adapted to handle the occlusion occurring during the recognition gracefully and in a principled manner. This requires adaptation of the models on a frame–by–frame basis. The techniques for implementing the adaptation are:

[1] the criterion used for selection was 7.6555 dB local SNR; it was calculated by comparing the clean and the noisy speech and assuming additivity of the speech and the noise in the power spectral domain

- *Marginalisation* – only the present data $\mathbf{x}_p$ is used to compute the likelihood of a model W. The relation between the probability of the "full" data $\mathbf{x}$ and the partial present data $\mathbf{x}_p$ for a model W is:

$$p(\mathbf{x}_p|W) = \int_{\Omega_{\mathbf{x}_m}} p(\mathbf{x}_p, \mathbf{x}_m|W) d\mathbf{x}_m \tag{3.1}$$

  The marginal probability $p(\mathbf{x}_p|W)$, instead of the "full" probability $p(\mathbf{x}|W)$ is then used in the subsequent stages of processing.

- *Data imputation* – the missing data is "filled in" (imputed) using the available knowledge in the form of the model $p(\mathbf{x}|W)$ and the present data $\mathbf{x}_p$. In order to do that, the conditional distribution of the missing data $\mathbf{x}_m$ is needed first:

$$p(\mathbf{x}_m|\mathbf{x}_p, W) = \frac{p(\mathbf{x}_p, \mathbf{x}_m|W)}{p(\mathbf{x}_p|W)} = \frac{p(\mathbf{x}_p, \mathbf{x}_m|W)}{\int_{\Omega_{\mathbf{x}_m}} p(\mathbf{x}_p, \mathbf{x}_m|W) d\mathbf{x}_m} \tag{3.2}$$

  Then, a single value $\hat{\mathbf{x}}_\mathbf{m}$ from the conditional distribution has to be chosen according to some criterion and used as a "plug in" replacement for the missing values $\mathbf{x}_m$. The choice can be made by minimising an error criterion. One of the most commonly used error measures is the mean square error (MSE). In that case, the minimal MSE (MMSE) value for $\mathbf{x}_m$ is the conditional expectation:[2]

$$\hat{\mathbf{x}}_{\mathbf{m}|\mathbf{W}} = \mathcal{E}\{\mathbf{x_m}|\mathbf{x_p}, \mathbf{W}\} = \int_{\mathbf{\Omega_{x_m}}} \mathbf{x_m} \mathbf{p}(\mathbf{x_m}|\mathbf{x_p}, \mathbf{W}) \mathbf{dx_m} \tag{3.3}$$

  Depending on the circumstances of application of the technique, other criteria may be used as well (Chapter 5). Once $\hat{\mathbf{x}}_\mathbf{m}$ is computed, it is used as a plug–in value instead of $\mathbf{x}_m$ and the "filled in" feature vector $(\mathbf{x}_m, \hat{\mathbf{x}}_\mathbf{m})$ is used for further processing.

The techniques for handling missing data stem from the statistical techniques for manipulating probability densities: marginalisation and conditioning. There is an intuitive connection between them. The marginal distribution $p(\mathbf{x}_p|W)$ can be seen as an "average" over all possible conditional imputations $p(\mathbf{x}_p|\mathbf{x}_m, W)$ weighted by their respective probabilities of occurrence $p(\mathbf{x}_m|W)$:

$$p(\mathbf{x}_p|W) = \int_{\Omega_{\mathbf{x}_m}} p(\mathbf{x}_p|\mathbf{x}_m, W) p(\mathbf{x}_m|W) d\mathbf{x}_m \tag{3.4}$$

3.4 Review of pattern matching methods for missing data

Normally for the purposes of speech recognition it will be assumed that the full (clean) data is available during the models training (parameter estimation). However, techniques for learning or model parameters estimation from incomplete data give insight into the possibilities for handling the missing data condition in principled way.

Little and Rubin (1997) formalised the missing data mechanism as two coupled statistical processes, one generating a feature vectors X and another one generating masks M that determine which parts of X are present/missing. Their the joint distribution is:

$$p(X, M|\theta, \phi) = p(M|X, \phi) p(X|\theta) \tag{3.5}$$

where θ are the parameters of the production process and ϕ are the parameters of the censoring process. Then, the following observations about the nature of the missing data mechanism can be made (Gelman et al., 1995; Ghahramani and Jordan, 1994b):

[2] for any random variable x with p.d.f. $p(x)$ the expectation $\mathcal{E}\{x\} = \int x p(x) dx$ is a value that minimises the mean square error MSE; ex: $MSE = \mathcal{E}\{(x-m)^2\} = \int (x-m)^2 p(x) dx$ where m is the unknown value; setting the first derivative of the MSE with respect to m to 0 gives $m = \int x p(x) dx$ i.e. $m = \mathcal{E}\{x\}$

(a) The mask M is independent of the data X. The missing data is said to be missing completely at random (MCAR): $p(M|X,\phi) = p(M|\phi)$

(b) The mask M depends on the present data X_p but not on the missing data X_m. The missing data is said to be missing at random (MAR): $p(M|X,\phi) = p(M|X_p, X_m, \phi) = p(M|X_p, \phi)$

(c) The mask M depends both on the present data X_p and on the missing data X_m. The missing data is said not to be missing at random (NMAR): $p(M|X,\phi) = p(M|X_p, X_m, \phi)$. This is sometimes referred to as a data "censoring".

In the case of robust ASR, X_p is the speech, and X_m is the noise observations. CASA (and other bottom–up techniques for speech estimation from noisy speech) assume that the mask M can be determined using the speech observations X_p only, and independent of the noise X_m. Therefore they assume the MAR model (case (b)) of censoring. Techniques that rely on both the (estimated) speech and noise in order to derive the mask (e.g. noise estimation for the purposes of SNR estimation for mask estimation) need both X_p and X_m to derive M. Therefore they assume the NMAR model (case (c)) of censoring.

The learning process is a search for parameters θ and ϕ that maximise the probability of the observed (present) data X_p and the mask M. The maximum likelihood (ML) methods maximise the likelihood:

$$\mathcal{L}(\theta, \phi|X_p, M) \propto P(X_p, M|\theta, \phi) \tag{3.6}$$

The maximum a posteriori (MAP) based methods maximise the a posteriori probability:

$$P(\theta, \phi|X_p, M) \propto P(X_p, M|\theta, \phi)P(\theta, \phi) \tag{3.7}$$

In both cases, the common factor:

$$P(X_p, M|\theta, \phi) = \int P(M|X_p, X_m, \phi)P(X_p, X_m|\theta)dX_m \tag{3.8}$$

contains the expression $P(M|X_p, X_m, \phi)$. In the cases of MCAR and MAR this expression is independent of X_m so at least it is $P(M|X_p, X_m, \phi) = P(M|X_p, \phi)$ (in the MCAR case even $P(M|X_p, X_m, \phi) = P(M|\phi)$). So $P(M|X_p, X_m, \phi)$ can be moved out of the integral giving:

$$P(X_p, M|\theta, \phi) = P(M|X_p, \phi)P(X_p|\theta) \tag{3.9}$$

The last equation is important as it allows to ignore the missing data mechanism if all that is needed is estimation of θ (the parameters that define the process generating the data). The θ that maximises $\mathcal{L}(\theta|X_p) \propto P(X_p|\theta)$ will also maximise $\mathcal{L}(\theta, \phi|X_p, M)$.

The MAP estimator has an additional factor $P(\theta, \phi)$ that connects the parameters of the data generating and the data masking process. One way around this is to assume that it is factorisable $P(\theta, \phi) = P(\theta)P(\phi)$. Then the data generating process parameters θ can be estimated without the need to estimate the masking process parameters ϕ.

Both ML and MAP can not ignore the data masking process in the NMAR case. Its parameters ϕ have to be estimated as well.

3.4.1 Parameters estimation with missing data for mixture models

We will consider mixture models in connection with the density based approach to learning (Ghahramani and Jordan, 1994a). Both are commonly used in the speech recognition statistical systems. The mixture model assumes the data was generated independently from a mixture of densities:

$$P(\mathbf{x}) = \sum_k P(\mathbf{x}|k, \theta_k)P(k) \tag{3.10}$$

where the components denoted by k have the parameters θ_k. The density based approach to learning estimates the joint density of the present and the missing data (and all variables, in fact)

as a first step[3]. All other relations of interest between any of the variables are derived from the joint density via conditioning and marginalisation. For example, the regression is the expectation over the conditional density.

The Expectation–Maximisation (EM) algorithm is usually the algorithm of choice for maximising the probability for the complete data (Dempster et al., 1977). For a given set if samples $X = \{\mathbf{x}_i\}$, the log–likelihood $\mathcal{L}$ (since the logarithm is a monotonic function, θ that maximises it maximises the likelihood, too) is:

$$\mathcal{L}(\theta|X) = \sum_i \log(\sum_k P(\mathbf{x}_i|k,\theta_k)P(k)) \tag{3.11}$$

where $\mathcal{L}(\theta|X) = \log p(X|\theta)$ is the log–likelihood of the model parameters θ given the data $X = \{\mathbf{x}_i\}$.

The original EM–algorithm itself introduces a "hidden" indicator variable $Z = \{z_{ik}\}$ that is 1 if and only if the vector $\mathbf{x}_i$ is generated by the mixture component k, and 0 otherwise. Then the "complete" log–likelihood $\mathcal{L}$ becomes:

$$\mathcal{L}_c(\theta|X) = \sum_i \sum_k z_{ik} \log(P(\mathbf{x}_i|\mathbf{z}_i,\theta)P(\mathbf{z}_i,\theta)) \tag{3.12}$$

The variable z_{ik} is itself missing data, since it can not be observed directly. The EM–algorithm states that $\mathcal{L}(\theta|X)$ can be maximised by iteratively alternating the following two steps:

$$\begin{aligned} Q(\theta|\theta^{(n)}) &= \mathcal{E}\{\mathcal{L}_c(\theta|X,Z)|X,\theta^{(n)}\} \quad \text{(E-step)} \\ \theta^{(n+1)} &= \underset{\theta}{argmax} Q(\theta|\theta^{(n)}) \quad \text{(M-step)} \end{aligned} \tag{3.13}$$

where (n) and $(n+1)$ denote two subsequent steps in the iteration. The extension when part of the data is missing is natural. The only difference is that the expectation in the E–step is taken with respect to the present (not the complete) data, in addition to the indicator variables Z.

$$\begin{aligned} Q(\theta|\theta^{(n)}) &= \mathcal{E}\{\mathcal{L}_c(\theta|X_p,X_m,Z)|X_p,\theta^{(n)}\} \quad \text{(E-step)} \\ \theta_k^{(n+1)} &= \underset{\theta}{argmax} Q(\theta|\theta^{(n)}) \quad \text{(M-step)} \end{aligned} \tag{3.14}$$

Mixtures of Gaussian distributions $\sum_k w_k \mathcal{N}(\mathbf{x};\mu_k,\Sigma_k)$ are the model commonly used in speech recognition. For this case, the function $Q(\theta|\theta^{(n)})$ (E–step) in Eq. (3.13) is:

$$Q(\theta|\theta^{(n)}) = \sum_i \sum_k P(k|\mathbf{x}_i)^{(n+1)} \log(w_k \mathcal{N}(\mathbf{x}_i;\mu_k,\Sigma_k)) \tag{3.15}$$

where the probability $P(k|\mathbf{x}_i)^{(n+1)}$ that a mixture k generated data $\mathbf{x}_i$ is:

$$P(k|\mathbf{x}_i)^{(n+1)} = \frac{w_k^{(n)} \mathcal{N}(\mathbf{x}_i;\mu_k^{(n)},\Sigma_k^{(n)})}{\sum_{k'} w_{k'}^{(n)} \mathcal{N}(\mathbf{x}_i;\mu_{k'}^{(n)},\Sigma_{k'}^{(n)})} \tag{3.16}$$

Differentiating $Q(\theta|\theta^{(n)})$ with respect to the "free" parameters $\theta = (w_k,\mu_k,\Sigma_k)$, setting the differentials to zero and solving the equations under the constraint $\sum_k w_k^{(n+1)} = 1$ (the M–step) yields the reestimation formulae for the mixture coefficients, means and variances of the Gaussians:

$$w_k^{(n+1)} = \frac{1}{N} \sum_i P(k|\mathbf{x}_i)^{(n+1)} \tag{3.17}$$

$$\mu_k^{(n+1)} = \frac{\sum_i P(k|\mathbf{x}_i)^{(n+1)} \mathbf{x}_i}{\sum_i P(k|\mathbf{x}_i)^{(n+1)}} \tag{3.18}$$

$$\Sigma_k^{(n+1)} = \frac{\sum_i P(k|\mathbf{x}_i)^{(n+1)} \mathbf{x}_i \mathbf{x}_i^T}{\sum_i P(k|\mathbf{x}_i)^{(n+1)}} - \mu_k^{(n+1)} (\mu_k^{(n+1)})^T \tag{3.19}$$

[3]sometimes the joint density estimation might be a harder problem then the one we are trying to solve

where N is the number of data vectors $X = \{\mathbf{x}_i\}_{i=1\ldots N}$.

In the case of missing data, it is not only the indicator variables that are missing, but parts of the data $\mathbf{x}_i$ as well. We will introduce the notation $\mu_{p,k}$ and $\mu_{m,k}$ to denote parts of the mean vector of Gaussian k belonging to the present and missing components, and $\Sigma_{pp,k}$, $\Sigma_{pm,k}$, $\Sigma_{mp,k}$ and $\Sigma_{mm,k}$ to denote parts of the covariance matrix of the Gaussian k with covariances between the present–present, present–missing, missing–present and missing–missing features. Similarly $\mathbf{x}_{p,i}$ and $\mathbf{x}_{m,i}$ denotes the present and missing components respectively of a feature vector $\mathbf{x}_i$. The log–likelihood in that case is:

$$\mathcal{L}_c(\theta|X_p, X_m, Z) = \sum_i \sum_k z_{ik} \log(P(\mathbf{x}_i|\mathbf{z}_i, \theta)) + \sum_i \sum_k z_{ik} \log(P(\mathbf{z}_i, \theta)) \tag{3.20}$$

Ignoring the second term (since only $P(\mathbf{x}_i|\mathbf{z}_i, \theta)$ is of interest to us), and decomposing the Gaussians to their present and missing components yields (Ghahramani and Jordan, 1994a):

$$\begin{aligned}
&\mathcal{L}_c(\theta|X_p, X_m, Z) = \\
&\quad \sum_i \sum_k z_{ik} [\frac{d}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma_k - \frac{1}{2}(\mathbf{x}_{p,i} - \mu_{p,j})^T \Sigma_{pp,j}^{-1} (\mathbf{x}_{p,i} - \mu_{p,j}) \\
&\quad - (\mathbf{x}_{p,i} - \mu_{p,j})^T \Sigma_{pm,j}^{-1} (\mathbf{x}_{m,i} - \mu_{m,j}) - \frac{1}{2}(\mathbf{x}_{m,i} - \mu_{m,j})^T \Sigma_{mm,j}^{-1} (\mathbf{x}_{m,i} - \mu_{m,j})
\end{aligned} \tag{3.21}$$

When computing $Q(\theta|\theta^{(n)})$ the first terms yields $\mathcal{E}\{z_{ik}|\mathbf{x}_{p,i}, \theta^{(n)}\}$ which (similarly to Eq. (3.16)) gives:

$$P(k|\mathbf{x}_{p,i})^{(n+1)} = \frac{w_k^{(n)} \mathcal{N}(\mathbf{x}_{p,i}; \mu_{p,k}^{(n)}, \Sigma_{pp,k}^{(n)})}{\sum_{k'} w_{k'}^{(n)} \mathcal{N}(\mathbf{x}_{p,i}; \mu_{p,k'}^{(n)}, \Sigma_{pp,k'}^{(n)})} \tag{3.22}$$

The second term of Eq. (3.21) yields $\mathcal{E}\{z_{ik}\mathbf{x}_{m,i}|\mathbf{x}_{p,i}\theta^{(n)}\}$ which can be computed as:

$$\begin{aligned}
\mathcal{E}\{z_{ik}\mathbf{x}_{m,i}|\mathbf{x}_{p,i}\theta^{(n)}\} &= P(k|\mathbf{x}_{p,i})^{(n+1)} \mathcal{E}\{\mathbf{x}_{m,i}|z_{ik} = 1, \mathbf{x}_{p,i}, \theta^{(n)}\} \\
&= P(k|\mathbf{x}_{p,i})^{(n+1)} (\mu_{m,k} + \Sigma_{mp,k}\Sigma_{pp,k}^{-1}(\mathbf{x}_{p,i} - \mu_{p,k}))
\end{aligned} \tag{3.23}$$

Similarly, the third term of Eq. (3.21) yields $\mathcal{E}\{z_{ik}\mathbf{x}_{m,i}\mathbf{x}_{m,i}^T|\mathbf{x}_{p,i}\theta^{(n)}\}$ which can be computed as:

$$\mathcal{E}\{z_{ik}\mathbf{x}_{m,i}\mathbf{x}_{m,i}^T|\mathbf{x}_{p,i}\theta^{(n)}\} = P(k|\mathbf{x}_{p,i})^{(n+1)} (\Sigma_{mm,k} - \Sigma_{mp,k}\Sigma_{pp,k}^{-1}\Sigma_{mp,k}^T + \hat{\mathbf{x}}_{m,i,k}\hat{\mathbf{x}}_{m,i,k}^T) \tag{3.24}$$

where $\hat{\mathbf{x}}_{m,i,k} = \mu_{m,k} + \Sigma_{mp,k}\Sigma_{pp,k}^{-1}(\mathbf{x}_{p,i} - \mu_{p,k})$ is the regression (expectation) of the missing given the present data for k-th Gaussian (used in Eq. (3.23), too). Once these terms are computed, the update formulae for the parameters of the model can be found the same way as in the complete data case.

Ghahramani and Jordan (1994a) also consider the case of EM for discrete variables with incomplete data.

3.4.2 Classification with missing data

Although the classification is just a special case of function approximation with the data vector elements divided into "inputs" and "outputs" and where the outputs are discrete variables (in most cases), it warrants special attention. There are hybrid speech recognition systems where as part of the process evaluation of the posterior probability of a class given the data is needed.

Classification with mixture models

Mixture models can handle classification with missing data naturally. Once the joint density of the data and the class labels $P(\mathbf{x}, C = j|\theta)$ is known, every relation between them can be derived readily. Most often the posterior probability $P(C = j|\mathbf{x}, \theta)$ is sought. One example for the joint

data – class labels density $P(\mathbf{x}, C = j|\theta)$ is a mixture of type (Ghahramani and Jordan, 1994a; Tresp et al., 1994):

$$P(\mathbf{x}, C = j|\theta) = \sum_k w_{jk}\pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k) \tag{3.25}$$

with Gaussian and multinomial components. If part of the vector $\mathbf{x}_i$ is missing, then the E–step of the EM algorithm is:

$$P(k|\mathbf{x}_{p,i}, C_i = j)^{(n+1)} = \frac{w_{jk}^{(n)}\pi_k^{(n)}\mathcal{N}(\mathbf{x}_{p,i}; \mu_{p,k}^{(n)}, \Sigma_{pp,k}^{(n)})}{\sum_{k'} w_{jk'}^{(n)} w_{k'}^{(n)} \mathcal{N}(\mathbf{x}_{p,i}; \mu_{p,k'}^{(n)}, \Sigma_{pp,k'}^{(n)})} \tag{3.26}$$

If the class label C_i is unknown then $w_{jk}^{(n)}$ factors vanish from the denominator and the numerator and the E–step is the same as for a Gaussian mixtures, Eq. (3.22). Then in the M–step the class label's j–th component is $\sum_k P(k|\mathbf{x}_{p,i}, C_i = j)^{(n+1)} w_{jk}$. The rest of the M–step is the same as at the Gaussian mixture model.

Somewhere in between the mixture model Eq. (3.25) and the feedforward networks discussed in the next subsection are the two layer radial bases functions (RBF) networks with Gaussian kernels (Bishop, 1995), used by Tresp et al. (1994) to classify with missing data. Each output y_i of the network is computed as:

$$y_j(\mathbf{x}) = \frac{\sum_k w_{kj}\pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)} \tag{3.27}$$

With suitable (unsupervised) training the network kernels in the first layer will estimate the data density–the denominator in the equation above. Similarly, the second layer will estimate the joint density of the data and the classes. It is then possible to compute the class posterior probability given a data vector $P(C_j|\mathbf{x}_p)$ needed for a forward pass when parts of the feature vector are missing (Ahmad and Tresp, 1993):

$$\begin{aligned} P(C_j|\mathbf{x}_p) &= \frac{\int \sum_k w_{kj}\pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k) d\mathbf{x}_m}{\int \sum_k \pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k) d\mathbf{x}_m} \\ &= \frac{\sum_k w_{kj}\pi_k \mathcal{N}(\mathbf{x}_p; \mu_{p,k}, \Sigma_{pp,k})}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_p; \mu_{p,k}, \Sigma_{pp,k})} = y_j(\mathbf{x}_p) \end{aligned} \tag{3.28}$$

Radial basis Boltzmann machines are another class of networks that share the property of input data density estimation as part of the training process. Consequently, this special type of Boltzmann machine can handle missing data in their inputs naturally and have been used for classification with missing data (Kappen and Nijman, 1995).

Classification with feedforward networks

Many feedforward networks have 1–of–N coding of the targets and are trained to minimise a quadratic error function. Richard and Lippmann (1991) have shown that in that case each output of the network y_i estimates the posterior probability $P(C_i|\mathbf{x})$ that the vector $\mathbf{x}$ comes from class C_i. We will consider the case when part of the input is missing only. It will be assumed that the class labels are always present. When part of the input is missing, the posterior probability $P(C|\mathbf{x}_p)$ of the class C given the present data $\mathbf{x}_p$ is (Ahmad and Tresp, 1993; Bishop, 1995):

$$P(C|\mathbf{x}_p) = \int P(C|\mathbf{x}) p(\mathbf{x}_m|\mathbf{x}_p) d\mathbf{x}_m \tag{3.29}$$

Intuitively, the form states that the estimate of the posterior on the basis present data $P(C|\mathbf{x}_p)$ is the average of the full data posterior $P(C|\mathbf{x})$ over all possible completions of the missing values $\mathbf{x}_m$, weighted by the probability that $\mathbf{x}_m$ could occur, given the present data $\mathbf{x}_p$. I.e. it is the expected conditional posterior given the present data $\mathcal{E}_{\mathbf{x}_m|\mathbf{x}_p}\{P(C|\mathbf{x})\}$.

The j–th network output $NN_j(\mathbf{x})$ estimates the class conditional density of the j–th class $NN_j(\mathbf{x}) \approx P(C_j|\mathbf{x})$ only. So the data density $p(\mathbf{x})$ remains unknown. One solution to the problem is to estimate the data density separately (Tresp et al., 1995). For example, the input distribution can be approximated using Parzen windows (Duda and Hart, 1973):

$$P(\mathbf{x}) = \frac{1}{N}\sum_i \mathcal{N}(\mathbf{x}; \mathbf{x}_i, \sigma) \tag{3.30}$$

where $\{\mathbf{x}_i\}$ for $i = 1 \dots N$ is the training data and σ is fixed. It is also obvious that $P(\mathbf{x}_p) = \frac{1}{N}\sum_i \mathcal{N}(\mathbf{x}_p; \mathbf{x}_{p,i}, \sigma)$. Then, the missing data integral Eq. (3.29) becomes:

$$P(C_j|\mathbf{x}_p) \quad = \quad \frac{1}{\frac{1}{N}\sum_i \mathcal{N}(\mathbf{x}_p; \mathbf{x}_{p,i}, \sigma)} \int NN_j(\mathbf{x})[\frac{1}{N}\sum_i \mathcal{N}(\mathbf{x}; \mathbf{x}_i, \sigma)]d\mathbf{x}_m, \tag{3.31}$$

where $\mathbf{x} = (\mathbf{x}_p, \mathbf{x}_m)$ is a test data vector that is to be classified and has the $\mathbf{x}_m$ components missing.

Assuming that the network prediction is constant over the "width" of the Gaussians, the $NN_j(\mathbf{x})$ comes out of the integral (as a constant) and the integral cancels out the missing data dimensions of $\mathcal{N}(\mathbf{x}; \mathbf{x}_i, \sigma)$ reducing it to $\mathcal{N}(\mathbf{x}_p; \mathbf{x}_{p,i}, \sigma)$:

$$P(C_j|\mathbf{x}_p) \approx \frac{\sum_i NN_j(\mathbf{x}_p, \mathbf{x}_{m,i})\mathcal{N}(\mathbf{x}_p; \mathbf{x}_{p,i}, \sigma)}{\sum_i \mathcal{N}(\mathbf{x}_p; \mathbf{x}_{p,i}, \sigma)} \tag{3.32}$$

The expression $NN_j(\mathbf{x}_p, \mathbf{x}_{m,i})$ means that this is the output of the network obtained feeding the present components of the observation data vector and missing components from the i–th training pattern. Therefore, in order to evaluate the missing data integral for a single input, a sum over all patterns from the training set has to be computed. This is nearly impossible to apply to a hybrid speech recognition system, since the number of training examples is very large. Instead, some form of clustering or semiparametric density estimation (e.g. Gaussians mixture) of the data density in the training set maybe employed. Then, instead of using the complete training set during the recognition, centroids of the clusters (means of the mixtures) will be used in the sum in Eq. (3.32). Raj et al. (1998) and Dupont (1998) used this approach.

If minimal assumptions only are to be made about the data density $p(\mathbf{x})$, then a closed form Eq. (B.14) (Appendix B) can be used to estimate the outputs of a single layer network with sigmoid transfer function. However, the assumption that an observation $\mathbf{o}$ makes all $\mathbf{x}_m$ in the interval $[0, \mathbf{o}_m]$ equally probable is clearly geared toward application in speech recognition with a certain type of features $\mathbf{x}$. As in our experiments no hybrid ASR system was used, this line of enquiry was not followed through.

Learning from incomplete data

The network parameters θ can be estimated once the probability of the complete data is known:

$$\mathcal{L} = \sum_i \log(P(C_j|\mathbf{x}_i, \theta)) + \sum_i \log(P(\mathbf{x}_i|\phi)) \tag{3.33}$$

where ϕ denotes the parameters of the data density. Taking into account Eq. (3.29), the gradient of the likelihood with respect to network parameters θ is (Tresp et al., 1994; Ghahramani and Jordan, 1994b):

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_i \frac{1}{p(C_i|\mathbf{x}_{p,i})} \int P(C_i|\mathbf{x}_{p,i}, \mathbf{x}_m, \theta)p(\mathbf{x}_m|\mathbf{x}_{p,i}, \phi)(t_i - P(C_i|\mathbf{x}_{p,i}, \mathbf{x}_m))\frac{\partial P(C_i|\mathbf{x}_{p,i}, \mathbf{x}_m)}{\partial \theta}d\mathbf{x}_m \tag{3.34}$$

where t_i is the desired output of the network for the data vector $\mathbf{x}_i$ (which belongs to class C_i). The integral has no close form solution and has to be approximated by a Monte Carlo simulation. The missing data components are drawn from the (known) input data distribution and then the integral is calculated.

Recurrent networks

Bengio and Gingras (1996) used recurrent networks with feedback for classification with missing data. In the "static version", network's unknown inputs are initialised with their unconditional mean, and their value is then updated by feedback links with delays, just as if these were hidden units. The network enters an oscillatory regime, effectively searching for the output giving minimal error (as measured by the error function) and imputing the "missing" hidden units in the process. On a small dataset, a learning algorithm using error backpropagation tested favourably compared to mixture model learning.

3.4.3 Missing data imputation for regression

The problem of regression (conditional average) $\mathcal{E}\{\mathbf{y}|\mathbf{x}\}$ is one commonly encountered in statistical analysis. When some of the input data $\mathbf{x}$ is missing, one way to proceed with the regression analysis is to impute the missing values (Little, 1992). The simplest method is to impute the missing components by their unconditional sample means,[4] computed from the data where these components are present. An improvement over this is to impute the missing values from a linear regression estimated between the missing and the present components in the complete data (the conditional mean). When $\mathbf{x}$ and $\mathbf{y}$ are highly correlated, then even better imputations may be obtained by using $\mathbf{y}$ in addition to the present components of the data to compute the regression.

However, the estimate obtained is biased and corrections need to be applied. The standard formulae for computing the errors on the complete data will not take into account the errors of the imputation. Multiple imputation (Rubin, 1987) is one solution to this problem. Instead of a single imputation, the data set is divided into subsets and multiple imputations are calculated. They are all in turn imputed and the complete data analysis is carried out giving multiple sets of regression parameters. The final parameters are their average.

3.5 Missing data for speech recognition: A review

The missing data approach has already been utilised for ASR. Cooke, Green, Anderson, and Abberley (1994b) reported on early experiments with application in speech recognition. A method for adapting a conventional Hidden Markov model (HMM) based ASR system[5] was developed which allows adaptation of the recogniser to an arbitrary pattern of occlusion in the observations. The method employed marginalisation of the missing features, and the state emission probability is calculated as $p(x_p|S)$. It was shown that the adaptation to the recogniser because of marginalisation is simple – it essentially boils down to parameter selection of the state p.d.f.s Gaussians in the known dimensions.[6] The training of the recogniser was on clean speech. The experiments were performed with HTK toolkit (Young and Woodland, 1993) on a TIMIT database (Garofolo and Pallet, 1989) with random occlusions (missing data) on proportions from 0 to 90%. Graceful degradation of performance was reported. Only after 60% of features were deleted, significant degradation did occur. The new method was compared to unconditional mean imputation and clearly performed much better. In the discussion, a CASA (Section 4.2) system was envisaged as a preprocessor for separation. It was also noted that adapted HMM models can act in the separation

[4] sometimes mere zeros are imputed, especially if the data is "preprocessed" to zero mean and unit variance

[5] Contemporary speech recognition systems fall into two broad categories: HMM based (Rabiner and Juang, 1993) and hybrid systems (Bourlard and Morgan, 1993). In the further discussion we will assume an HMM system with emission probability $p(\mathbf{x}|S)$ modelled as a mixture of diagonal Gaussians $p(\mathbf{x}|S) = \sum_k P(k|S)p(\mathbf{x}_p|k,S)p(\mathbf{x}_m|k,S)$, unless stated otherwise.

[6] Chapter 5 discusses this in detail

process like "schemas". The motivation comes from the evidence to suggest that in ASA there is flow of information from the top to the bottom (in addition to bottom–up processing), in a coupled and self–reinforcing manner (Bregman, 1990).

The idea of using missing data for recognition of speech was extended further by introducing training with randomly deleted (missing) data (Cooke et al., 1994c). A self–organising, topology preserving Kohonen network was used for the experiments with a modified learning algorithms which takes into account only the present components of the feature vector. The experiments used PLP and ratemap(Cooke, 1991) representation and occluded data both during training and recognition on a subset of the TIMIT database. There is little degradation for up to 50% deletions, with the ratemap representation outperforming PLP, especially at deletions of more then 50% of the data. It was speculated that this is because of the correlations between the features in the ratemap vector.[7] It was also noted that a further constraint can be utilised: the (observed) total energy of the mixture implies that none of the components has greater energy then this.

This idea was further developed in (Cooke et al., 1994a). The motivation comes from auditory experiments showing that perception of a particular sound can "fill in" gaps in the evidence, provided that there isn't evidence against. This is termed "auditory induction" (Warren et al., 1994) and experiments have shown that people are not aware of gaps in the sounds if there is sufficient energy to allow the possibility of a particular sound (the "phoneme restoration" effect (Warren, 1970)). As a Kohonen net was used for classification, the "auditory induction" was implemented by giving less weight to the hypothesis that expect more energy then the total energy of the mixture of sounds at a particular place. The improvement was notable at high percentages of data deletions. Experiments on correlated deletions ware also carried out, as better approximation to CASA. The relative importance of spectral peaks over the spectral valleys was noted for robustness: an ASR system relying on information in spectral valleys will not be robust as they "fill" with noise first (see Section 2.7.7). It was also found that even for clean speech using the peaks only (instead of the full spectrum) improves recognition. It can be speculated that this may be for two reasons:

- lessening the sensitivity to F_0 which is speaker dependent and hinders the ASR performance
- peaks are less correlated which makes modelling the speech with diagonal covariance matrices in the Gaussians more realistic

Green et al. (1995) reported on incorporating the "auditory induction" idea in an HMM based ASR system originally used in (Cooke et al., 1994b). The state emission probability was computed as:

$$\sum_k P(k|S)p(x_p|k,S)\int_{-\infty}^{\mathbf{o}_m} p(\mathbf{x}_m|k,S)d\mathbf{x}_m \tag{3.35}$$

Further, instead of a random deletion pattern, single and triple digits mixed with babble noise, both from the NOISEX database (Varga et al., 1992) were used. In order to assess the potential of the technique with realistic noise, but without available CASA system, the present features were found by comparison of the noisy with the clean speech. We will refer to the masks derived this way as an *a–priori masks* (panel (c) of Figure 3.1). The comparison gives the local SNR in a particular time–frequency point, as opposed to the global SNR, which is an average computed over the whole utterance. The local SNR was thresholded at values ranging from -10dB to 10dB and time–frequency points with local SNR below the threshold were considered missing. Comparisons of the simulated ASR system with 0dB threshold with listeners performance on the same task gave curves parallel to each other. However, listeners perform with virtually no degradation until 0dB global SNR, and fall off to the levels of chance only at global SNR of -15dB.

Several missing data imputation schemes for adapting the HMM recogniser (in addition to Eq. (3.35)) coupled with different deletion patterns were reported in (Cooke et al., 1996). The missing data imputation $\hat{\mathbf{x}}_m$ was performed as:

- unconditional mean imputation $\hat{\mathbf{x}}_m = \mu_m$

[7] the PLP features are much more independent in the feature vector

- conditional mean imputation $\hat{\mathbf{x}}_m = \mu_m + \Sigma_{pm}^T \Sigma_{pp}^{-1}(\mathbf{x}_p - \mu_p)$
- conditional mean imputation together with conditional variance to weigh the distribution spread

The experiments were carried out on a Resource management (RM) task (Price et al., 1988). It was found that performance of the missing data techniques degrades with "large block" across time and frequency deletions. This is also the case with deletions derived from the a–priori mask. Investigation showed that there are frames where no information at all is available across the whole spectrum. The covariance weighting giving more weight to the states in the frames depending on the amount of available data was also tested. The above techniques, as well as other issues like data orthogonalisation were discussed in more details in (Morris et al., 1998). Using context dependent models with tying, tests were carried out on the same RM task with feature vectors complemented with first and second order differences. The spectral peaks were again found to be advantageous.

3.5.1 Relation to the MAX model of speech and noise combination

It has been already noted by several researchers that assigning the energy in the mixture to one of the sources only is a viable approximation and also significantly eases the computations (Nadas et al., 1989; Varga and Moore, 1991; Kadirkamanathan, 1992; Rose et al., 1994). As seen on Figure 3.2, the more energetic speech features "peak" over the noise, while the less energetic ones gradually submerge under the noise as the global SNR increases.

The MAX environment model (Nadas et al., 1989) models the resulting noisy speech $\mathbf{x}$ as:

$$\mathbf{x} = \max(\mathbf{s}, \mathbf{n}) \tag{3.36}$$

Assuming noise and speech independence, for the cumulative probability $P_X(\mathbf{x})$ of the mixture[8] we have:

$$\begin{aligned} P_X(\mathbf{x}) &= Prob\{X \leq \mathbf{x}\} = Prob\{S \leq \mathbf{x}, N \leq \mathbf{x}\} \\ &= Prob\{S \leq \mathbf{x}\} Prob\{N \leq \mathbf{x}\} = P_S(\mathbf{x}) P_N(\mathbf{x}) \end{aligned} \tag{3.37}$$

After taking the derivative with respect to $\mathbf{x}$, the density of X is expressed as:

$$p_X(\mathbf{x}) = p_S(\mathbf{x}) P_N(\mathbf{x}) + P_S(\mathbf{x}) p_N(\mathbf{x}) \tag{3.38}$$

Once this relation is known, and assuming certain parametric models for the speech and the noise, the parameters of the models can be derived from the noisy data. The EM for mixtures (Section 3.4.1) was used in (Nadas et al., 1989; Rose et al., 1994) to compute the reestimation formulae for speech modelled with Gaussian mixture model and noise modelled with single Gaussian. The formulae essentially express the intuition that the information content of the signals below the noise is low and they should be given less weight when updates are calculated. HMM decomposition (Varga and Moore, 1991) (Section 2.8.2) can be used to deal with multistate noise models.

The missing data approach effectively utilises a MAX environmental model. However, it is assumed that no prior noise model will be available. Then Eq. (3.38) degenerates to $p_S(x)$ when the feature x is present, and $1/x \int_0^x p_S(u)du$ when it is missing. In the latter case, noise is assumed to be uniformly distributed between 0 and x.

3.5.2 Relation to noise masking

Holmes and Sedgwick (1986) modelled the masking of speech by noise (Section 2.6.3) similarly. The noise process has a fixed threshold value known in advance. Then, the masked features contribute $P_S(x)$ to the probability (x is the observed value), while the unmasked features contribute $p_S(x)$.

[8] in this section only the random variables will be marked by capital letters, their realisations with small letters; the probability density function of variable Q evaluated at point $\mathbf{q}$ is $p_Q(q)$, and cumulative probability function is $P_Q(\mathbf{q})$

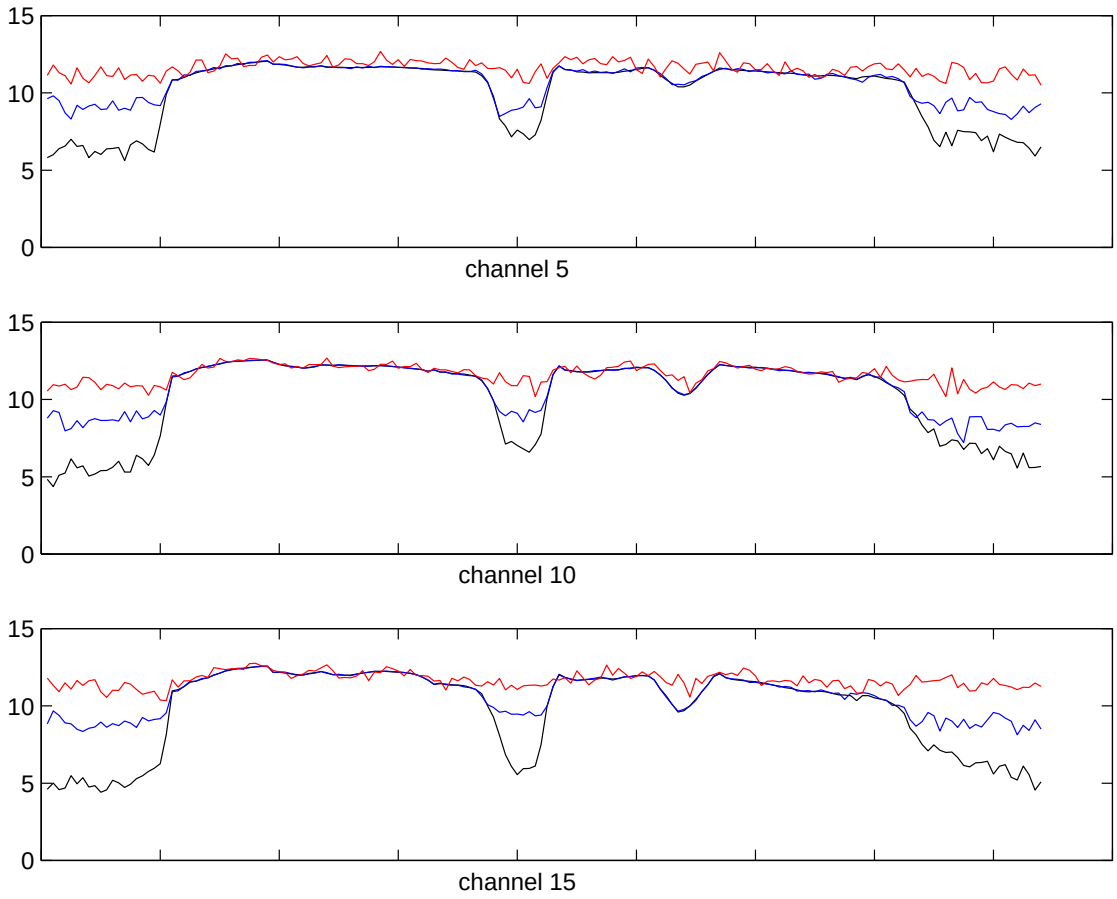


Figure 3.2: Log–magnitude of channels 5 (top panel), 10 (middle panel) and 15 (bottom panel) of a 24–channel Mel–scale filterbank of clean speech (black bottom line), and the same speech mixed with factory noise at 20dB (blue middle line) and 0dB global SNR (red). The horizontal axis is the time. The vertical axis is the log–magnitude.

Training with noisy speech is possible, as well. The mask value for each channel B is the highest noise value found in the training data for that channel. If the fraction of masked features for a channel is F, and μ' is the sample mean of the unmasked samples, then the mean μ and variance σ^2 are:

$$\begin{aligned} \mu &= \frac{\mu' erf^{-1}(F) - BQ(F)}{erf^{-1}(F) - Q(F)} \\ \sigma &= \frac{B - \mu}{erf^{-1}(F)} \end{aligned} \tag{3.39}$$

where $Q(x) = \mathcal{N}(erf^{-1}(x), 0, 1)$.

Similarly to noise masking, Brendborg and Lindberg (1997) investigated two approaches to robustness in the context of a HMM system:

- mean value masking – Gaussians that have means smaller then a threshold were considered sensitive to noise and prevented from scoring very low probability scores
- dimensionality reduction – Gaussians with means smaller then a threshold were ignored

The second technique was reported to give better results. It is equivalent to putting a default missing data mask to every Gaussian in the mixture. The first technique is motivated by similar

concerns as the techniques from the next Section: preventing extremely low scores for outliers in the distribution caused by points dominated by noise (and therefore not drawn from the speech distribution). For sonorant sounds, results similar to PMC without utilising explicit noise model were reported.

3.5.3 Missing feature compensation based on the acoustic evidence

If a state p.d.f. scores very low for a particular observation, all paths passing through that state will have their scores depressed to the extent that it is unlikely that any of them will win. This is because the overall probability of a model is product of the probabilities in each frame. An extremely low probability at some frame effectively discards the model, regardless of strength of the previous or subsequent acoustic evidence. Several factors contribute to this:

- the Gaussian p.d.f. used for state emission probability modelling have very "slim" tails–they quickly fall–off when moving away from their mean
- in the noisy speech, the unreliable (missing) features are not generated by the speech source, but by the noise source; so they may fall on the tail of the speech p.d.f.
- the Gaussians in the mixture are usually diagonal, the features are independent and the final score is a product of the individual features scores; a feature lying on the tails of several Gaussians and scoring very low diminishes the discrimination between the states, regardless of the evidence from the other features

Two techniques have emerged to address this problem: acoustic back–off and the UNION model.

Acoustic back–off

de Veth, Cranen, and Boves (1998) devised an acoustic back–off scheme in order to control the damage: the state distribution is bounded by how low it can score. There is a certain analogy with multigram language models, where a certain probability mass is reserved for the tuples never seen in the training data.[9] There is also a connection with the well known problem of "outliers" in statistics: the problem occurring when the data sample from which the distributions are inferred is not representative enough. Therefore, points that were very rare in the training sample will score very low probabilities. Difference of many orders of magnitude between the "regular" points and "outliers' in the data space may be a poor model of the real process. Reserving certain probability mass for low–frequency points (and thus establishing a lower bound for the probability of the "outlier") is one common technique. So the "backed–off" state p.d.f. $p`(\mathbf{x}|S)$ is:

$$p'(\mathbf{x}|S) = \alpha p(\mathbf{x}|S) + (1-\alpha)\mathbf{p}_0 \tag{3.40}$$

Experimental results in (de Veth et al., 1998) indicate improved robustness when tested with artificially induced "disturbance" of the features.

The UNION model

The union model (Ming and Smith, 2000; Ming et al., 1999) originates from attempts to merge the partial evidence in the context of multiband/multistream speech recognition system (Bourlard et al., 1996; Tibrewala and Hermansky, 1998). As mentioned above, the acoustic scores in the present systems are multiplied together loosing the discriminability between the models when there is an outlier. Assuming feature independence, in the UNION model the probability of observing $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is given by the recurrent relation:

$$P(\mathbf{x}) = P(\vee_{i=1}^{n} x_i) = P(\vee_{i=1}^{n-1} x_i) + P(x_n) - P(\vee_{i=1}^{n-1} x_i)P(x_n) \tag{3.41}$$

[9] assigning probability zero to the unseen tuples would rule out any possibility of recognition, regardless how good the acoustic evidence for recognition is

where $P(\vee_{i=1}^{n} x_i) = P(x_1 \vee x_2 \vee \ldots \vee x_n)$ or equivalently:

$$P(\mathbf{x}) = 1 - \prod_{i=1}^{n}(1 - P(x_i)) \tag{3.42}$$

If the feature $\mathbf{x}$ is a spectral frame (and the individual features are the frequency bands), then this is the "product of error rule": the probability of error is product of the probabilities of error in all frequency bands (Allen, 1994). In this form, the aim of the method is straightforward: when x_i is an outlier and scores extremely low acoustic score, $1 - P(x_i) \approx 1$ and it doesn't disturb the overall acoustic score. While the back-off tries to limit the damage done by the low scores, the union model uses the low scores to its advantage.

It was noted that grouping each band separately greatly damages the discriminability. So groups of bands are combined with the $\wedge$ operator between them (thus accumulating the discriminability), and with the $\vee$ operator between them. This is the segment–based UNION model.

However, both back–off and the UNION model do not add any new constraints in the recognition search. They can not handle the case when some noise patch matches some speech model relatively well and therefore can not be discounted by its acoustic score.

3.5.4 Missing data imputation

Missing data may be reconstructed independently of the speech recogniser. Once reconstructed, the complete data can be fed to the recogniser which need not change at all. This approach is very attractive, and it has already been used (Section 2.6) for speech enhancement. The difference is that the same techniques that are largely used for recognition (like clustering and modelling the data distributions as Gaussian mixtures) are now utilised for reconstruction of the missing features.

(Raj et al., 1998) clustered the input data, and the cluster with maximum score for the present data was used for filling in the missing values. In the second technique, the correlations across time between the missing features and the most highly correlated present features were used in the data imputation process. Dupont (1998) used the data imputation as a preprocessor to a hybrid HMM/ANN system. The separation was carried out via thresholding of the estimated SNR (assuming additive noise model). The data distribution was estimated separately with Gaussians mixture model (GMM). Then, the missing features were compensated by imputing the conditional mean.

Renevey and Drygajlo (2000b,a) integrated together feature separation, spectral subtraction, PMC compensation and data imputation. The data p.d.f. was estimated with diagonal GMM $\sum_k P(k)\mathcal{N}(\mathbf{x}, \mu, \sigma^2; k)$ independently of the recogniser. Under the additive noise assumption, an on–line noise estimation was carried out and the noise p.d.f. was computed. The distribution of the noise was assumed to be Gaussian. Then, the probability of each channel being noisy can be computed for a given observation. By thresholding, the features are separated on present (probability that noise in that channel is greater then the threshold) or missing (probability that noise in that channel is smaller then the threshold). The data GMM model was first compensated with PMC (Eq. (2.23)) using the running noise model estimate. Next, the responsibilities of the compensated GMM were calculated:

$$P(k_{pmc}|\mathbf{x}) = \frac{P(k)\mathcal{N}(\mathbf{x}; \mu_{pmc}, \sigma^2_{pmc}; k)}{\sum_{k'} P(k')\mathcal{N}(\mathbf{x}; \mu_{pmc}, \sigma^2_{pmc}; k')} \tag{3.43}$$

Then in (Renevey and Drygajlo, 2000b) the missing features were compensated by imputing the expected value of the compensated conditional distribution:

$$\hat{\mathbf{x}}_m = \mathcal{E}_{\mathbf{x}_m|\mathbf{x}_p}\{\mathbf{x}_m\} = \sum_k P(k_{pmc}|\mathbf{x}_p)\mu_{m,k} \tag{3.44}$$

The present features $\mathbf{x}_p$ were compensated with spectral subtraction.

In (Renevey and Drygajlo, 2000a) instead of dividing the features on present and missing from the onset and compensating them separately, the probability that a channel is noisy $\Psi(\mathbf{x})$ was used as a "soft" measure between the conditional mean of the compensated GMM and the noisy observation:

$$\hat{\mathbf{x}}_m = [1 - \Psi(\mathbf{x})]\mathcal{E}_{\mathbf{x}_m|\mathbf{x}_p}\{\mathbf{x}_m\} + \Psi(\mathbf{x})\mathbf{x} = [1 - \Psi(\mathbf{x})] \sum_k P(k_{pmc}|\mathbf{x}_p)\mu_{m,k} + \Psi(\mathbf{x})\mathbf{x} \tag{3.45}$$

The techniques showed an improvement over the baseline on the task of recognition digit strings from TIdigits database (Leonard, 1984) mixed with noise from NOISEX-92 database (Varga et al., 1992).

Park and Kim (2000) reported on data imputation from a GMM model of the wideband speech for narrowband (telephone) to wideband speech enhancement.

3.5.5 Stochastic features

Several researchers have noted that if a measure of the feature reliability is available, then it is desirable to integrate over the domain of the unreliable feature, weighting the integral with the reliability measure. In the extreme, if the variance of the feature is infinite, the technique degenerates into marginalisation.

Garner and Holmes (1998) used this idea to incorporate formant features into a conventional HMM–based ASR system. Both the estimate of the reliability of the tracker and the formant features were rigorously incorporated in the HMM model.

Similarly, the unreliability of the features after spectral subtraction was used to weight the evidence during the Viterbi search in an HMM system and during the dynamic programming (DP) search in a dynamic time warping (DTW) ASR system (Yoma et al., 1998). nad H. Pao et al. (1998) named those features "stochastic features", and again used their variance as a weighting factor along the integration path.

3.5.6 Missing data combined with other techniques

The missing data approach has been combined naturally with several other techniques for robust ASR on a speaker verification task (El-Maliki, 2000). It was also used to deal in a principled way with the "negative spectrum" problem arising in spectral subtraction. The features where the noise estimation failed can be considered missing and marginalised (Drygajlo and El-Maliki, 1998a). Similarly, it naturally combines with PMC: only the missing features need to be compensated with PMC (Renevey and Drygajlo, 1999). Compared to "plain" PMC, an improvement on the ASR task of digit strings recognition in noise was reported at all SNRs and for all tested noises. (Drygajlo and El-Maliki, 1998b,c) used a missing data based system, together with non–linear and adaptive generalised SS and MMSE spectral amplitude estimator for detection of the missing features (El-Maliki and Drygajlo, 1999) for robust speaker verification. Padmanabhan and Picheny (2000) utilised the idea of missing data within a multiscale graphical model designed to address several shortcomings of the current systems. Droppo et al. (2002) learnt a separate conditional distribution $p(noisy|clean)$ for different SNRs and noises (from noisy data) and then used the conditional average of the state likelihood (inferred from clean data) during the decoding.

3.5.7 Missing data in speech perception modelling

A missing data recogniser was used to model the perception of severely spectrally reduced sine–wave speech (Barker and Cooke, 1997; Barker, 1998). The sine–wave speech represents the natural speech by a small number time varying sinusoids (somewhat similar to formant synthesis). It is a crude approximation of the natural speech, and yet it remains intelligible.[10] Since it is a greatly reduced representation, sine–wave speech was posed as a challenge to the bottom–up grouping on

[10] with some training on the part of the listeners

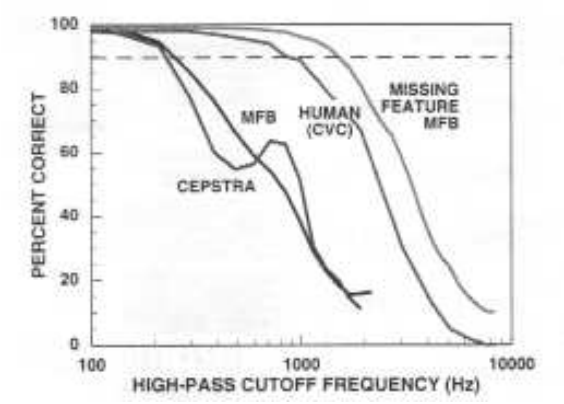


Figure 3.3: Decrease in correctness of HSR ("HUMAN"), MD ASR (" MISSING FEATURE MFB"), filterbank ("MFB") and cepstra ("CEPSTRA") based ASR with highpass filtered speech (reproduced from Lippmann and Carlson (1997))

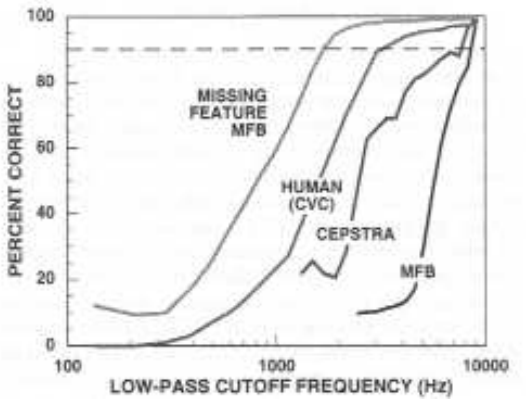


Figure 3.4: Decrease in correctness of HSR ("HUMAN"), MD ASR (" MISSING FEATURE MFB"), filterbank ("MFB") and cepstra ("CEPSTRA") based ASR with lowpass filtered speech (reproduced from Lippmann and Carlson (1997))

the basis of primitive features in the speech (Remez et al., 1994). (Barker, 1998) used a missing data recogniser to recognise sine–wave speech successfully with models trained on clean speech. Spectral peaks were used as features for recognition. During the recognition, spectral peak $x_{i,t}$ was identified in channel i at time frame t, according to:

$$peak(i,t) = \begin{cases} 1 & \text{if } x_{i,t} > x_{i,t-1} \text{ and } x_{i,t} > x_{i,t+1} \\ 0 & \text{otherwise} \end{cases} \tag{3.46}$$

The features in the peak positions were used for recognition, while the rest of the features in the vector were marginalised. A small improvement was also noted with models trained on the peaks from the clean speech with Viterbi training, despite the data sparseness (only a small fraction of the data are spectral peaks) compared to the "whole spectrum" models. Since the features used (64–channel ratemap) have fine frequency resolution and resolve the harmonics of the fundamental, selecting the spectral peaks effectively amounts to spectrum sampling at the multiples of the fundamental frequency F_0 – similarly to the double vowel identification model below.

Experiments with low and high–pass filtering showed a gradual decrease in the performance for missing data based recogniser similar to one observed with humans (Lippmann and Carlson, 1997) (Figures (3.3) and (3.4)). Provided that:

- the missing bands are known in advance, and the models are adapted correspondingly (in this case a low and high frequency filter were used, with known cut–off frequencies),
- there is no contextual information which humans can make use of, and machines can not (in this case nonsense CVC syllables ware used),

Under these conditions, the curves of performance decrease for humans and missing data based recogniser are parallel in shape. It should be taken into account that the human performance is for much harder task of nonsense syllables recognition, while machine recognition is for digits recognition task (perplexity of 6900 vs. 10). The recogniser used marginalisation without additional constraints. The identity of the missing features was known a–priori.

A joint psychophysical and modelling study assessing the intelligibility of band–pass filtered speech for humans and a missing data recogniser (Cunningham and Cooke, 1999) indicated that:

- The intelligibility of the speech filtered through gammatone bandpass filters with centre frequencies in the 1-3kHz range remains high.
- The performance increases further when noise (up to a certain level) is added to an appropriately selected bands. It was speculated that the added noise acts as a counter–evidence explaining the spectral restoration effect.
- a missing data recogniser employed on the same task (the mask was known in advance) showed a pattern of performance similar to the humans, although at much lower level. The performance rose as the number of channels increased. However, for the higher channels the gap remained much larger then for the lower channels, indicating a deficiency in the feature extraction at the higher channels. Using the auditory induction constrained didn't improve the results for the recogniser. A gap in the performance at 750Hz was notable both with human listeners and the missing data recogniser.

de Cheveigne and Kawahara (1999) constructed a missing data based model of vowel identification. On a simple task of identification of five synthetic vowels,[11] relying on the shape of short term smoothed spectrum[12] was found lacking. With increase of F_0, the effects due to truncation and aliasing could not be ignored, and adversely affected the distance measure between the template T_i and the target example T. A missing data model where the short term spectrum was sampled at multiples of estimated $\hat{F}_0$ and matched against the template at these points only, was found to be resistant to increases in F_0. The new distance measure was implemented as a weighting function $W(f)$ to the squared spectral distance $D(T, T_i)$:

$$\begin{aligned} W(f) &= \sum_{n=0}^{\infty} \delta(f - n\hat{F}_0) \\ D(T, T_i) &= \int [T(f) - T_i(f)]^2 W(f) df \end{aligned} \tag{3.47}$$

A fundamental frequency tracker is needed for F_0 estimation. It was discussed that if it can estimate the p.d.f. of the F_0[13] instead of a single value, it would provide natural adaptation when F_0 can not be estimated or is unlikely to be estimated well.

3.6 Summary

In this chapter the idea that parts of the speech spectrum may be obscured by sounds from other sources in a multisource acoustic environment and thus effectively removed from the recognition process was introduced. It is termed missing data, or missing features in speech. Arguments supporting the idea, coming from observations of how humans handle natural auditory scenes, experiments on humans with artificial stimuli, physiological evidence from various levels in the auditory chain and arguments from a signal processing perspective were put forward. On this basis, the missing data approach to robust ASR was formulated. It envisages that the recognition should be performed in two distinct steps:

- identification and grouping of the evidence coming from the speech source of interest
- adaptation of the recogniser, and recognition of the partial speech

It was speculated that the approach should be adaptable and robust to various speech degradations.

Next, the techniques for pattern classification (training and recognition) were reviewed. They can be systematised into two broad classes:

[11] well separable in the $F1 - F2$ plane

[12] derived by Fourier transform of the spectrum, truncation of the coefficients above $2F_0$ and inverse Fourier transform

[13] the Dirac δ function is one extreme case – the F_0 estimation process could assume less constrained parametric form for the p.d.f. and estimate its parameters, e.g. assuming a Gaussian p.d.f. and estimating the variance in addition to the mean

- techniques relying on marginalisation of the missing data
- techniques using the knowledge of the present data to impute estimates of the missing data

Some pattern classifiers (typically those estimating the data p.d.f. as part of the learning process) are more amenable to adaptation for handling missing data then the others (typically estimating the conditional densities only, not the joint density between the data and the classes).

In the last section previous missing data studies in connection with speech and speaker recognition and speech perception were reviewed. Relations to other models for achieving robustness were also highlighted, and the ways of combining the missing data approach with other techniques for robust ASR was discussed.

Chapter 4

Missing data identification

4.1 Introduction

The first task in the proposed missing data based robust ASR system is to identify (with a certain degree of confidence) the reliable regions of features on which the recognition is going to be based.

The way humans do this seems to be by auditory scene analysis (ASA) (Bregman, 1990), utilising principles (prior knowledge) reaching back to the physics of sound. Computational ASA (CASA) (Brown and Cooke, 1994) tries to devise models for separation of the speech from the mixture of sounds in a similar manner to ASA. The techniques will be discussed in the next section.

Sound sources are physically independent entities and therefore the corresponding signals in the mixture will be independent by default.[1] With loose assumptions about the nature of the mixing process, this knowledge can be used to devise a transformation that will increase (some measurement of) the independence between the sounds in the transformed mixture, compared to the original mixture of sounds. This class of techniques will be considered in Section 4.3.

The identification of the reliable parts of the spectrum can be also achieved by measuring the local SNR in each time–frequency point. Since neither the clean speech nor the noise are available, a model for their combination has to be assumed. Then, either the clean speech or the noise estimate are derived from the noisy speech. The methods for achieving this are discussed in Section 4.4.

4.2 Auditory scene analysis

Auditory Scene Analysis (Bregman, 1990) refers to the ability of the auditory system to segregate/decompose mixtures of sounds that enter the ears into their corresponding constituents originating from the same source. It is a theory of hearing. It lays down the principles governing the process of putting together the signals coming from the same physical source into a coherent perceptual stream. This process in audition seems less obvious then a similar one in vision, where it is immediately clear that many objects will enter the field of vision. So, most of the potential applications in audition (e.g. ASR) assume quiet, single source environment where the problem of forming perceptual streams that correspond to the auditory "objects" (sound sources) is "solved" apriori, as all signals originate from that source.

Today there seems to be a consensus between the researchers that (Cooke and Ellis, 1999):

- The processes of perceptual organisation operate on some kind of time–frequency representation. Typically these representations are the firing rates of the auditory nerve.
- There are bottom–up (BU) rules giving rise to cues about which parts of the speech in the time–frequency plane might originate from the same physical source. This is called primitive grouping.

[1] they may not be in music, for instance

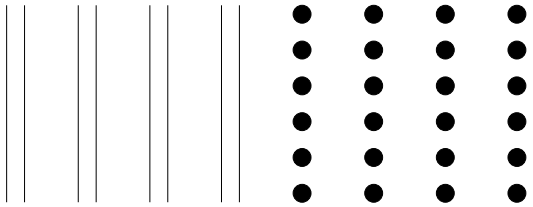


Figure 4.1: Illustration of the law of proximity: the closer lines tend to form pairs; we tend to see vertical stripes of dots since the horizontal distance between the dots is bigger then the vertical (after Katz (1951, pp. 24))

- There is stored knowledge about the familiar sound patterns, e.g. speech. The representations of the familiar patterns are called *schemas* and the mechanisms *schema–driven*. The schemas need not be complete speech units, nor even speech. "High–level" speech schemas (if existed) would be similar to the speech models in the ASR. The problem of segregation of sources can be solved (to a certain extent) even using schemas alone. Again, if schemas were complete speech units, the HMM decomposition (Section 2.8.2) technique would be an example of completely schema driven separation.

The bottom–up rules are of special interest for robust ASR. They can be considered to be source independent apriori constraints about how likely is that a group of features comes from the same source. They facilitate the process of *streaming*, in which disparate signals reaching our ears are grouped together as coming from separate sound objects/sources. The BU constraints are not utilised in present ASR systems. Both in audition and in vision these rules draw on a school of thought termed *Gestalt psychology* and developed in the beginning of the 20-th century (Koffka, 1935; Ellis, 1955; Kohler, 1947). Gestalt psychologists explored the principles that humans use to group sensory evidence together to form coherent objects in general (but with interest primarily in vision). The rules they derived are applicable to audition, too (Katz, 1951):

- the law of *proximity*: with all other things equal, the elements closest to each other tend to form groups. Figure 4.1 is an example of visual proximity. The acoustic components tend to be group together according to their proximity on the time–frequency plane, too.
- the law of *similarity*: when more then one element type is present, similar elements tend to form groups. Figure 4.2 illustrates this law in vision. In audition, sounds with similar pitch, timbre, intensity or spatial location tend to form groups.
- the law of *closed forms*: with all other things equal, the lines enclosing a surface tend to group together. Figure 4.3 is an example of the law applied in vision. The formulation of the law is clearly motivated by vision (e.g. "lines", "surface"). In audition, its meaning would be that the grouping tends to support whole (complete) perceptual forms. The phoneme restoration effect (Warren, 1970) is one such example. Listeners are unaware of absence

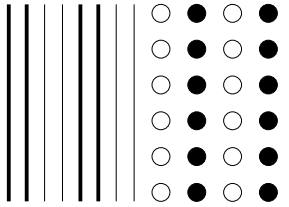


Figure 4.2: Illustration of the law of similarity: similar objects (the thick vs. thin lines, the full vs. the empty circles) tend to form groups (after Katz (1951, pp. 25))

Figure 4.3: Illustration of the law of closed forms: lines enclosing a surface in (b) tend to form a group while they group exactly the opposite way in (a) (after Katz (1951, pp. 26))

of short segments of speech when they are replaced by louder noise. Similarly, when parts of speech are obscured by noise bursts, the listeners report hearing the whole sentence and noise bursts, not a sentence interrupted by noise bursts. The sentence is perceived as complete. The auditory system seems to selectively pick the evidence to support a whole sentence hypothesis, as if searching for a simpler, rather than more complicated explanation. Another example is hearing the pitch of a complex sound to be at a frequency containing no energy at all. Taking the resolved higher harmonics into account, the auditory system judges that the evidence outweighs the lack of energy at the frequency of the fundamental.

- the law of *"good" contour*, or *common fate*: parts that have a "good" contour, or common faith, tend to group together. Figure 4.4 is a visual example. In audition, sounds tend to change slowly. Speech is smoothly changing signal, because it is produced by a physical system of slowly moving articulators. The auditory system expects continuous sound contours–sharp discontinuity in frequency, intensity or spatial location may signal a new sound source entering the auditory scene.

- the law of *common movement*: elements get grouped together when they move in the same time and/or similar manner. In hearing, common start/stop time (onset/offset) of the harmonics, common change in their amplitude (amplitude modulation–AM), or change in their frequency at the same time (frequency modulation–FM) tend to indicate that components belong to the same source.

- the law of *experience*: the comprehension of symbolic forms depends on the circumstances under which they were learnt. This law looks slightly at odds with the previous five. It can be considered as an acknowledgement by the Gestalt psychologists that not only bottom–up process shape our perception (which is what the previous laws are about). Top–down processes (dependent on the "schemas" or models acquired through experience) also play a part in it.

It should be noted that in the real world the grouping principles compete mutually. Since real world objects are much more complex then mere lines and dots (used here to illustrate the laws),

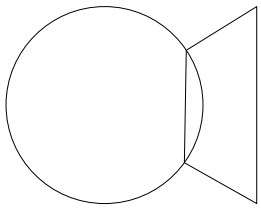


Figure 4.4: Illustration of the law of good contour or common faith: we see a circle and a trapezoid because parts of each have a common destiny (after Katz (1951, pp. 26))

all or most of the principles have a say in the process of binding the sensory input to the perceptual streams. Similarly, in audition, pure tones are rarely heard in the real world. The natural sound reaching our ears every moment can be considered as a cacophony of pure tones. Further, the cues work in concert to either mutually support a particular grouping or contradict each other. There is also the issue of quantification – *how much* does a cue support particular grouping. There must be a mechanism to resolve the contradictions and come up with a single solution to the binding problem. It is interesting that the Gestalt psychologists have observed that the solution may change (and even oscillate) over time, especially when several competing groupings seem equally possible (Kohler, 1947, pp. 171).

Just as "visual illusions" offer an insight into the mechanisms of visual perception, researchers in audition have designed experiments which reveal aspects of auditory processing. In a recent ASA review Cooke and Ellis (1999) identified the following following cues as likely to be significant:

- synchronous transitions across frequency regions – common onset or offset
- correlated envelopes in different frequency channels, i.e. correlated frequency modulation
- unresolved harmonics – channel envelopes with periodicity at F_0
- resolved harmonics – peaks in the spectrum at frequencies which are multiples of F_0
- interaural time difference arising from the different path lengths of the sound to the ears
- interaural level difference due to head shadowing
- across–time similarity of whole–event attributes such as pitch, timbre
- long interval periodicity giving rise to the perception of rhythm

If we envisage a system where the various grouping cues reinforce or contradict each possible grouping, something similar to a generalised Viterbi search in ASR may finally resolve the conflict and come up with the most likely solution to the binding problem (Barker et al., 2000). In speech the bottom–up processes are intermingled with the top–down schema driven processes making it harder at assess their respective contributions to the binding.

4.2.1 Computational Auditory Scene Analysis

Computational auditory scene analysis (CASA) (Brown and Cooke, 1994)[2] is a new discipline involved in building computational models with techniques based on conclusions of the perceptual studies of ASA (as performed by humans). In the past most of modelling effort in the speech community was on tackling the problem of ASR alone. The CASA models are much more general and aim to model the hearing process in general. They are of varying complexity and are built with various aims: from low–level simulations of neurophysiological processes, through mid–level simulations of some of the simpler tasks facing listeners (like double vowel segregation), to systems attempting to separate mixtures of whole sentences (thus incorporating the notion of time in the system) and/or real–world noises and acting as front–ends to ASR systems.

All systems roughly consist of two major parts. The first part performs some kind of time–frequency analysis analogous to the one performed by the ear as an initial step. It is usually through a filterbank with impulse response roughly matching the one of the human ear. In some systems the neural response of the auditory nerve to the stimulus is modelled, too. Next, some of the cues considered important for grouping (onset/offset, harmonicity, etc.) are computed and represented in some form. The second major part of the system differs, and the systems fall into two major groups (Cooke and Ellis, 1999):

- Weintraub (1985); Cooke (1991); Brown (1992); Wang and Brown (1999) group the information in a bottom–up fashion in the next phase. The elements are grouped together if there is evidence for that.

[2] see Rosenthal and Okuno (1998) and references therein for a recent cross–section of the field

- Cooke et al. (1993); Ellis (1996); Nakatani et al. (1998); Godsmark and Brown (1999) continue by generating hypothesis about the possible groupings and matching them with the acoustic evidence. Sometimes the systems are termed as "blackboard architectures" or explanation based systems.

Bottom–up systems

Weintraub (1985) attempted separation of two simultaneous speakers using the estimated pitch period of the voices. A hand crafted seven state Markov model, with states corresponding to silence, periodic, non–periodic, onset, offset, increasing and decreasing periodicity, estimated the power spectrum. A dual pitch tracking algorithm was used to devise and assign the signal energy to the spectrums of the both voices.

Cooke (1991) computed time–frequency tracks called "synchrony strands" from the output of the auditory periphery model. Their formation was guided by local similarity and continuity constraints. They encompassed the dominant periodic components. The grouping algorithm used the harmonicity cue for the lower frequency channels and amplitude modulation for the higher frequency channels.

Brown (1992) used similar decomposition into synchronous partials. In addition, the pitch of each partial was computed by combining a summary autocorrelation across the channels with the local autocorrelation, giving rise to a pitch contour for each partial. The systems then searched for groups with common pitch contours. Among them, the ones with common onsets were given preference in grouping. Figure 4.5 depicts an example of grouping, segregation and the resulting representation of the speech speech (the interfering noise is a siren).

The oscillatory correlation model of Wang and Brown (1999) departs from the symbolic representations used in the previous (and the next) models. The low – level representation is nerve firing probability model, while the mid – level representation involves pooled (across channels) correlation for pitch detection, and cross-correlation between the adjacent channels. However, the search stage, where the grouping occurs, is replaced with a two layer network of oscillators. The first layer consists of relaxation oscillators excited locally and inhibited globally. The oscillators in the second layer are linked by lateral connections. Together, both layers implement the cues of proximity (elements close in time and frequency are grouped together) and good continuation. The first layer produces smaller structures, while the second layer groups these structures into streams. The neurophysiological findings give this architecture a neurobiological foundation. The model is also consistent with the parallel and distributed processing paradigm.

Top–down systems

Cooke et al. (1993) and Godsmark and Brown (1999) used a "blackboard" architecture for their CASA systems. The blackboard system has a global data structure (the "blackboard"), accessible by the modules that implement various rules (the "experts"). If some expert concludes that the current data on the blackboard gives it a reason to act, it does so under a control of the blackboards' monitor module (the "scheduler"). The scheduler orders the experts' actions in time. The result of an action is change in the blackboards' configuration. The experts in the systems implemented grouping by common F_0, common amplitude modulation and common onset/offset time.

Ellis (1996) utilised a "prediction – driven" architecture in his system. It segregates the sounds by "explaining" the predictions generated by the internal "audio world model" with the acoustic evidence. This is essentially a search for the most probable hypothesis in the space of possible hypothesis–explanations of the observed evidence. The search is mostly heuristic. The mid – level sound elements are: *noise clouds* of unstructured noise energy; tonal elements with perceived periodicity – *wefts*; and *transients* i.e. rapid bursts without pitch or any other structure. The system works in terms of prediction and reconciliation: depending on the current evidence, predictions are made about the speech in the future time instant. The predictions are probabilistic in the the sense that the mean/expected value is generated together with a margin of uncertainty (the expected error). Each hypothesis ties together a set of mid – level elements. As new evidence

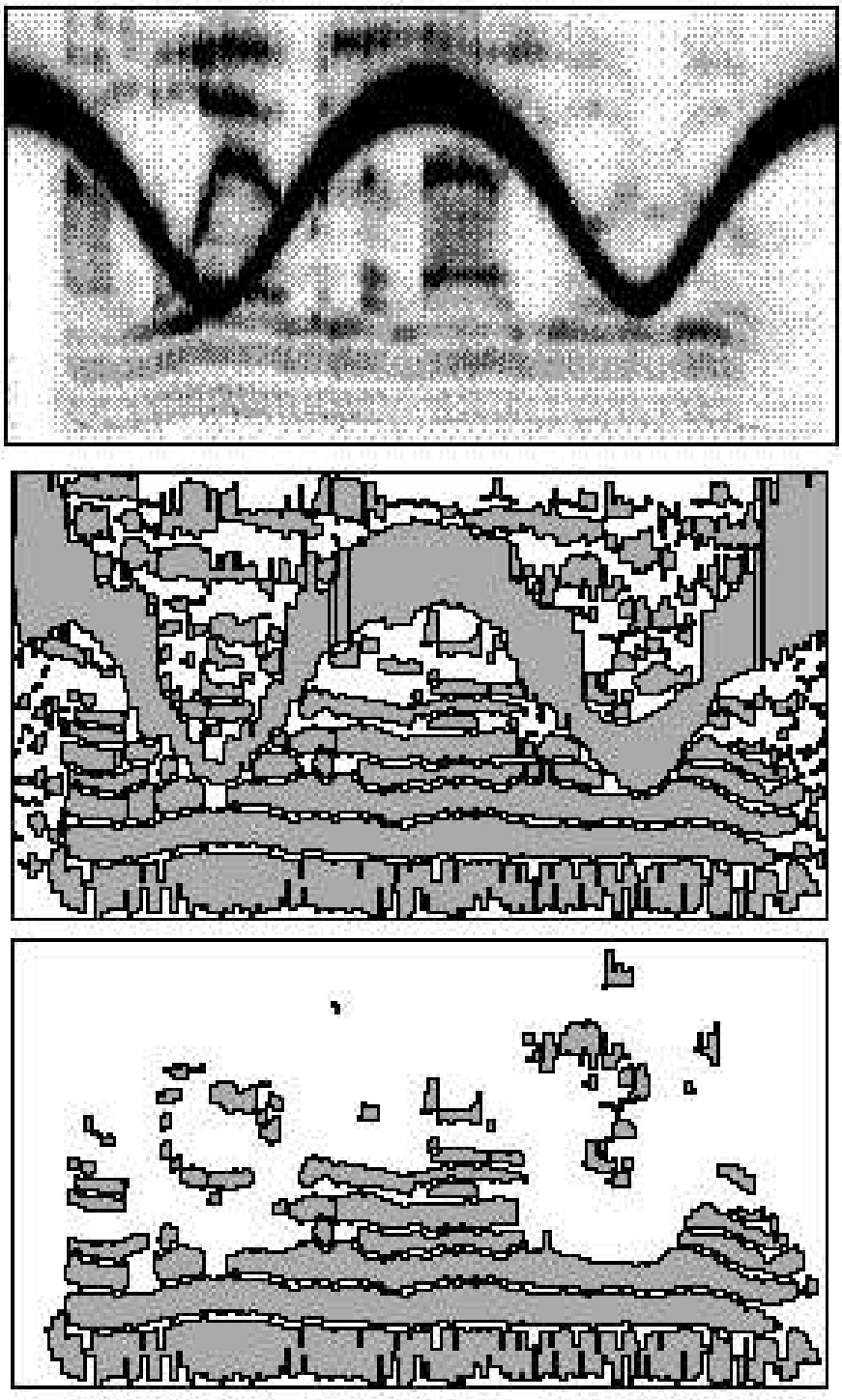


Figure 4.5: The time – frequency – firing rate representation of an utterance mixed with siren (top); symbolic auditory time – frequency representation produced by Brown (1992) system (middle); time – frequency representation after grouping and removal of the siren (bottom) (reproduced from Cooke et al. (1994b)).

arrives, if the energy of the signal is within the bounds of the expected signal then the grouping doesn't change. If there is a surplus of energy, additional elements are hypothesised to account for the energy. If the observed signal lacks energy, then some of the elements will be terminated in a consistent manner.

This is not dissimilar to the system by Nakatani et al. (1998). Here, another paradigm for the blackboard architecture is formulated: the data is explained through a "cooperation" of independent "agents", specialists for particular task. There are modules ("agencies") for creation and destruction of agents depending of whether there is surplus or lack of energy compared to the prediction. The architecture is termed "residual driven". The system is binaural and uses binaural harmonicity and localisation cues.

Performance

Despite the progress made toward a CASA system that will approach human performance, it is widely recognised (the authors of the systems included) that the present systems have a huge gap to bridge. Even the "second generation" of "top–down" systems which take into consideration certain high level constraints do not approach human audition. Still, in all cases a worthwhile improvement in SNR was noted. Lately, it has been argued (Roweis, 2000) that a statistical approach may be better suited to the problem of reconciling the different and sometime conflicting grouping cues.

4.2.2 Integration of CASA and ASR

Enhancement

The integration of CASA and ASR happened quite early – one of the very first systems by Weintraub (1985) assessed the quality of the separation by running the separated speech through an ASR system and noting the resulting accuracy. The same path has been taken by the most subsequent systems: the CASA system separates the speech from the "background" first, the speech is resynthesised from the fragments assigned to it, and the ASR system tries to recognise this speech next. It is not hard to see why this approach of integration is so popular:

- the task is clearly separated into two independent stages, making the construction of the systems easier
- systems that were not designed with integration in mind, can by simply integrated this way
- once CASA is considered as a "speech enhancement" process and the result of the analysis is speech, a number of criteria (objective and subjective) can be employed to assess the quality of enhancement (SNR, intelligibility, accuracy, etc.)

Typically, the approach yields an improvement compared to the recognition of the noisy speech (Okuno et al., 1999), but it has not been proven that it performs significantly better then other techniques for robust ASR (e.g. noise estimation, blind source separation).

The problems of the "enhancement path"

The "enhancement" combination of CASA and ASR has serious shortcomings. It is quite certain that speech schemas play a role in speech perception. There are experimental examples of schemas defeating the bottom–up grouping rules, as in the "duplex phenomena" where parts of the stimulus are interpreted as belonging to more then one speech source. The ASR system contains a detailed library of high level schemas – the speech models. When the CASA system is used merely as a preprocessor prior to ASR, the models are not utilised during the separation process.

Further, separation via CASA and subsequent resynthesis introduces various distortions in the speech signal that is input to the ASR system. The ASR systems operate best in matched conditions: when the speech they recognise is (statistically) the same as the material they were

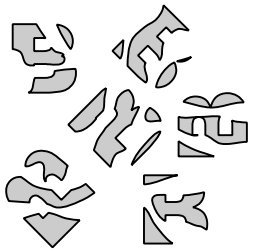


Figure 4.6: Visual occlusion (check Figure 4.7 for a hint – after Bregman (1990)).

trained with. The resynthesised speech may be very different from the clean speech, as the scope and the nature of the distortion introduced by CASA can not be easily assessed. This can be ameliorated by training the ASR system on a noisy speech that has been processed with the CASA system.

A notable problem with the "speech reconstruction" application of CASA to ASR is that CASA systems typically assign the energy of a point in the spectrum to one source only (the principle of exclusive allocation). When other sources are resynthesised, there is a hole in their spectrum. This was observed by several researchers (ex: Ellis (1998, pp. 4)) and does not arise only when separation is carried out with CASA, but with ICA (Section 4.3), too (ex: Choi et al. (1999, pp. 3)[3]). Unfortunately, the models that ASR systems use are based on coding the shape of the whole spectrum. Holes in the spectrum (that never occurred during the training) give rise to outliers to the p.d.f. of the spectrum shape. In a typical setup of a contemporary ASR system, an outlier in any dimension can seriously reduce the discriminability between the models.

The problem is the one of *auditory occlusion*: the locally loudest source obscures all locally quieter sources in the time–frequency representation of the signal. Section 3.2 lists the arguments in favour of this observation which is not so intuitive (occlusion in vision seems much more natural). Compare for example Figure 4.6 and Figure 4.7. In the former there is no reason to believe that the black lines and the grey fill of the letter shapes continue behind the white background since the white colour can not obscure black colour. However, the reverse is true, and the shapes reveal themselves naturally on the latter figure.

Using partial evidence from CASA for ASR

Cooke, Green, Anderson, and Abberley (1994b); Cooke, Crawford, and Green (1994a); Cooke, Green, and Crawford (1994c); Green, Cooke, and Crawford (1995) proposed an alternative integration of CASA with ASR: adapting the ASR system to be able to use partial evidence as delivered by CASA. The underlying assumption is that speech is redundant enough so that it contains enough information for recognition even if occluded. Chapter 3 discusses the assumptions and the subsequent work, while the following chapters in this work will detail the techniques and experimental results. It is obvious that the nature of adaptation depends on the architecture of the ASR system. Cooke et al. (1994b) approach is suited to a HMM based ASR system. Berthomier et al. (1998) used a multiband hybrid (ANN/ASR) system for recognition of partial speech delivered by a CASA system. The CASA system utilised the harmonicity cue alone for identification of the noisy bands. Two to the power of the number of bands recognisers were trained off–line to cover every possible combination of noisy bands. After CASA detected the noisy bands, a matched classifier was selected to perform the classification using the clean bands only.

But simply ignoring the evidence coming from the other sources may not be the best that can be achieved. The nature of the *masker* (the object occluding other objects) can put constraints on the nature of the *maskees* (the obscured objects). I.e. by knowing the masker one may not be able

[3]Wu et al. (1998a) have taken advantage of the exclusive allocation to bootstrap their ICA algorithm

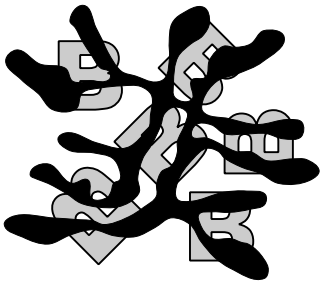


Figure 4.7: Visual occlusion with a hint (compare to Figure 4.6; after Bregman (1990)).

to determine the "correct" maskee, but it may be able to determine which maskees *did not* occur. The underlying principle of "auditory induction" was incorporated into a HMM ASR system by Green et al. (1995), and Holmes and Sedgwick (1986) proposed the same technique earlier coming from entirely different premises and motivation. In any time–frequency representation of speech, the masker puts a higher–bound constraint on what the energy of maskees could have been in that time–frequency point. Barker et al. (2000) used the constraint in a full–blown integrated Viterbi search for simultaneously finding the most probable grouping and the most probable sentence in the presence of noise.[4]

4.3 Independent component analysis for blind source separation

Simulating ASA is not the only way of recovering a signal from a mixture of signals. It can be seen as a specialisation of a well researched problem in the signal processing community – the problem of *blind source separation* (BSS). The problem of BSS is defined as separation (recovery) of the signals produced by several sources from their linear mixture, without any knowledge about the sources themselves (hence the term "blind"). In a typical BSS model there are N receivers picking up the mixture of the M source signals ($N \geq M$). This is related to the problem of blind deconvolution, which in the simplest case tries to find an inverse of an unknown filter which is convolved with one source (there are also extensions for multiple signals) by observing the resulting signal only. It is also somewhat related to the principal component analysis (PCA): geometrically, PCA seeks an orthogonal axis on which the observations are projected, such that the projections are uncorrelated (but not necessarily independent). Taking higher order statistics into account, the independent component analysis (ICA) tries to recover the axis of projection not necessarily geometrically orthogonal, but on which the data projections are statistically independent.

In the case of instantaneous mixing, and neglecting the noise in the process, the independent component analysis (ICA) assumes the mixing model (see (Hyvarinen, 1999; Lee, 1998) and references therein):

$$\mathbf{x} = A\mathbf{s} \tag{4.1}$$

where $\mathbf{s} = (s_1, \ldots s_n)^T$ are the sources which are independent, A is the unknown mixing matrix and $\mathbf{x} = (x_1, \ldots x_m)^T$ is the observed mixture. The independence of the sources is a stronger requirement then uncorrelation, since it implies:[5]

$$\mathcal{E}\{g_1(x_i)g_2(x_j)\} - \mathcal{E}\{g_1(x_i)\}\mathcal{E}\{g_2(x_j)\} = 0 \quad \text{for } i \neq j \tag{4.2}$$

instead of mere $\mathcal{E}\{x_i x_j\} - \mathcal{E}\{x_i\}\mathcal{E}\{x_j\} = 0$ for $i \neq j$ for any functions g_1 and q_2.

[4] discussed in more detail in Section 7.2.1

[5] $\mathcal{E}$ is the expectation, i.e. $\mathcal{E}\{g(x)\} = \int g(x)p(x)dx$

The aim of ICA is given a set of measurements $\{\mathbf{x}(t)\}$ to recover the the original signals $\{\mathbf{s}(t)\}$. Without getting too far into the problem of identifiability of the model, it can be said that at most one source can be Gaussian, usually there have to be more sensors then sources (although some solutions for the reverse case are not impossible) and the mixing matrix must be a full column rank. Even then, the matrix A can be recovered only up to a column permutation (the model puts no constraint on the order of the sources) and up to a multiplicative constant for each column (since it can be cancelled by dividing the corresponding source with it).

Methods for ICA consist of two major parts:

- an objective function to measure the amount of independence between the transformed measurements
- an algorithm that is used to perform the optimisation

As an illustration, we will present the maximum likelihood objective function (since it best suits the probabilistic framework in which the ASR systems are discussed here), and gradient descent based optimisation (since it is the simplest and most often used). We will also assume $M = N$ for simplicity. Since the sources $\mathbf{s}$ are independent, $p(\mathbf{s}) = \prod_{i=1}^{M} p(s_i)$. The log–probability of one observation $\{\mathbf{x}\}$ for a given matrix A is:

$$\begin{aligned} \log p(\mathbf{x}|A) &= \log \int p(\mathbf{x}|A,\mathbf{s})p(\mathbf{s})d\mathbf{s} \\ &= \log \int [\prod_j \delta(x_j - \sum_i A_{ji}s_i)][\prod_i p(s_i)]d\mathbf{s} \\ &= -\log|A| + \sum_i \log[p_{s_i}(\sum_j (A^{-1})_{ij}x_j)] \end{aligned} \tag{4.3}$$

where $\delta(x)$ is the Dirac impulse function,[6] and $p(s_i) = p_{s_i}(x)$ is the p.d.f. of a random variable s_i.[7]

However, in majority of cases the matrix of interest isn't A, but its inverse $A^{-1} = W$, since the knowledge of W allows for recovery of the sources $\mathbf{s}$ from the observations $\mathbf{x}$. So the log–probability of a single observations is:

$$\log p(\mathbf{x}|W) = \log|W| + \sum_i \log p_{s_i}(\sum_j W_{ij}x_j) \tag{4.4}$$

Denoting the "unmixed" sources (i.e. our estimate of) $u_i = \sum_j W_{ij}x_j$ and differentiating with respect to W:

$$\frac{\partial}{\partial W}\log p(\mathbf{x}|W) = W^{-T} + \begin{bmatrix} \frac{1}{p_{s_1}(u_1)}\frac{\partial p_{s_1}(u_1)}{\partial u_1} \\ \vdots \\ \frac{1}{p_{s_N}(u_N)}\frac{\partial p_{s_N}(u_N)}{\partial u_N} \end{bmatrix} \cdot [x_1 \dots x_N] \tag{4.5}$$

It has been argued that if the gradient ascent follows the so called "natural" gradient $\left\{\frac{\partial}{\partial W}\log p(\mathbf{x}|W)\right\} \cdot W^T W$ (Lee, 1998, pp. 56), the convergence is faster (and W need not be inverted to compute the update ΔW):

$$\Delta W \propto \{1 + \begin{bmatrix} \frac{1}{p_{s_1}(u_1)}\frac{\partial p_{s_1}(u_1)}{\partial u_1} \\ \vdots \\ \frac{1}{p_{s_N}(u_N)}\frac{\partial p_{s_N}(u_N)}{\partial u_N} \end{bmatrix} \cdot [u_1 \dots u_N] \} \cdot W \tag{4.6}$$

The same update formula can be derived for other objective functions: the entropy of $p_{\mathbf{s}}(\mathbf{u})$ and the mutual information between the $p_{\mathbf{s}}(\mathbf{u})$ and $\prod_i p_{s_i}(u_i)$. Hyvarinen (1999) is a detailed survey of different ICA variants.

[6] $\delta(x) = 0$ for $x \neq 0$ and $\int \delta(x)dx = 1$

[7] with the usual abuse of notation

ICA for speech separation

Separation of sound sources is one of the classical applications of ICA. Usually the signals are separated in time domain and entirely independently from the ASR system (Lee et al., 1997; Choi et al., 1999), and then fed into the system. Extended algorithms, handling delays in addition to the mixing between the sources are usually employed in the experiments. This setup is similar to a real life situation of several sources in a reverberant room. There are known problems with this approach:

(a) the "separated speech" contains low energy regions ("holes" in the spectrum) where the other source was dominant. A conventional recogniser can't handle that naturally. Choi et al. (1999) resorted to heuristics like thresholding. A missing data recogniser could handle this easily, if the locations of the "holes" were known.

(b) even after the separation, the interfering source can be heard in the background. The recogniser picks this nevertheless and resulting in a huge number of insertion errors. This "cross–hearing" is due to poor separation–this is closely related to the next problem.

(c) the quality of the separation is limited because the speech model that is used is crude. Even methods that do not explicitly assume the p.d.f. of the sources, do so implicitly (e.g. the squashing non–linearity in (Bell and Sejnowski, 1995) represents the cumulative p.d.f. of the sources, and its a–priori choice amounts to assumption about the form of sources p.d.f.). When assuming some speech distribution, Laplacian p.d.f. $p(x) = \frac{\alpha}{2}e^{-\alpha|x|}$ (for $\alpha > 0$) is usually chosen to express the sparsity of speech distribution. In contrast, the ASR recogniser not only has an extensive model about the speech distribution(s), but also a dynamic model of how the speech source (articulators) change and evolve over time. This is not utilised in the separation phase. It has been reported (on non–ASR task) that separation can gain significantly from a known signal distribution (Torkkola, 1996).

These problems are in addition to the problems already inherent to ICA that make the ASR application problematic:

- needs at least two microphones
- the resulting unmixing matrix is undetermined up to a scaling factor per column and column permutation

The latter problems have been addressed by Ikeda and Murata (1999) with some success, as discussed in the next section. However, it seems that the question of whether it may be better to use a more appropriate model than ICA in the first place is still open.

ICA in the spectral domain

A possible solution to some of the problems may be to perform the separation in the spectral domain. The accuracy of the separation seems to be much better there (Zibulevsky and Pearlmutter, 1999)[8]. Ikeda and Murata (1999) reported on ICA for spectrogram–like speech features. The ambiguity of the scaling factor was resolved by projecting the signals back onto the observations space and comparing the projections with the true observations. The ambiguity of the permutation was handled by imposing continuity constraints on the separated signals.

Wu et al. (1998a) used the missing data assumption that only one of the sources is dominant in the mixture to speed up the iterative process at no loss of accuracy. Both methods (Ikeda and Murata, 1999; Wu et al., 1998a) have also been applied to the case of more signals then sensors. In fact, nothing in the derivation of the learning rule in Eq. (4.6) requires equal or grater number of sensors then sources.

[8]The conclusion is based on Figure 7 on page 13. The figure compares the newly proposed method with few other ICA methods. However, it is notable that all methods exhibit relative separation errors in the spectral domain almost an order of magnitude smaller than in the time domain

However, ICA application on the spectrograms introduces a new problem: treating the bands in the spectrogram as independent sources is quite unjustified. But if this constraint is not enforced at all, and the assumption that the p.d.f. of the sources is factorisable is removed, then instantaneous ICA degenerates into a Maximum Likelihood Linear Regression (MLLR–commonly used for speaker adaptation) with gradient descent search for the best (un)mixing matrix (instead of the more usual EM algorithm).

An example of a more accurate model would be:

- for each feature vector $\mathbf{x}$, two subsets of features exist: one subset coming from the speech source, and the other subset generated by the other (noise) source
- there is an accurate model of the speech source, and no model or very weak model (some a–priori assumptions like Laplacian density or on–line single channel noise estimation) for the other source
- the subsets change on a per–frame basis
- we have prior knowledge that the subsets are usually localised in the time–frequency plane, i.e. it is more probable that a whole patch belongs to one source

The Multidimensional ICA model (MICA) (Cardoso, 1998) is an extension of ICA for handling multidimensional signals. This is equivalent to assuming that in the unmixed space groups of features are going to come from the same source and thus are not going to be independent. MICA model rewrites the ICA model (Eq. (4.1)) as sum of independent components model (similarly to PCA). Unlike ICA, the MICA model is uniquely determined. However, we are not aware of a general algorithm for MICA. The model was tested by performing ICA on the data first, and then manually selecting the non–independent components to be "merged". Also, it was assumed that the subsets of the components are constant. In the auditory scene the subsets change on a per–frame basis.

4.3.1 ICA and CASA

Given that both CASA and ICA have been applied to the same task of speech separation from a mixture of sounds, it is interesting to see how they compare. Although both rest on the same premise of independence of the physical sources, ICA is more data driven, while CASA is more knowledge driven. CASA typically operates with one sensor (or two if binaural cues are used), while ICA needs two sensors or more. Further, ICA usually needs at least as many sensors as sources.

van der Kouwe, Wang, and Brown (1999) compared CASA and ICA on a 100 speech and noise mixtures from Cooke (1991) (there are 10 voiced utterances mixed with 10 noises). The CASA system used was a monaural one by Wang and Brown (1999). The ICA algorithm performed joint approximate diagonalisation of eigen matrices (JADE) (Cardoso (1997)) on two linear mixtures of the speech and the noise. Comparison of the SNR before and after the separation showed that CASA performed better then ICA on 2 of the noises, while ICA performed better on the remaining 8 noises. CASA favoured locally narrowband, continuous and structured noises (1 kHz tone and siren), while ICA performed better on the rest of the noises.

Okuno et al. (1999) combined CASA and ICA trying to cover their respective weaknesses and combine their advantages. The combined system acted as a speech enhancement preprocessor to an ASR system. When used alone, ICA leaded to better accuracy in two–speaker scenes, while CASA was more successful with three or more speakers. Their combination achieved better accuracy then either of them alone.

4.4 Noise and Local Signal-to-Noise Ratio estimation for separation

Single channel noise and local SNR estimation[9] techniques can be used for identification of the reliable parts of the spectrum. Their estimate can be either thresholded yielding hard missing/present data separation, or used as reliability measure of the points in the time–frequency (T–F) plane. We are going to consider the techniques which use a mixture of the signal and the noise to obtain their estimates. Typically they assume the additive speech and noise combination in time $x(t) = s(t) + n(t)$ and power spectrum $X^2(w) = S^2(w) + N^2(w)$ domains. They are adaptive without explicit speech/silence detection.

Martin (1993, 2001) tracks the minimal envelope of the noisy speech x smoothed power spectrum $\bar{Y}_x(t)$:

$$\begin{aligned} Y_x(t) &= Y_x(t-1) + x^2(t) - x^2(t-1) \\ \bar{Y}_x(t) &= \alpha \bar{Y}_x(t-1) + (1-\alpha) Y_x(t) \end{aligned} \tag{4.7}$$

Next, the time window of approximately $L = 0.625$ sec is divided in few W (ex: $W = 4$) smaller windows, and the minimum of the smoothed spectrum $\bar{Y}_x(t)$ in each of the windows is determined. If the sequence of the windows minima monotonically increases, then it is assumed that the noise increases rapidly and the minimum of the last window is the noise power. Otherwise, the smallest value of all windows minimas is the noise power. This value is multiplied by an overestimation factor (from 1.3 to 2, depending on the length of the windows used for power estimation and minima calculation), and is bounded from above by the power of the speech plus noise mixture. The estimator is biased when there is no speech present.

Ris and Dupont (2001) used a variant of the same algorithm of minima tracking. For each frequency band, on the basis of N consecutive frames, n (typically $n = N/5$) minimal values were averaged to estimate the noise level in each band.

Hirsch and Enrichter (1995) proposed two algorithms for noise estimation. Both operate on the outputs of a filterbank in the spectral magnitude domain.

The first algorithm calculates weighted first order average for each channel separately, on a per–sample basis. When the energy in the channel exceeds a certain threshold (the threshold is set to be the last computed average scaled by an overestimation factor of $\beta = 1.5$ to 2.5), it is considered that a speech segment starts and the recursive computation is stopped. The calculated average thus far is taken as the value of the noise energy at that moment:

$$N(t,w) = \begin{cases} \alpha N(t-1,w) + (1-\alpha) X(t-1,w) & \text{if } X(t,w) \leq \beta N(t-1,w) \\ \alpha N(t-1,w) & \text{otherwise} \end{cases} \tag{4.8}$$

Ris and Dupont (2001) noted that the above method overestimates the noise in low SNR. Therefore a second order recursion was added to compute the variance of the noise estimate, in addition to its mean:

$$var_N(t,w) = \alpha \; var(t-1,w) + (1-\alpha)(X(t,w) - N(t,w))^2 \tag{4.9}$$

Further, the frames were grouped in time segments of several frames, the estimates of the mean and the variance of the noise were computed on a per segment basis, and for all the frames in the new segment Eq. (4.8) and (4.9) were applied in the frames where:

$$|X(t,w) - N(t-1,w)| \leq k \cdot \sigma_N(w) \tag{4.10}$$

where $\sigma_N(w)$ is the deviation of the previous segment.

The second algorithm reported by Hirsch (1993) is based on computing the histograms (with roughly 40 bins) of the noisy speech energy which is below the threshold computed with Eq. (4.8)

[9] local SNR estimation refers to estimated SNR in each point in a T–F plane, as opposed to some average of the SNR along the time and/or frequency axis

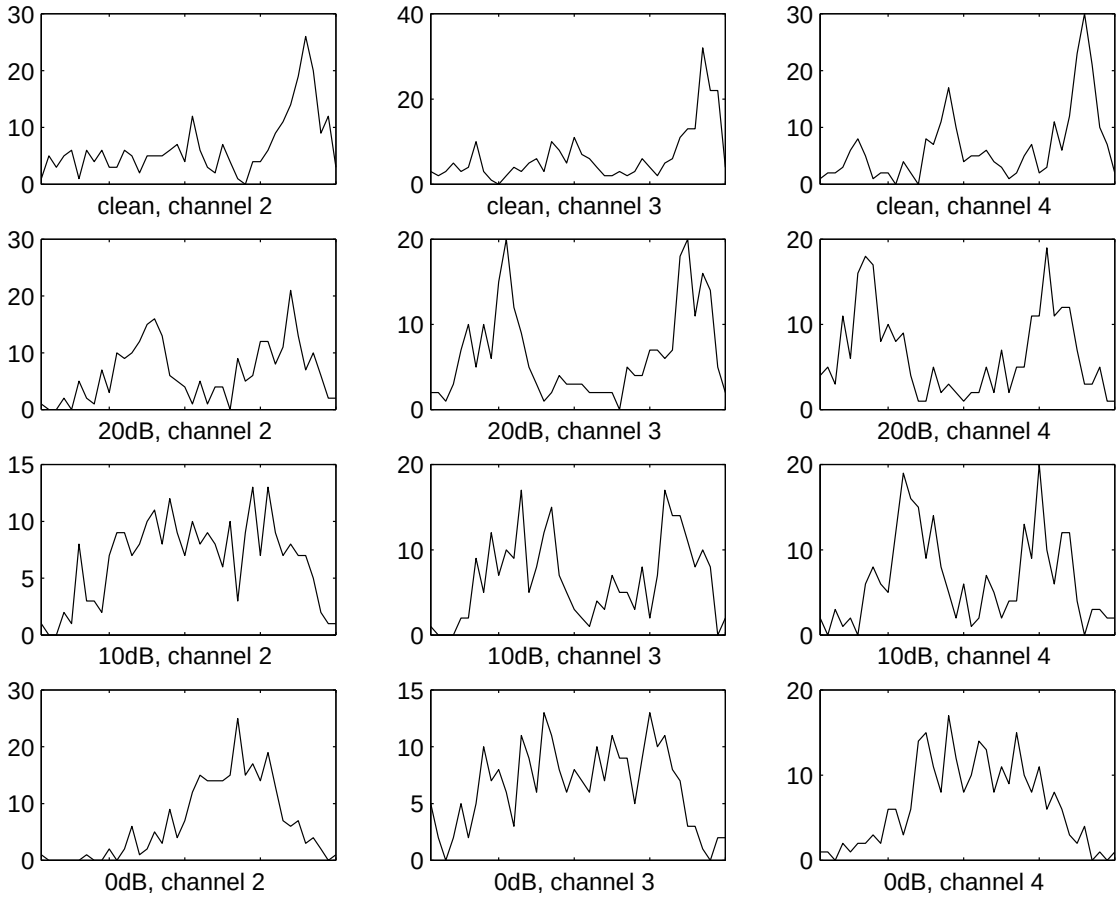


Figure 4.8: Histograms of the second, third and fourth channel of a 24 channel log–filterbank in clean conditions (top row), mixed with factory noise at global SNR at 20dB (second row), 10dB (third row) and 0 dB (bottom row). The two bumps notable at 20 dB and 10 dB tend to merge into a single one at 0 dB.

in each subband over a 400ms time window. The values above a threshold are not take into account – they are considered to belong to speech, not noise. The histograms are computed using the values below threshold for each channel separately. The maximum of the distribution is taken to be the value of the noise energy in a particular channel.

Similarly, it was noticed that the histograms both of the clean and the noisy subbands feature two distinct peaks (Ris and Dupont, 2001)[10] (Figure 4.8). The low energy mode is related to the presumed speech pauses or noisy frames, and the high energy mode is related to the speech frames. The modes are well separated for clean speech and mid–SNRs, and get closer as the SNR decreases. For low SNR they finally merge into one single mode. Therefore, the distributions in each band were modelled by two Gaussians and EM and K–means clustering algorithms were used to fit a two Gaussians mixture to the histograms. The difference between the means of the Gaussians is directly related to the local SNR of the noisy speech in each band.

A method based on subbands quantile[11] filtering was reported by Stahl, Fischer, and Bippus (2000). Again it is assumed that the speech pauses in the subbands are going to be filled with noise as the noise level increases. Quantile with $q = 0.55$ was used for the experiments as the

[10] McAulay and Malpass (1980) mention the same idea, the first reference seems to be Roberts (1978); rediscovered and reported in length by Ris and Dupont (2001)

[11] q–quantile is the minimum for $q = 0$, the median for $q = 0.5$ and the maximum for $q = 1$

subband noise estimate. It should be noted that if the window of the analysis is not long enough to encompass enough speech silences, then the quantile is essentially going to estimate the speech (not the noise), similarly to median smoothing commonly used in image processing to decrease the effect of the impulsive noise.

All methods mentioned above rely on the assumption that there are going to be enough frames with silent speech that are going to be filled with noise as the SNR decreases. However, if the frequency resolution is good enough to resolve the harmonics of the F_0, which are going to correspond to the spectral peaks (see Section 3.5.7), the spectral valleys between them are going to contain low speech energy. They are going to fill with noise first, and provide enough data for noise estimation. Therefore, the reliance on speech pauses can be lessened and the segments over which the histograms and the averages are computed can be shortened. Ris and Dupont (2001) used a 64ms long window for this purpose.

Meyer, Simmer, and Kammeyer (1999) assessed the performance of some of the above algorithms. The task was estimation of the noise given the mixture of clean speech and slowly amplitude–modulated noise. All algorithms perform bad with rapidly increasing noise and better with decreasing noise. In the latter case, Martin (1993)'s algorithm performed better then Hirsch's weighted algorithm at SNRs close to zero.

Ris and Dupont (2001) tested some of the methods described above with six types of noises, both artificial and realistic. They measured the mean square error (MSE) between the true and estimated noise level in the 700 Hz to 1600 Hz region. The results did not indicate any of the the techniques to be a clear winner. Depending on the noise type, different noise estimation methods came best. The only consensus seems to be that all noise estimation methods preferred better frequency then time resolution.

Any of the above techniques can be used used in an ASR system which can perform classification with partial data. In that case:

- the negative spectrum (improbable, assuming enough smoothing of the power spectrum) can be treated as missing (Drygajlo and El-Maliki, 1998a)
- a threshold for the estimated local SNR can be set to to separate the features as present or missing
- the estimated SNR can be used as a measure of how reliable the feature is

In this sense, the local SNR estimation techniques can be used for speech separation.

4.5 Summary

Techniques for speech source separation were discussed in this chapter. CASA uses the low level properties of sound that are believed to be used by humans to facilitate separation. ICA assumes independence of the sources and the way they combine to yield the observations, and then transforms the signals trying to enhance their independence in the transformed space. The local noise and SNR estimation estimation techniques use well known heuristics about the speech and the noise to build models of the noise from limited data.

None of the techniques claims to be the complete solution to the problem of separation. They all have problems hindering their application. However, each on its own may deliver certain constraints that would make the ASR search for the best explanation of the data more accurate.

For example, having the missing data of speech in mind, on–line noise estimation might provide some noise model estimate. ICA might use that model together with the speech models of the recogniser to update the probability of two sets of features coming from different sources by assessing their independence. CASA might deliver another update of what features are likely to have come from the speech source.[12]

[12]the speech model $p(\mathbf{x}_p, \mathbf{x}_m|S)$ can not be used to assess the independence of the features $\mathbf{x}_p$ and $\mathbf{x}_m$ because the $\mathbf{x}_m$ features have not come from the speech source.; i.e. $p(\mathbf{x}_p, \mathbf{x}_m)$ does not model the joint distribution of the speech and the noise.

Chapter 5

Robust ASR with missing data in an HMM system

5.1 Introduction

The aim of this chapter is to show how techniques for handling missing and unreliable data can be integrated in today's standard ASR systems. These systems model the speech source as a Hidden Markov model (HMM). They assume that, during recognition, the data the HMM is matched against was generated by a single speech source. However, as discussed previously, in most real–life situations this is not true. The observations picked up by a microphone are a mixture of several sound sources. Human audition has an intriguing property of being able to attend to one source in the mixture alone. The ASR systems fail completely in such conditions. This is not a fault of the speech HMM. Its parameters were inferred during training with the assumption that all features in the multidimensional observation were generated by the speech source. As argued in Chapter 3, a range of phenomena like: (a) easy handling of bandwidth restrictions and severe alterations in time–frequency (T–F) plane; (b) evidence of winner–takes–all physiological processes; (c) psychoacoustics findings maybe offer an insight into some of the principles that underline the robustness of the human speech recognition (HSR).

In this chapter the same principles are applied to a HMM based ASR system. After a brief introduction into the structure of such system, the missing data (MD) model for speech recognition is introduced. In addition to the "standard" model of the speech source, the MD model envisages a model of the *mask*. The mask model determines the robustness of the speech features. Two techniques can be used to implement the MD model: marginalisation and imputation. They are rooted in two complementary statistical approaches for treating missing data. The MD model accounts for a non–stationary environment by frame-by-frame on–line adaptation. The aim of the adaptation is to match to the speech HMM the features that originated from that speech source only. Features originating from other sources can be matched to their models, if such are available. If not, the MD recogniser can make the best out of the available data. A variant of the MD model can be used for separation, as the separation and the recognition are tied together.

5.2 An outline of an HMM based ASR system

The operation of a typical contemporary continuous density HMM based ASR system is depicted on Figure 5.1 (see any ASR textbook for further references, e.g. Rabiner and Juang (1993)). Each HMM model represents a single speech unit. The speech units are usually context sensitive phones for larger vocabulary tasks (e.g. dictation), and whole words for small vocabulary, command–and–control type tasks. The architecture of the HMMs (number of states, topology, the parametric form of the state p.d.f.s) is decided apriori. The free parameters are the transition probabilities

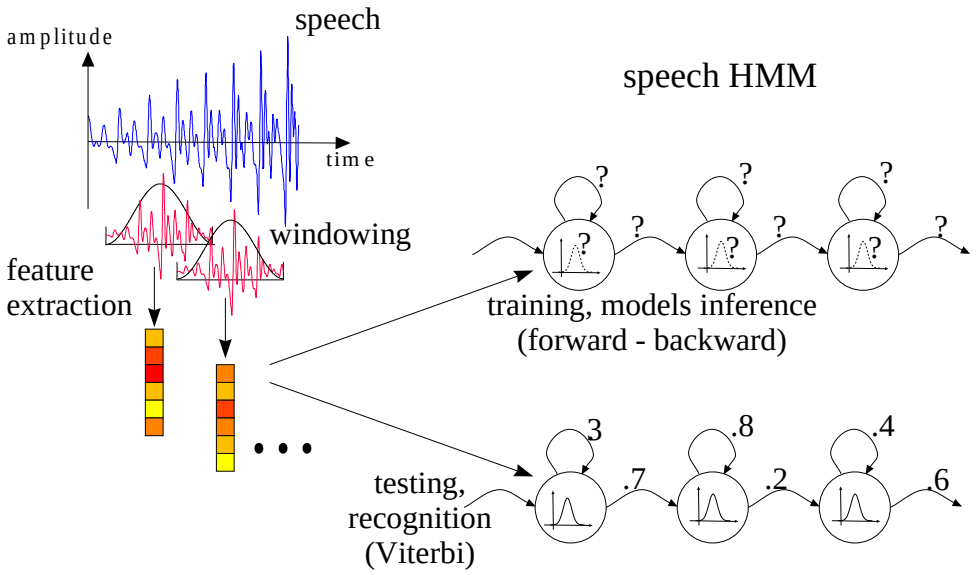


Figure 5.1: Scheme of operation of a typical HMM based ASR system

and the parameters of the state p.d.f.s. In addition, when HMMs are subword units, a *phonetic dictionary* maps strings of subword units into words. In all but the simplest tasks there exists a *grammar* – a set of rules about how the words are put together into sentences.

There are two modes of operation: *training* (inference of the unknown system parameters) and *testing* (recognition). In both cases the incoming speech is divided into overlapping *frames*. Each frame is *windowed* (multiplied with a windowing function) and transformed to yield a *feature vector*. Then during:

Training time: The transcription of the utterance hence the corresponding HMM sequence is known in advance (but the boundaries between the HMMs need not be known). The HMMs of all units in the utterance are concatenated into a *composite HMM*. An efficient recursive *forward–backward* algorithm iteratively updates the free parameters of the composite HMM to increase[1] the likelihood of the training data feature vectors. The algorithm belongs to the Expectation–Maximisation (EM) class of algorithms (Dempster et al. (1977) – see Section 3.4.1). The hidden variable is the state sequence that the source went through while producing the observations.

Testing time: For an isolated words ASR with whole word HMMs, the likelihood of each HMM producing the data can be computed. This is known as the *forward probability*. The HMM with the highest score is the recognised word. However, this mode of operation is hard to extend to connected word recognition. So instead of taking into account all possible paths through the HMM in order to compute the likelihood that the data was produced by the HMM, only the path with the highest likelihood is computed. This is known as a *Viterbi* search. In practise, the likelihood of the best path (called *Viterbi path*) is so much greater then the likelihoods of all other paths, that picking that score alone instead of adding all scores makes little difference. The advantage of the Viterbi search is that it extends naturally to efficiently handle the connected word recognition task.

It is possible to use the Viterbi search during the training, too. In each iteration the data is *aligned* to the states on the Viterbi path. All the data deemed to be generated by one state is

[1] more precise, not to decrease

pooled together and the parameters of the state p.d.f. are updated to increase the likelihood of that data. The transition probabilities can be updated by merely counting the number of times the particular states transition appears. The process is repeated iteratively, obtaining increasingly better parameters and better alignment in each iteration. This is known as *Viterbi training*. It is used infrequently because the forward–backward training is efficient enough. Viterbi training can be used to quickly retrain the models when only the feature extraction module of the ASR system changes, or to adapt the system on–line during recognition.[2]

5.3 The missing data model for robust speech recognition

The missing data (MD) model for robust speech recognition assumes that:

- local patches in some time–frequency representation of the speech spectrum remain mostly unaffected by the other sounds even at poor global SNRs
- they can be identified with a certain probability
- there is sufficient quantity of information there for recognition of the partial speech

So, instead of the usual single source ASR Viterbi decoding:

$$\begin{aligned} W^* &= \underset{W}{argmax} P(W|O) = \underset{W}{argmax} P(O|W)P(W) \\ &= \underset{W}{argmax} \sum_{all\ Q} P(O|Q,W)P(Q|W)P(W) \\ &\approx \underset{W}{argmax}\ \underset{Q}{argmax} P(O|Q,W)P(Q|W)P(W) \\ &= \underset{W}{argmax} P(O|Q^*,W)P(Q^*|W)P(W) \end{aligned} \tag{5.1}$$

where O is a sequence of observations (data vectors), W is the hypothesised word, W^* is the most likely W, Q is a path through the HMM model and Q^* is the most likely Q, the search in the MD model is:

$$\begin{aligned} W^* &= \underset{W}{argmax} P(W|O) = \underset{W}{argmax} P(O|W)P(W) \\ &= \underset{W}{argmax} \sum_{all\ Q} \sum_{all\ M} P(O|Q,M,W)P(M|Q,W)P(Q|W)P(W) \\ &\approx \underset{W}{argmax}\ \underset{Q}{argmax} \sum_{all\ M} P(O|M,Q,W)P(M|Q,W)P(Q|W)P(W) \\ &= \underset{W}{argmax} \sum_{all\ M} P(O|M,Q^*,W)P(M|Q^*,W)P(Q^*|W)P(W) \end{aligned} \tag{5.2}$$

where M is a *mask* determining which features were generated by the source that is decoded. A simple example of a mask would be a binary matrix. The MD model makes provision for multiple sources via the mask. The mask captures the prior information about:

- how reliable the features are
- which combinations of features tend to "stick together" above the noise

[2]The relation between Viterbi and Baum-Welch training is similar to the one between the K-means and EM algorithms for fitting mixtures of Gaussian to data. The former algorithms assume hard decision, i.e. the data was either generated or not generated by the state/mixture. The latter allow for every point to have been generated by every state/mixture with a certain probability.

Features deemed to have been generated by the (speech) source of interest will be refereed to as "present" and indicated accordingly by the binary mask. Features deemed to have been generated by other source(s) which are not of direct interest for the ASR system will be refereed to as "missing".

An alternative form for Eq. (5.2) is:

$$\begin{aligned} W^* &= \underset{W}{argmax} P(W|O) = \underset{W}{argmax} P(O|W)P(W) \\ &= \underset{W}{argmax} \sum_{all\ Q} \sum_{all\ M} P(O|Q,M,W)P(Q|M,W)P(M|W)P(W) \\ &\approx \underset{W}{argmax} \sum_{all\ M} P(O|M,Q^*,W)P(Q^*|W)P(M|W)P(W) \end{aligned} \tag{5.3}$$

where it is assumed that the path Q is independent of the mask M (hence $P(Q|M,W) = P(Q|W)$). In this form the mask M is conditioned on the word W, but not on the state path Q.

The factors in the expression above represent:

$P(W)$	the probability of the word regardless of the acoustic observations (comes from the language model)
$P(Q^*\|W)$	the probability of the most likely path Q^*
$P(M\|Q^*,W)$	(or $P(M\|W)$) the probability of the mask (determining which features were generated by the state or the word)
$P(O\|M,Q^*,W)$	the probability of the partial observations

Figure 5.2 shows a mask example. In the top row, the left panel shows a clean speech TIdigits (Leonard, 1984) digits utterance ("1159"). It is mixed with NOISEX factory noise (Varga et al., 1992) shown in the middle panel. The right panel depicts the noisy speech at global SNR of 0dB. The representation is a smoothed 64–channel auditory filter bank (centre frequencies spaced linearly in ERB-rate from 50 to 8000Hz), computed every 10ms (Cooke, 1991).

The panels in the middle depict three different masks. The red areas in the masks indicate presence of speech in the noisy signal, while the blue regions indicate absence of speech. The masks are derived assuming additivity of the speech and the noise in the power spectral domain. The left one is derived by comparing the clean and the noisy speech, and selecting as speech points those with local SNR of 7dB and more. The middle one is derived from a local SNR estimate, based on a stationary noise estimate from the noisy speech (with the same 7dB threshold criterion). The right one is derived by comparing the speech estimate used for spectral subtraction (SS) and noisy speech and treating the points where speech estimate is smaller then the noisy speech as non–speech. The panels in the bottom row show the noisy speech as "seen through" the corresponding masks. They illustrate that even at SNR of 0dB, with on average equally loud speech and noise, large parts of speech spectrum are unaffected.

Separation with the MD model

The MD model for robust ASR integrates the separation of the speech and the noise with the recognition of the speech. Finding the mask amounts to separation of the speech and the noise. The ML mask M^* is:

$$\begin{aligned} M^* &= \underset{M}{argmax} P(M|O) = \underset{M}{argmax} P(O,M) \\ &= \underset{M}{argmax} \sum_{all\ W} \sum_{all\ Q} P(O|M,Q,W)P(M|Q,W)P(Q|W)P(W) \\ &\approx \underset{M}{argmax} P(O|M,Q^*,W^*)P(M|Q^*,W^*)P(Q^*|W^*)P(W^*) \end{aligned} \tag{5.4}$$

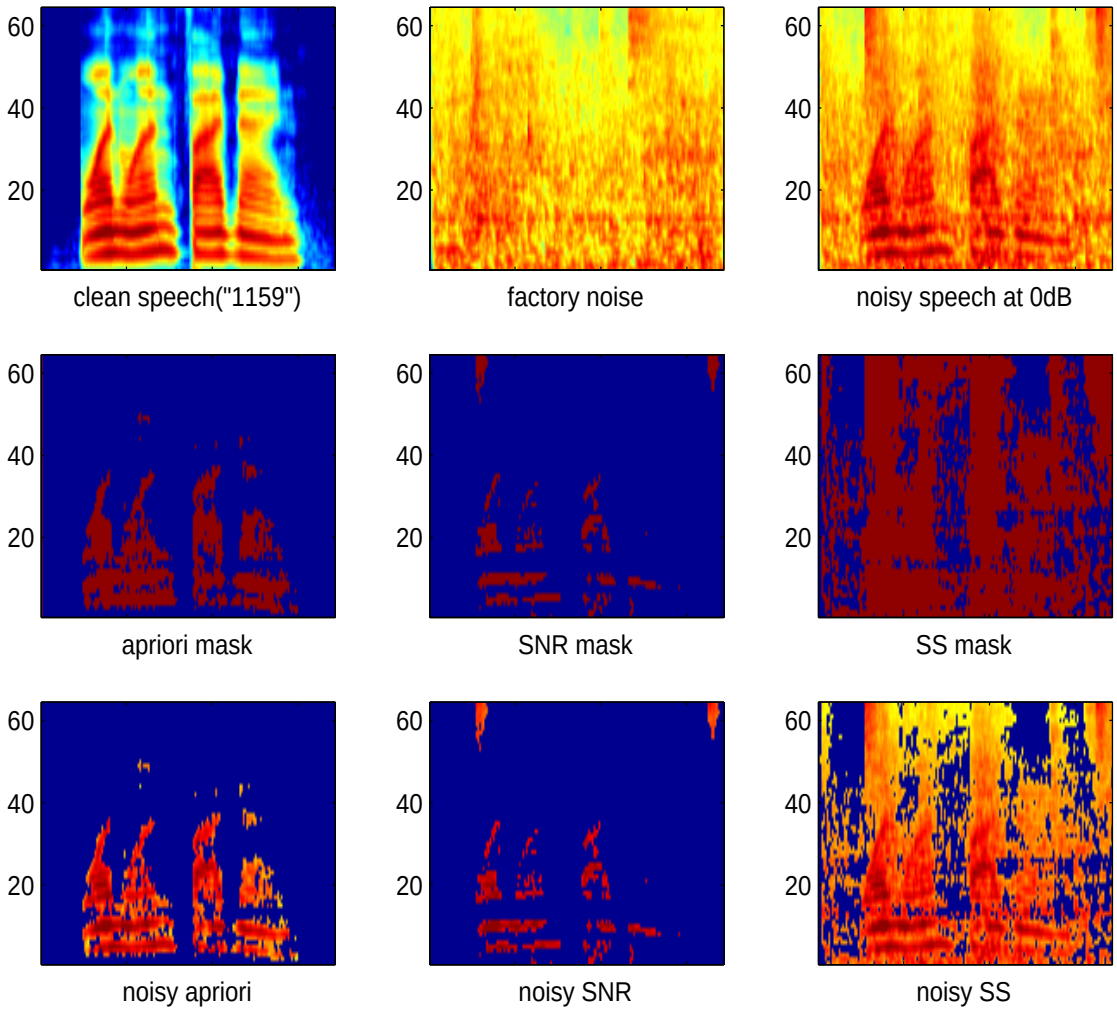


Figure 5.2: Top row: clean speech (spoken digits "1159") on the left, factory noise (middle) and noisy speech created by mixing the clean speech and the noise at 0dB global SNR. Middle row: "oracle" mask (left), estimated mask (middle) and the mask indicating the regions where speech estimation failed (right). Bottom row: noisy speech as "seen through" the corresponding masks.

or:

$$M^* \approx \underset{M}{argmax} P(O|M, Q^*, W^*) P(Q^*|W^*) P(M|W^*) P(W^*) \tag{5.5}$$

The problem of separation is reduced to Viterbi search (dynamic programming), too.

The integration of the separation and the recognition with the MD model is possible for a simple model of the acoustic environment. The combination of speech and noise in a all–or–nothing manner (each feature is generated by the speech or the noise exclusively) is a significant simplification. It may not be a viable one for a complicated environment with convolutional noise and/or reverberation. But for additive noise it is largely appropriate for several reasons already discussed in Section 3.2. Experimental comparisons with techniques like PMC do not indicate performance loss due to this simplification (Renevey, 2000, pp. 149). Further:

- The small difference in performance of multiconditional models on seen and unseen noises (Hirsch and Pearce, 2000) may suggest that there is insufficient structure in the noises to infer strong

noise models/constraints. This supports the argument that the rich speech structure should be the primary source of constraints needed for separation.

- The speech and the noise are independent sources (i.e. the mutual information is zero). The only way in which the noise source constrains the speech source is through the environmental function that combines the observations generated from both sources into a noisy observation. The MD model can capture most of this information via the usage of *counterevidence* (Section 5.5.7), if the variance of the noise is reasonably large, as is the case for most of the noises.

Chapter 7 contains more detailed discussion about the relative merits on the noise models and counterevidence in Section 7.2.4.

5.4 Modelling the mask

The introduction of a mask decomposes the problem of robust ASR[3] into two subproblems: separation and recognition. Contrary to some other models (e.g. HMM decomposition), where decoding one source implies decoding all sources, the mask allows for decoding of one source only while disregarding the other sources present in the auditory scene.

At present it is unclear which is the most appropriate form for the mask model.

While experimenting, it is useful to use *oracle* masks (Figure (5.2), left panel of the middle row). They can be computed by comparing the clean speech and noise[4]. This is helpful while evaluating different strategies for computing the likelihood of the partial data $P(O|M, Q, W)$, since the influence of imperfect separation is minimised. It can also provide an idea about the upper limits on the performance that can be achieved. The oracle mask M^* has probability of 1 and is independent of Q^* and W.

For "production" ASR, local SNR estimation based on noise estimation (Figure (5.2), middle panel of the middle row) was used to compute the mask model in the experiments (see Chapter 6). Ideally, CASA would be the method of choice for building this model. Barker et al. (2001b) have successfully merged the former with elements of the latter (harmonicity based masks) in their system. Seltzer et al. (2000) used a static classifier for mask estimation with good results, roughly half way between the noise estimation results and results with "oracle" masks. Roweis (2000) used speaker dependent HMMs to learn speaker dependent mask models solely for the purpose of separation and (successful) reconstruction.

Faced with integrating several sources of evidence, the most plausible route seems a statistical model. Chapter 7 discusses further the issues pertinent to building a mask model and the properties that would be desirable for the model to have.

5.4.1 Computing the sum over all possible masks

Approximation

The summation for *all* M in Eqs. (5.2) and (5.3) is over all possible masks. In general, the number of possible masks per frame is prohibitively large: two to the power of number of features. However, under assumptions similar to ones leading to the Viterbi approximation in Eq. (5.1):

(a) the most likely (ML)[5] mask M^* will be much more likely than any other mask M, i.e. $P(M^*|Q^*, W) \gg P(M|Q^*, W)$ for $M \neq M^*$

(b) the likelihoods of the data that the corresponding masks will give rise to ($P(O|M^*, Q^*, W)$ and $P(O|M, Q^*, W)$) will behave similarly, i.e. $P(O|M^*, Q^*, W) \gg P(O|M, Q^*, W)$ for $M \neq M^*$

[3] when viewed as decoding one source while listening to several

[4] or clean and noisy speech – but then a model of acoustic environment is needed to derive the SNR

[5] ML will be used both for "most likely" and "maximum likelihood"

The summation over all possible masks (weighted by their probability) may be approximated by selection of the most probable mask M^*:

$$W^* \approx \underset{W}{argmax} P(O|M^*, Q^*, W)P(M^*|Q^*, W)P(Q^*|W)P(W) \tag{5.6}$$

or its analog for Eq. (5.3):

$$W^* \approx \underset{W}{argmax} P(O|M^*, Q^*, W)P(Q^*|W)P(M^*|W)P(W) \tag{5.7}$$

This means that is is sufficient to select the HMM model/"word" W^* whose most likely mask M^* gives rise to the biggest of the partial data likelihoods on the most likely path Q^*.

The motive for approximation of the sum over all masks M with selection of the most probable mask M^* is the same as to the original Viterbi approximation[6] (Eq. (5.1)). The likelihood of the features not generated by this (or any other HMM model available) is small and may be neglected in the sum.

Exact calculation in a special case

In the special case when both the state and the mask p.d.f.s are sums of factorisable distributions, and when the *independence assumptions* are taken into account (see Section 5.5), an efficient computation of the sum over all possible masks in Eq. (5.2) is possible (Appendix D, Eq. (D.5)):

$$\begin{aligned}
W^* &= \underset{W}{argmax}\left\{\prod_{t=1}^{T} p(\mathbf{o}(t)|q^*(t), W)\right\} \cdot P(Q^*|W)P(W) \\
&= \underset{W}{argmax}\left\{\prod_{t=1}^{T} \sum_{all\,\mathbf{m}} p(\mathbf{o}(t), \mathbf{m}(t)|q^*(t), W)\right\} \cdot P(Q^*|W)P(W) \\
&= \underset{W}{argmax}\left\{\prod_{t=1}^{T} \sum_{k} P(k) \prod_{i} [p_i(o_i(t)|k, m_i(t) = 0, q^*(t), W)p(m_i(t) = 0|q^*(t), W) \right. \\
&\left. + p_i(o_i(t)|k, m_i(t) = 1, q^*(t), W)p(m_i(t) = 0|q^*(t), W)]\right\} \cdot P(Q^*|W)P(W)
\end{aligned} \tag{5.8}$$

The result is intuitive: the contribution of the present and missing features to the likelihood should be weighted by the probability of them being present or missing. A discrete mask is a special case where the probabilities of a feature present/missing is either 0 or 1. The missing features (observations generated by the non–speech source) can still contribute to the likelihood of the speech source model – their contribution can be exploited as *counterevidence* (Section 5.5.7).

5.5 Computing the likelihood of the partial observations

Assuming that $O = \{\mathbf{o}(t)\}_{t=1\ldots T}$, $Q = \{q(t)\}_{t=1\ldots T}$ and $M = \{\mathbf{m}(t)\}_{t=1\ldots T}$ where t is the time frame, T is the total number of frames, $\mathbf{o}(t)$ is the observation vector, $q(t)$ is the state the speech source was in at time t and $\mathbf{m}(t)$ is the framewise mask for frame t, the mask $\mathbf{m}(t)$ divides the feature vector $\mathbf{x}$ into a *present* part $\mathbf{x}_p$ and a *missing* part $\mathbf{x}_m$ at time t.[7] This means that the observation $\mathbf{o}(t)$ was only partly generated by the speech source in state $q(t)$. The feature subvector $\mathbf{o}_p(t)$ of the vector $\mathbf{o}(t)$ was generated by the speech source that is being decoded (and a model is certainly available). The remaining subvector $\mathbf{o}_m(t)$ was generated by some other, possibly noise, source. A model of this source might, or might not, be available.

Taking into account the *independence assumptions*:

[6] a "folk conjecture" in the ASR community is that transition probabilities are unimportant in the overall likelihood compared to the emission probabilities; however, it is the "ordering" of the HMM states enforced by fixing most of the transition probabilities to zero that yields advantage over a mere mixture model.

[7] $\mathbf{o}(t)$ is the realisation of the random variable $\mathbf{x}$ at time t

- the observations are independent, i.e. $\mathbf{o}(t)$ is independent of $\mathbf{o}(t-1)$
- each observation $\mathbf{o}(t)$ is dependent only on the state $q(t)$

The likelihood of the partial data $P(O|M,Q,W)$ from Eq. (5.2)becomes:

$$P(O|M,Q,W) = \prod_{t=1}^{T} P(\mathbf{o}(t)|\mathbf{m}(t),q(t),W) \tag{5.9}$$

The conditioning on the mask $\mathbf{m}(t)$ determines which features of the feature vector $\mathbf{o}(t)$ were produced by the speech source when in state $q(t)$. It divides the observation random variable $\mathbf{x}$ into a present part $\mathbf{x}_p$ and a missing part $\mathbf{x}_m$:

$$\mathbf{x} = (\mathbf{x}_p, \mathbf{x}_m) \tag{5.10}$$

As mentioned in Chapter 3, knowing the probability distribution of the whole observation, there are two possible ways to compute the probability of the partial observation: *marginalisation* and *imputation*. In the case of marginalisation the probability distribution itself is adapted to yield the distribution of the partial data. In the case of imputation, the conditional distribution of the unobserved given the observed data is computed from the joint distribution, and a "suitable" point on the curve is picked as a plug–in replacement of the unobserved data. Then the joint data distribution can be used to asses the probability "full" data vector. These strategies are discussed next in the context of an HMM based ASR system.

5.5.1 Marginalisation in an HMM based MD ASR system

Dropping the conditioning on W and time index t where not necessary and replacing the probability distribution P with probability density distribution p to suit a continuous random variable $\mathbf{x}$, the factor $P(\mathbf{o}|\mathbf{m},q)$ from Eq. (5.9) becomes (Ahmad and Tresp, 1993):

$$p(\mathbf{o}|\mathbf{m},q) = p(\mathbf{o}_p|q) = \int p(\mathbf{o}_p,\mathbf{x}_m|q)d\mathbf{x}_m \tag{5.11}$$

where $\mathbf{x}_m$ is a subvector of $\mathbf{x}$ completely determined by the mask $\mathbf{m}$. Depending on $\mathbf{m}$, different sets of features need to be marginalised in each frame.

The state p.d.f. in a typical HMM system is a mixture of multivariate *diagonal* Gaussians:

$$p(\mathbf{x}|q) = \sum_{k=1}^{K} P(k|q)p(\mathbf{x}|k,q) = \sum_{k=1}^{K} P(k|q)\prod_{i=1}^{N} p(x_i|k,q) \tag{5.12}$$

where $p(x_i|k,q)$ is a univariate Gaussian:

$$p(x_i|k,q) = \frac{1}{\sqrt{2\pi\sigma_{i,k}^2}}\ \exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_{i,k}}{\sigma_{i,k}}\right)^2\right\} \tag{5.13}$$

The assumption that the state p.d.f. can be modelled by a Gaussian mixture with diagonal covariance matrices is crucial for practical implementation of a MD ASR system. Both for MD and non–MD ASR system the number of free parameters (and consequently the amount of data needed for their estimation) is significantly smaller. Further, sufficient number of components in the mixture can approximate not only rotation of the distribution (that full covariance matrix can model as well), but also non–Gaussian distributions and/or multimodal distributions (that full covariance matrix can not model). In addition, while still

$$p(\mathbf{x}_p,\mathbf{x}_m|q) \neq p(\mathbf{x}_p|q)p(\mathbf{x}_m|q), \tag{5.14}$$

"inside" each mixture component

$$p(\mathbf{x}_p, \mathbf{x}_m|k, q) = p(\mathbf{x}_p|k, q)p(\mathbf{x}_m|k, q), \tag{5.15}$$

since the independence (between any two features) makes

$$p(\mathbf{x}_m|\mathbf{x}_p, k, q) = p(\mathbf{x}_m|k, q). \tag{5.16}$$

This is important for efficient computation of the marginal state p.d.f., which has to be performed in every frame for all states in the MD HMM system. The marginal of the general multivariate normal distribution:

$$\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\} \tag{5.17}$$

is again multivariate Normal $\mathcal{N}(\mathbf{x}_p; \mu_p, \Sigma_{pp})$. Its parameters μ_p and Σ_{pp} are readily available from the parameters of the joint distribution:

$$\mu = \begin{bmatrix}\mu_p \\ \mu_m\end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix}\Sigma_{pp} & \Sigma_{pm} \\ \Sigma_{pm} & \Sigma_{mm}\end{bmatrix} \tag{5.18}$$

However, for efficient computation the inverted covariance of the partial data $(\Sigma_{pp})^{-1}$ is needed (instead of the readily available Σ_{pp}). As it can not be easily derived from Σ^{-1} with a general covariance matrix[8], a matrix inversion per state per frame is needed to adapt the HMM system in every frame. This is very costly.

But, if all pairs of features are mutually independent, implying a diagonal covariance matrix (only the variances are non–zero) the state p.d.f. marginal Eq. (5.11) is:

$$\begin{aligned} p(\mathbf{x}_p|q) &= \int p(\mathbf{x}_p, \mathbf{x}_m|q)d\mathbf{x}_m = \int \left\{\sum_{k=1}^{K} P(k)p(\mathbf{x}_p, \mathbf{x}_m|k, q)\right\} d\mathbf{x}_m \\ &= \int \sum_{k=1}^{K} P(k)p(\mathbf{x}_p|k, q)p(\mathbf{x}_m|k, q)d\mathbf{x}_m = \sum_{k=1}^{K} P(k)p(\mathbf{x}_p|k, q) \underbrace{\int p(\mathbf{x}_m|k, q)d\mathbf{x}_m}_{1} \\ &= \sum_{k=1}^{K} P(k)p(\mathbf{x}_p|k, q) = \sum_{k=1}^{K} P(k) \prod_{i \in present} p(x_i|k, q) \end{aligned} \tag{5.19}$$

This form lends itself to efficient calculation. It simply states that only the contributions (to the likelihood) of the present features need to be taken into account.

5.5.2 Imputation in an HMM based MD ASR system

The second strategy for computing the likelihood of partial observations $p(\mathbf{o}|\mathbf{m}, q)$ from Eq. (5.9) (see Section 5.5) is to "fill in", *impute* the missing observations using the knowledge of the data density, and then continue as if the imputed features were the "true" (but the unobserved) ones.[9] The non–speech observations $\mathbf{o}_m(t)$ are disregarded as they are not generated from the speech source. Instead, estimates $\hat{\mathbf{o}}_m(t)$ are obtained and used further. It seems natural to use some form of conditional data distribution $p(\mathbf{x}_m(t)|\mathbf{x}_p(t))$ to come up with a "sensible" $\hat{\mathbf{o}}_m(t)$ to impute.

In the context of an HMM based recogniser, there are two possible conditional distributions that can be used to draw the imputed values from: the data distribution $p(\mathbf{x}_m|\mathbf{x}_p)$ or the state conditioned data distribution $p(\mathbf{x}_m|\mathbf{x}_p, q)$. In the former case there is a single value which is

[8] For example, $(\Sigma^{-1})_{pp} = (\Sigma_{pp} - \Sigma_{pm}\Sigma_{mm}^{-1}\Sigma_{mp})^{-1}$, $(\Sigma^{-1})_{mm} = (\Sigma_{mm} - \Sigma_{mp}\Sigma_{pp}^{-1}\Sigma_{pm})^{-1}$, etc.

[9] imputation makes sense only with generative models; if the data density is not inferred during the training, then it needs to be inferred separately solely for the purpose of imputation, as is the case with the hybrid, non–HMM based ASR systems, e.g. (Dupont, 1998)

imputed. This case will be labelled Global Data Imputation (GDI). In the latter case as many values can be imputed as there are states resulting in as many different frames. This may seem as unsurmountable difficulty. However, during the Viterbi search, when estimating the probability that a particular frame was generated by a particular state, not all frames need to be assessed. It is enough to compute the likelihood of the frame whose missing values were imputed from that state conditional p.d.f. The rationalisation being that it is already assumed that the speech source was in that particular state. This case will be termed State conditioned Data Imputation (SDI).

5.5.3 Global data imputation

Computing the data distribution $p(\mathbf{x}_m|\mathbf{x}_p)$ in an HMM based system is straightforward, as all state conditioned distributions are available. Hence:

$$\begin{aligned} p(\mathbf{x}_m|\mathbf{x}_p) &= \sum_{all\ q} p(\mathbf{x}_m, q|\mathbf{x}_p) = \sum_{all\ q} p(\mathbf{x}_m|\mathbf{x}_p, q)p(q|\mathbf{x}_p) = \sum_{all\ q} p(\mathbf{x}_m|\mathbf{x}_p, q)\frac{p(\mathbf{x}_p|q)P(q)}{p(\mathbf{x}_p)} \\ &= \sum_{all\ q} p(\mathbf{x}_m|\mathbf{x}_p, q)\frac{p(\mathbf{x}_p|q)P(q)}{\sum\limits_{all\ q'} p(\mathbf{x}_p|q')P(q')} = \frac{\sum\limits_{all\ q} p(\mathbf{x}_p|q)P(q)p(\mathbf{x}_m|\mathbf{x}_p, q)}{\sum\limits_{all\ q} p(\mathbf{x}_p|q)P(q)} \end{aligned} \tag{5.20}$$

The form simply states that the conditional p.d.f. $p(\mathbf{x}_m|\mathbf{x}_p)$ is a sum of state conditioned conditional p.d.f.s $p(\mathbf{x}_m|\mathbf{x}_p, q)$, weighted by a factor $\frac{p(\mathbf{x}_p|q)P(q)}{\sum_{all\ q'} p(\mathbf{x}_p|q')P(q')}$.

5.5.4 "Probability of a state"

The prior probability of the state q, $P(q)$, can either be computed exactly, by the recursive:[10]

$$P(q) = \sum_{t=1}^{T} P(q(t) = q) = \sum_{t=1}^{T} \sum_{all q'} P(q(t-1) = q')P(q(t) = q|q(t-1) = q') \tag{5.21}$$

or approximately, as the relative frequency with which the state q appears in the state aligned training data. In the case of a *straight–through* HMM (HMM with no skip states, which is commonly used for speech recognition) the latter can be derived from the transition probabilities[11] as:

$$P(q) = \frac{1/[1 - s(q)]}{\sum_{all\ q'} 1/[1 - s(q')]} \tag{5.22}$$

where $s(q)$ is the probability that the speech source stays in state q once it is in it $P(q(t) = q|q(t-1) = q)$, the sum $\sum_{all\ q'}$ is over all states (across HMM models) and it is assumed that there is no grammar (every word/HMM sequence is equally likely).

In almost all continuous–density (CD) HMM based ASR systems the state p.d.f.s $p(\mathbf{x}|q)$ are mixtures of diagonal Gaussians (Eq. (5.12)). Considering the somewhat more general case of a

[10] abusing the notation for consistency – q stands for a particular state (a realisation of a random variable) while $q(t)$ stands for the state the source is in time t (a random variable)

[11] The transition probabilities may be derived themselves from the state aligned training data. Absence of skip state makes the reverse computation possible: the expected number of frames $N(q)$ generated from state q is $\mathcal{E}\{N(q)\} = 1 + \sum_{n=0}^{\infty} n[s(q)]^n[1 - s(q)]$, where the 1 comes from the straight through HMM (the model must pass through each state at least once) and the infinite sum is the expected number of frames the source remains in state q once it is in it. The infinite sum evaluates to $s(q)/[1 - s(q)]$, giving rise to $\mathcal{E}\{N(q)\} = 1/[1 - s(q)]$ and leading to Eq. (5.22) which uses the expected number of frames $P(q) = \mathcal{E}\{N(q)\}/\sum_{q'} \mathcal{E}\{N(q')\}$ instead of relative frequency

mixture of factorisable distributions, we get:

$$\begin{aligned} p(\mathbf{x}_m|\mathbf{x}_p) &= \frac{\sum\limits_{all\ q} p(\mathbf{x}_p|q)P(q)p(\mathbf{x}_m|\mathbf{x}_p,q)}{\sum\limits_{all\ q} p(\mathbf{x}_p|q)P(q)} = \frac{\sum\limits_{all\ q} P(q)\sum_k P(k|q)p(\mathbf{x}_p|k,q)p(\mathbf{x}_m|\mathbf{x}_p,k,q)}{\sum\limits_{all\ q} P(q)\sum_k P(k|q)p(\mathbf{x}_p|k,q)} \\ &= \frac{\sum\limits_{all\ q} P(q)\sum\limits_k P(k|q)p(\mathbf{x}_p|k,q)p(\mathbf{x}_m|k,q)}{\sum\limits_{all\ q} P(q)\sum\limits_k P(k|q)p(\mathbf{x}_p|k,q)} \end{aligned} \tag{5.23}$$

where the independence assumption that inside the mixture $p(\mathbf{x}_m,\mathbf{x}_p|k,q) = p(\mathbf{x}_m|k,q)p(\mathbf{x}_p|k,q)$ (or $p(\mathbf{x}_m|\mathbf{x}_p,k,q) = p(\mathbf{x}_m|k,q)$) has been used.

The conditional distribution is a weighted and normalised (weights sum to unity) sum of the individual factors $p(\mathbf{x}_m|k,q)$. The weights depend on the prior probability of the mixture $P(k|q)$ (state dependent) in addition to the prior state probability $P(q)$ and the probability of the present data $p(\mathbf{x}_p|k,q)$.

It is not immediately clear how to choose a point from this manifold. The "natural" criterion maybe to choose the global maximum, as it is the most likely point. However, it can not be easily determined. Another point of choice may be the conditional mean. It has an appealing property that it minimises the quadratic error, and is easily computable:

$$\begin{aligned} \mathcal{E}_{\mathbf{x}_m|\mathbf{x}_p}\{\mathbf{x}_m\} &= \int p(\mathbf{x}_m|\mathbf{x}_p)\mathbf{x}_m d\mathbf{x}_m = \frac{\sum\limits_{all\ q} P(q)\sum\limits_k P(k|q)p(\mathbf{x}_p|k,q)\overbrace{\int p(\mathbf{x}_m|k,q)\mathbf{x}_m d\mathbf{x}_m}^{\mu_{m,k,q}}}{\sum\limits_{all\ q} P(q)\sum\limits_k P(k|q)p(\mathbf{x}_p|k,q)} \\ &= \frac{\sum\limits_{all\ q} P(q)\sum\limits_k P(k|q)p(\mathbf{x}_p|k,q)\mu_{m,k,q}}{\sum\limits_{all\ q} P(q)\sum\limits_k P(k|q)p(\mathbf{x}_p|k,q)} \end{aligned} \tag{5.24}$$

The form is a weighed sum of the means of the missing components. The drawback is that if the Gaussians are far apart and the distribution is multimodal, the mean may fall in a region of very low probability. If, however, most of the Gaussians are "stacked" close together for the purpose of approximating a non–Gaussian distribution which has few modes, imputing the mean maybe appropriate.

The "imputed" missing observations $\hat{\mathbf{o}}_m = \mathcal{E}_{\mathbf{x}_m|\mathbf{o}_p}\{\mathbf{x}_m\}$ can be used instead of the noisy ones $\mathbf{o}_m$ and subsequent recognition can continue with the "complete" data vector $\hat{\mathbf{o}} = (\mathbf{o}_p, \hat{\mathbf{o}}_m)$ instead of $\mathbf{o}$.

5.5.5 State dependent data imputation

When computing the emission probability for a state q it is assumed that the source is in that state. Hence the state conditioned p.d.f. $p(\mathbf{x}|q)$ is used for computing the probability that the observation was generated by that state. Analogous, if some of the observations are missing, and it is assumed that the source us in state q, it maybe preferable to use the state conditioned conditional data density $p(\mathbf{x}_m|\mathbf{x}_p,q)$ instead of the conditional data density $p(\mathbf{x}_m|\mathbf{x}_p)$, for the purpose of imputation. The form is similar but simpler then Eq. (5.20):

$$p(\mathbf{x}_m|\mathbf{x}_p,q) = \frac{p(\mathbf{x}_m,\mathbf{x}_p|q)}{p(\mathbf{x}_p|q)} = \frac{p(\mathbf{x}_m,\mathbf{x}_p|q)}{\int p(\mathbf{x}_m,\mathbf{x}_p|q)d\mathbf{x}_m} \tag{5.25}$$

Again taking into account that usually in a HMM based ASR system the state p.d.f.s $p(\mathbf{x}|q)$ are mixtures of diagonal Gaussians (Eq. (5.12)), and considering the somewhat more general case of

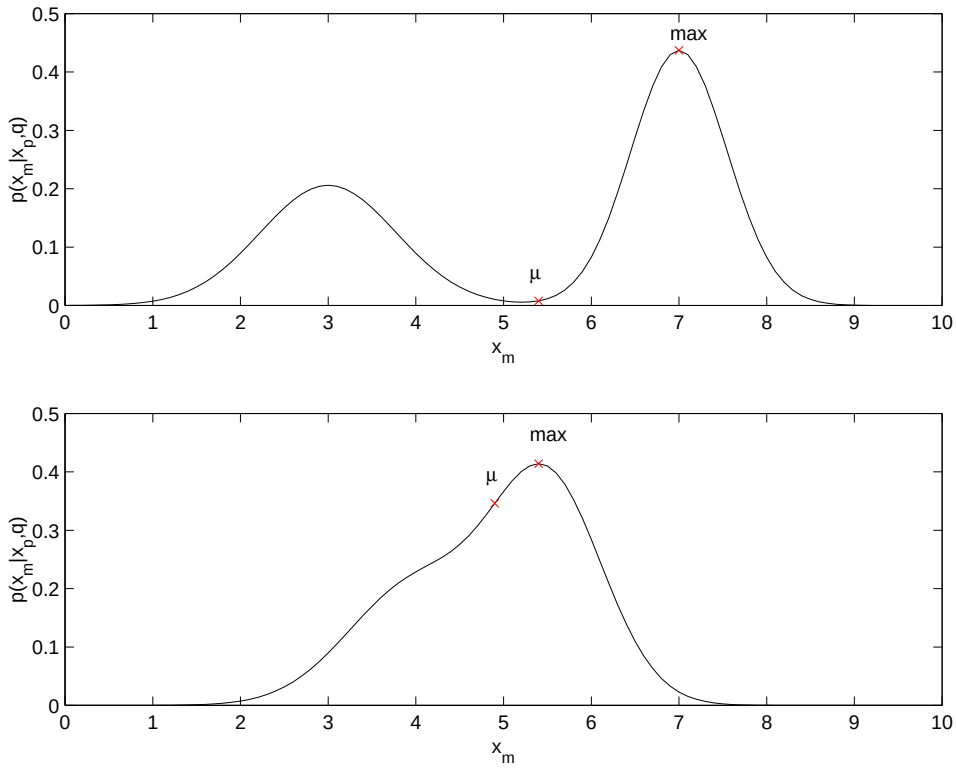


Figure 5.3: Picking a point appropriate for imputation from the conditional p.d.f. can be tricky. The upper panel depicts a conditional p.d.f. $p(x_m|x_p, q) = 0.4\mathcal{N}(x_m; 3, 0.6) + 0.6\mathcal{N}(x_m; 7, 0.3)$ (i.e. $p(k = 1|o_p, q) = 0.4$, $p(k = 2|o_p, q) = 0.6$). The lower panel depicts a conditional p.d.f. $p(x_m|x_p, q) = 0.4\mathcal{N}(x_m; 4, 0.6) + 0.6\mathcal{N}(x_m; 5.5, 0.4)$. The "$\mu$" and "max" symbols note the corresponding (conditional) mean and the global maximum.

mixture of a factorisable distributions, the previous expression becomes:

$$p(\mathbf{x}_m|\mathbf{x}_p, q) = \frac{\sum_k P(k|q)p(\mathbf{x}_p|k, q)p(\mathbf{x}_m|k, q)}{p(\mathbf{x}_p|q)} = \sum_k p(k|\mathbf{x}_p, q)p(\mathbf{x}_m|k, q) \tag{5.26}$$

where

$$p(k|\mathbf{x}_p, q) = \frac{p(\mathbf{x}_p|k, q)}{p(\mathbf{x}_p|q)} = \frac{P(k|q)p(\mathbf{x}_p|k, q)}{\sum_{k'} P(k'|q)p(\mathbf{x}_p|k', q)} \tag{5.27}$$

is the *responsibility* of the k-th mixture for the present data $\mathbf{x}_p$.

Choosing a criterion for picking a single point from this function for imputation is again not obvious. The highest mode (the global maximum) maybe most desirable (as the most likely value), but it is not easily computable (Carreira-Perpiñán, 1999). Another choice is minimising the quadratic error which implies using the conditional mean:

$$\mathcal{E}_{\mathbf{x}_m|\mathbf{x}_p,q}\{\mathbf{x}_m\} = \int p(\mathbf{x}_m|\mathbf{x}_p, q)\mathbf{x}_m d\mathbf{x}_m = \sum_k p(k|\mathbf{x}_p, q) \underbrace{\int p(\mathbf{x}_m|k, q)\mathbf{x}_m d\mathbf{x}_m}_{\mu_{m,k,q}} \tag{5.28}$$

Again, if the Gaussians in the mixture are far apart and the state p.d.f. is multimodal, the mean may fall in the region of very low probability, as shown on the upper panel of Fig. (5.3). However,

if the Gaussians approximate possibly a non–Gaussian distribution which is not too multimodal, the mean may be a good choice (as illustrated on the lower panel of Fig. (5.3)).

In both cases the global maximum is close to the means of the individual Gaussians. In the upper panel of Fig. (5.3) (components far apart), it almost coincides with the highest mean in the mixture. It in the lower panel of Fig. (5.3) (components stacked together), it is a bit further from the highest mean. In any case, the means of the individual Gaussians may be a good starting point for the search for the global maximum (Carreira-Perpiñán, 1999). But still, the global maximum does not coincide with the "true value" all the time (just most of the time), so the reconstruction is not from perfect.

After the $\hat{\mathbf{o}}_{m,q}$ is computed, the emission probability of each state can be computed as $p(\hat{\mathbf{o}}_{m,q}, \mathbf{o}_p|q)$ and the decoding can continue.

5.5.6 Marginalisation or imputation?

It is of interest to consider under which conditions one of the techniques is more preferable to the other.

Marginalisation is computationally cheaper and there are no problems like the choice of criterion for picking a point on the conditional p.d.f. in the imputation. In our experiments the ASR accuracy was always better with marginalisation then with imputation.

Imputation provides reconstruction of the unobserved data in addition to recognition. The reconstruction task is not trivial. Many imputed values will give rise to the "correct likelihood", although only one is "correct". While marginalisation effectively averages the likelihood over all of the possible imputations (see Eq. (5.29)below), the imputation process has to pick only a single value to be imputed. The fact that many other are also possible (if less probable) is disregarded. Still, some non–HMM ASR systems may require complete(d) feature vectors because they can not be adapted. Sometimes the adaptation is not trivial (as is the case with the hybrid ones), or the recogniser is entirely separate subsystem treated like a "black box". In that case imputation can be the method of choice.

It is also possible to use both techniques in a complementary manner, if both recognition and reconstruction are required: marginalisation can be used to obtain the most likely state sequence. Once the sequence is known, the conditional state p.d.f.s may be used for imputation.

It seems that both techniques are somewhat dependent on the amount of data the mask "lets in". In the ASR experiments it was notable that for marginalisation it is better to impose a more stringent assessment about the reliability of the features. While for imputation it was better to loosen the criterion and treat more of the features as more reliable (compared to marginalisation). It seems it is hard to impute sensibly the majority of the features in the vector if only small minority of them are present. In the most extreme cases when where only couple of features are present, it may be preferable to delete the whole frame altogether (*frame deletion*), rather then trying to salvage it. The exact relationship between the data "quality" and "quantity" in the context of robust ASR is not clear at present.

As mentioned in Section 3.3.2, there is an intuitive connection between the marginalisation of the state conditioned p.d.f. $p(\mathbf{x}_p, \mathbf{x}_m|q)$ and the imputation from the state conditioned conditional distribution $p(\mathbf{x}_m|\mathbf{x}_p, q)$. The marginal $p(\mathbf{x}_p|q)$ can be considered an average over all possible conditional imputations $p(\mathbf{x}_p|\mathbf{x}_m, q)$ weighted by their respective probabilities of occurrence $p(\mathbf{x}_m|q)$:

$$p(\mathbf{x}_p|q) = \int p(\mathbf{x}_p|\mathbf{x}_m, q)p(\mathbf{x}_m|q)d\mathbf{x}_m \tag{5.29}$$

5.5.7 Counterevidence

Although the missing observations $\mathbf{o}_m(t)$ were not generated by the speech source that is decoded, they can still be used to derive information about which states are *unlikely* to have generated the observations. This is an important constraint when no noise model is available. The commonly used models of speech and noise mixing:

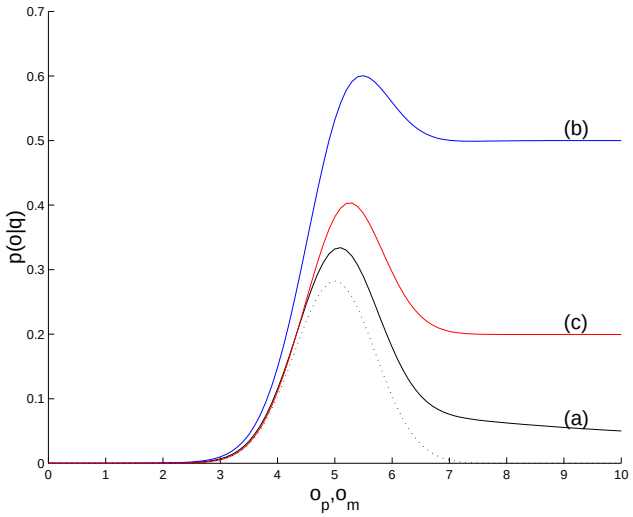


Figure 5.4: Plot of $p(o|q) = \sum_m p(o|m,q)p(m|q)$ with several possible measures of counterevidence $p(o_m|q)$: (a) "average likelihood" $p(o_m|q) = 1/o_m \int_0^{o_m} p(x_m|q)dx_m$ (b) "probability" $p(o_m|q) = \int_0^{o_m} p(x_m|q)dx_m$ (c) "self–weighted likelihood" $p(o_m|q) = \int_0^{o_m} [p(x_m|q)]^2 dx_m$; the dashed line is the original Gaussian $p(x|q) = \mathcal{N}(x; 5, 0.5)$ and the mask probabilities are $p(m = 0) = p(m = 1) = 0.5$

- additive acoustic environment model – in time domain, and approximately in the power spectral domain with sufficient smoothing: $x = s + n$
- MAX acoustic environment model (Nadas et al., 1989) – in the log spectral or log filterbank domain: $x = \max\{s, n\}$

both imply that the values of the clean speech must be *below* the observed values of the noisy mixture. This can be used to score the states on how likely they are to have produced an observation below $\mathbf{o}_m(t)$ (Holmes and Sedgwick, 1986; Cooke et al., 1994a; Green et al., 1995):

$$p(\mathbf{o}_m|q) = \int_{-\infty}^{\mathbf{o}_m} p(\mathbf{x}_m|q)d\mathbf{x}_m \tag{5.30}$$

where $p(\mathbf{x}_m|q) = \int p(\mathbf{x}_p, \mathbf{x}_m|q)d\mathbf{x}_p$ is itself a marginal p.d.f.

Additionally, in the log–spectral domain the energies should be positive, assuming that only the compressive domain range $[1, +\infty)$ of the logarithm function is used. So a stricter:

$$p(\mathbf{o}_m|q) = \int_{0}^{\mathbf{o}_m} p(\mathbf{x}_m|q)d\mathbf{x}_m \tag{5.31}$$

can be used.[12] For imputation, bounding the marginal in the denominator of Eq. 5.26 and constraining the imputed values to fall within the range also improves the results (Cooke et al., 2001).

The marginal p.d.f.s for $\mathbf{x}_p$ and $\mathbf{x}_m$ need not be computed separately. The additional knowledge about the admissible range of $\mathbf{x}_m$ can be utilised directly:

$$\begin{aligned} p(\mathbf{o}_p, \mathbf{x}_m \in [0, \mathbf{o}_m]|q) &= \int_0^{\mathbf{o}_m} p(\mathbf{x}_p, \mathbf{x}_m|q)d\mathbf{x}_m \\ &= \sum_k P(k|q)p(\mathbf{o}_p|k,q) \int_0^{\mathbf{o}_m} p(\mathbf{x}_m|k,q)d\mathbf{x}_m \end{aligned} \tag{5.32}$$

[12] strictly, the p.d.f. should be truncated at 0, but the truncated forms are inconvenient to work with and the probability mass left of 0 is negligible in most of the cases

where it is assumed that $p(\mathbf{x}|q)$ is a sum of factorisable distributions, $p(\mathbf{x}|q) = \sum_k P(k|q)p(\mathbf{x}|k,q)$.

For the special case of Gaussians with diagonal covariance matrices it can be evaluated by:

$$p(\mathbf{o}_p, \mathbf{x}_m \in [0, \mathbf{o}_m]|q) \quad = \quad 0.5 \sum_k P(k|q) \prod_p \mathcal{N}(o_p; \mu_{p,k,q}, \sigma^2_{p,k,q}) \prod_m \left\{ erf(\frac{o_m - \mu_{m,k}}{\sqrt{2}\ \sigma_{m,k,q}}) - erf(\frac{-\mu_{m,k}}{\sqrt{2}\ \sigma_{m,k,q}}) \right\} \tag{5.33}$$

where the error function $erf(x)$ is defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{5.34}$$

The assumption that Gaussians in the mixture are diagonal allows for the closed form solution.

In the case of non–diagonal Gaussians the integral in Eq. (5.32) doesn't have a convenient closed form solution. The covariance can be "diagonalised" with a suitable linear transform of the variables of integration. But then the variables of integration appear in the bounds. So it is not possible to decompose the multivariate integral into a form involving only one dimensional ones (like in Eq. (5.33)). The solution has to be either approximated (as in Morris et al. (1998)) or evaluated with Monte-Carlo type methods (Genz, 1992, 1993) which are unsuitable for ASR due to computational constraints.[13]

Counterevidence with marginalisation

Using counterevidence with marginalisation is straightforward. Instead of $p(\mathbf{o}|\mathbf{m}, q) = p(\mathbf{o}_p|q)$ (Eq. (5.11)) for the state p.d.f. of the HMM system, a more constrained:

$$p(\mathbf{o}|\mathbf{m}, q) = p(\mathbf{o}_p, \mathbf{x}_m \in [0, \mathbf{o}_m]) \tag{5.35}$$

is used.

Counterevidence with data imputation

Counterevidence can be used twice with data imputation (we will apply it only to state dependent data imputation, Section 5.5.5).

Firstly, when the conditional p.d.f. is computed, instead of $p(\mathbf{x}_m|\mathbf{x}_p, q)$ (Eq. (5.25)) we have:

$$p(\mathbf{x}_m|\mathbf{x}_p, \mathbf{x}_m \in [0, \mathbf{o}_m], q) = \frac{p(\mathbf{x}_m, \mathbf{x}_p|q)}{p(\mathbf{x}_p, \mathbf{x}_m \in [0, \mathbf{o}_m]|q)} \tag{5.36}$$

The integral in the denominator becomes a bounded one taking into account the bounds constraint.

Secondly, when a point from the conditional p.d.f. above is drawn, it has to be in the $[0, \mathbf{o}_m]$ interval. Regardless of whether a mean or a mode is imputed, one has to consider the case when the chosen point is not in the interval of $[0, \mathbf{o}_m]$. In our experiments (see Section 6.3.2) the highest point among all Gaussians in the mixture that is within the bounds was chosen as a value to be imputed.

The relative merits of evidence and counterevidence

There is an inherent problem in mixing the contributions of the present and the missing features together: the former are likelihoods, i.e. points on the p.d.f. curve (and can take any value), while the latter are true probabilities, i.e. points on the c.d.f. curve. If the mask is discrete (i.e. probabilities of present and missing data $p(m = 0)$ and $p(m = 1)$ are either 0 or 1) the scores used in the Viterbi search during the decoding need not be correct up to a scaling factor, as long as

[13] the form is the same as one discussed in the Appendix C which arises in the case of a linear transform of the feature vector

the factor is the same for all states. The optimal path does not change when the likelihoods of all states are scaled by the same factor. Hence using the curve (a) or (b) from Fig. (5.4) makes no difference to the winning path if the mask is discrete . Even with non–discrete mask the difference in the contributions of the counterevidence to different states is small.

Figure 5.4 plots an example of three different ways of mixing the evidence from the present and missing data, together with the original Gaussian $p(x|q)$. It's notable that the shape of the curves is virtually identical. The main difference is in the scores at high valued observations o (disregarding the scaling factor between them). Curve (c) was found to perform poorer then (a) and (b) in our experiments. Both (a) and (b) perform identically for ordinary Viterbi single–source search. Barker et al. (2000) reported problems with (b) in the multisource decoder, as different paths in this decoder see the data differently (as present or missing, with discrete mask). The problem was circumvented by either using an empirically established scaling constant, or by using the "average likelihood" (curve (a)).

The problem of the relative contributions of the present and missing data is a reminiscent of the problem of using the acoustic and the language model in the ASR system together, when they were estimated separately.[14] The probabilities given by the acoustic model are overoptimistic, most probably due to the assumption that the frames are independent. The "fudge factor" γ (in addition to some other empirically derived parameters, like the word insertion penalty) is used as in:

$$W^* = \underset{W}{argmax} P(O|W)P(W)^{\gamma} \tag{5.37}$$

instead of Eq. (5.1) to weight the relative contributions between the evidence from the acoustic and the language models. These "adjustment factors" are usually tuned to minimise the word error rate (WER) on a separate, "development" test set. Then their "optimal" values are used in the final evaluation of the ASR system on a different "evaluation" test set.

If both the mask and the speech models are estimated jointly, then usage of such empirical factors may be avoided. The average likelihood (a) maybe the safest choice for expressing the counterevidence, as it at least keeps the score "dimensionally correct", was shown to work well in the multisource decoder (Barker et al., 2000) and doesn't hinder the performance of the single source Viterbi decoder as the "self–weighted likelihood" (c) does.

5.6 Summary

Techniques that cater for missing and unreliable data were integrated into an HMM based ASR system in this chapter. The adaptation enables the system that models a single speech source to handle observations that are a mixture of several sources. This is achieved without explicit models for all of them, nor their decoding in parallel. The approach is inspired by the HSR which seems able to attend to one source in the mixture alone, neglecting the others. The MD model for speech recognition introduces the notion of a mask as a random variable, whose constraints can be captured by an appropriate model. It is further discussed how marginalisation and data imputation can be used to implement the adaptation of the speech HMM. The changes needed in the HMM based system are fairly straightforward. The role of counterevidence and how it fits in an HMM system is also explored. The implementation makes use of the function of acoustic environment to circumvent the need for explicit noise models, while still capturing most of the constraints. It also makes use of the special forms both of the speech and the mask model (factorisable or sum of factorisable p.d.f.s) to achieve efficient computation of the mask conditioned likelihoods.

[14] in addition, most often, when the acoustic model is estimated the wrong criterion is optimised: the likelihood is maximised instead of minimising the word error rate

Chapter 6

Experiments

6.1 Introduction

The aim of this chapter is to present the results of the experiments with the Missing Data (MD) ASR system. The connected digits task was chosen to test the techniques which is a de–facto standard test for robust ASR. It has the advantage that whole word models suffice (there is no need for phonetic dictionary) and there is no language model. Arguably, it is still a non–trivial task, while leaving out components of the ASR system that have less influence on the robustness of the system. The speech data was artificially contaminated with various noises at different SNRs. The MD ASR system was an HMM based one, employing a textbook training procedures (Young and Woodland, 1993) during models training and a textbook in–house Viterbi decoder during the recognition (in several different implementations). The MD system was tested in various configurations. The features were constrained to be in frequency domain, as the mask estimation and recognition used the same features. The experiments included filterbanks (24 channels), ratemaps (32 and 64 channels) both with or without the first derivatives. The mask estimation in the experiments still makes use of a noise estimate. A weak noise model was estimated on–line during the recognition. It was mostly stationary, i.e. constant for the duration of the utterance. Apriori mask (which requires knowledge of the clean speech) was used to get an indication of the performance that may be achieved with very good separation. Both marginalisation and data imputation were tried for computing the likelihood of the partial data in the initial experiments. In the latter set of experiments the marginalisation technique only was used, as it is faster and data reconstruction was not required by the task.

6.2 Description of the MD ASR system and the corpora

An HMMs based ASR system in various configurations, adapted to handle missing and unreliable data, was used for all experiments reported in this chapter. The system was trained and tested on the TIdigits database (Leonard, 1984) mixed with Lynx helicopter and factory noise from the NOISEX database (Varga et al., 1992), as well as TIdigits' noisy variant Aurora 2 (Hirsch and Pearce, 2000).

The TIdigits database consists of digit strings containing between one and seven digits ("1" to "9", "oh", "zero") recorded in quiet conditions and sampled at 20 kHz. The male and female corpora (leaving out the digits strings spoken by children) were used together (no gender dependent modelling) both for training and testing. The "canonical" TIdigits trainset (all clean speech) was used for training.

Lynx helicopter noise from NOISEX was used as an example of stationary, and the factory noise as an example of non–stationary noise. They were both added with random starting points at SNRs from +20dB to -5dB to a subset of the TIdigits test set consisting of 240 digit strings which were used for testing in the experiments with the NOISEX noises.

The Aurora 2 database was used in the later set of experiments. It is based on the TIdigits database. The TIdigits sentences have been downsampled from 20 kHz to 8kHz and various distortions have been artificially added. The subset with additive noise contains speech mixed with 8 different noise types: suburban train, babble (crowd of people), car, exhibition hall, restaurant, street, airport, train station. The noise signals were added at SNRs from 20 dB to -5 dB. They all contain both stationary and non-stationary components to various degrees: from car noise and exhibition hall which are mostly stationary, to street and airport noises which are very non–stationary. The first four noises are used during the multiconditional training regime (training with data contaminated with noise) for inferring noisy models. All eight noises feature in the testing set. The first four noises form the subset testa; the last four form the subset testb.

The ASR system used was a "textbook" HMM based one. Each digit ("1" to "9", "zero" and "oh") was modelled by a single "straight–through" HMM with "self" and "next" non–zero transition only. The number of states was the same for all digits and varied from 8 for the TIdigits+NOISEX experiments to 16 for the Aurora 2 experiments. Each state had a Gaussian mixture p.d.f. with up to 10 Gaussians in the mixture. The grammar consisted of a silence (represented with a single model) followed by any number of digits followed by silence. Occasionally a distinction was made between the "long silence" on the beginning and the end of the sentences and a "short", interword silence.

The small vocabulary task of connected digits recognition[1] was considered a convenient platform for exploration of the robustness aspects of the ASR. Arguably it is still a non–trivial task, while the grammar is minimal and there is no need for a phonetic dictionary. The aim was to reduce the impact of the components deemed unlikely to be the "core technology" of a robust ASR system (phonetic dictionary, complex language model).

The system was trained using the HTK (Young and Woodland, 1993) in various versions (from 1.5 to 3.0). For testing an "in–house" vectorised Matlab based decoder was used for the former, and the CASA toolkit (Barker, 2000) (CTK) for the later set of experiments.

6.3 Experiments with NOISEX factory and Lynx helicopter noises

6.3.1 Speech/noise separation

The separation of the speech and noise in the time–frequency plane was accomplished by deriving a mask $\mathbf{m}(t)$ (as described in Section 5.3). As every point is assumed to be either speech or noise only, a (non–stationary) binary mask is enough to define the separation completely. Further text will assume a notation where the a mask of $m_i(t) = 1$ signals speech, while a mask of $m_i(t) = 0$ signals noise (i being the frequency band). The noise and SNR estimation were carried out in spectral magnitude domain, after the binning of the FFT magnitude and computing the magnitude of the filters, but before the compressive non–linearity in the front–end.

Spectral Subtraction based masks (SS)

The spectral subtraction (SS) based masks were derived by assuming the points where the (non–adaptive) spectral subtraction failed and ended with negative spectral magnitude, are noise (Drygajlo and El-Maliki, 1998a). All the other points were considered speech:

$$m_i(t) = \begin{cases} 1 & \text{if } s_i(t) + n_i(t) \geq \hat{n}_i, \\ 0 & \text{otherwise,} \end{cases} \tag{6.1}$$

where $s_i(t) + n_i(t)$ is the noisy speech feature i at time t, and $\hat{\mathbf{n}}$ is a stationary noise estimate (constant over the duration of the whole utterance) computed as a mean of the first 10 frames in the utterance.

[1] Morris et al. (1998) have reported on Missing Data experiments on a medium vocabulary Resource Management task with artificial random masks

Signal-to-Noise Ratio based masks (SNR)

The noise $\hat{\mathbf{n}}$ was estimated as before, as a mean of the first 10 frames in the sentence. It was further assumed that the speech and the noise are additive in the spectral magnitude domain. These assumptions make thresholding of the SNR estimate possible (which is much more accurate then the SNR estimate itself). The speech mask is computed as as:

$$m_i(t) = \begin{cases} 1 & \text{if } s_i(t) + n_i(t) \geq \hat{n}_i(1 + 10^{\frac{THR}{20}}), \\ 0 & \text{otherwise,} \end{cases} \tag{6.2}$$

where THR is the threshold in [dB]. A threshold of $THR = 7dB$ was found to work well with a range of SNRs, and the results were not too sensitive to the choice of this value.

Oracle/apriori masks (APR)

The oracle or apriori masks assume knowledge of the clean speech $\mathbf{s}(t)$. It is further assumed that the speech and noise are additive in the spectral magnitude domain. With those assumptions in hand it is possible compute the apriori SNR estimate[2] and/or to threshold it. The speech mask was obtained as:

$$m_i(t) = \begin{cases} 1 & \text{if } s_i(t) + n_i(t) < s_i(t)(1 + 10^{\frac{-THR}{20}}), \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

Again a threshold of $THR = 7dB$ was found to work well with a range of SNRs.

6.3.2 Computing the likelihood of the partially observed data

Computing the likelihood of the partially observed data (see Section 5.5) $p(\mathbf{o}(t)|\mathbf{m}(t), q(t), W)$ is all that is needed to adapt an HMM system (trained on non–censored data) to handle the partial data. The system was used both with marginalisation and imputation of the missing data. The following techniques were tried:

Marginalisation (MG)

Marginalisation of the missing data:

$$p(\mathbf{o}(t)|\mathbf{m}(t), q(t), W) = p(\mathbf{o}_p(t)|q(t), W) \tag{6.4}$$

where $p(\mathbf{o}_p(t)|q(t), W)$ is computed as in Eq. (5.19). With the diagonal GMM state p.d.f.s used this amounts to disregarding the missing features.

It was further noted that if a noise estimate is available (e.g. with SS or SNR masks), subtracting the noise slightly improves the accuracy.

Bounded marginalisation (BMG)

Marginalisation of the missing data while taking into account the counter–evidence constraint:

$$p(\mathbf{o}(t)|\mathbf{m}(t), q(t), W) = p(\mathbf{o}_p(t), \mathbf{x}_m(t) \in [0, \mathbf{o}_m]|q(t), W) \tag{6.5}$$

(Eq. (5.35)) where $p(\mathbf{o}_p(t), \mathbf{x}_m(t) \in [0, \mathbf{o}_m]|q(t), W)$ is computed as in Eqs. (5.32) and (5.33). With the diagonal GMM state p.d.f.s used for each missing feature $x_i(t)$ the area beneath the p.d.f. and between 0 and $o_i(t)$ is computed.

[2] the SNR is still estimate – although the speech $\mathbf{s}$ is known, it is still an assumption that the speech and noise are additive in the particular domain

State–based data imputation (SDI)

Imputation of the missing features by drawing a point from the conditional state p.d.f.:

$$p(\mathbf{x}_m(t)|\mathbf{o}_p(t), q(t)) = \frac{p(\mathbf{x}_m(t), \mathbf{o}_p(t)|q(t))}{p(\mathbf{o}_p(t)|q(t))} \tag{6.6}$$

In all experiments with SDI the mean of the p.d.f. was computed:

$$\hat{\mathbf{o}}_m(t) = \mathcal{E}_{\mathbf{x}_m|\mathbf{o}_p(t),q(t)}\{\mathbf{x}_m\} \tag{6.7}$$

as in Eq. (5.26) and all subsequent processing was on the vector $(\mathbf{x}_p(t), \hat{\mathbf{o}}_m(t))$ which has no data missing.

In every time frame the possible imputations of all states are computed. Hence there are as many versions of the frame as there are states. During the emission probability calculation, for each state only the likelihood of the feature vector with $\mathbf{x}_m$'s filled from the p.d.f. of the same state is computed i.e. $p(\mathbf{o}_p(t), \hat{\mathbf{o}}_m(t)|q(t))$.

Bounded state–based data imputation (BSDI)

Imputation of the missing features by drawing a point from the conditional state p.d.f. constrained in the range of $[0, \mathbf{o}(t)]$ (see Section 5.5.7). Firstly the conditional state p.d.f. was computed as:

$$p(\mathbf{x}_m|\mathbf{o}_p(t), \mathbf{x}_m \in [0, \mathbf{o}_m(t)], q(t)) = \frac{p(\mathbf{x}_m, \mathbf{o}_p(t)|q(t))}{p(\mathbf{o}_p(t), \mathbf{x}_m \in [0, \mathbf{o}_m(t)]|q(t))} \tag{6.8}$$

as in Eq. (5.36).

Secondly, a point from the conditional p.d.f. was chosen as $\hat{\mathbf{o}}_m(t)$. It was noted that in most cases the Gaussians in the mixture were far away in at least one dimension, resulting in a multimodal conditional p.d.f. For each Gaussian in the mixture, the point with the highest density that is within the bounds $[0, \mathbf{o}_m(t)]$ was chosen as a possible candidate for $\hat{\mathbf{o}}_m(t)$. Then among all points the one with the highest overall posterior probability (i.e. taking into account the "responsibilities" together with the likelihood) was chosen as value for $\hat{\mathbf{o}}_m(t)$. All subsequent processing was on the "completed" vector $(\mathbf{x}_p(t), \hat{\mathbf{o}}_m(t))$.

6.3.3 Results with 64 channel ratemap features

The acoustic vectors consisted of smooth outputs of from 64–channel auditory filter bank (centre frequencies spaced linearly in ERB scale from 50 to 8000Hz), computed every 10ms (Cooke, 1991). HTK (Young and Woodland, 1993) was used for training, and a local MATLAB decoder for recognition. Twelve models ('1'–'9', 'oh', 'zero' and 'silence') consisting of 8 no–skip, straight–through states with observations modelled with a 10 component diagonal Gaussian mixture were trained on clean speech. Stationary Lynx helicopter noise as well as non–stationary factory noise from the NOISEX database was added (with random start points) at SNRs from +20dB to -5dB to a subset of the TIdigits test set consisting of 240 digit strings used for testing. In all graphs, the X-axis is the SNR, decreasing from clean, 20 dB, to -5 dB in 5 dB steps, while the Y-axis depicts the recognition accuracy. In all cases the factory noise proved to be a harder task due to its non–stationarity.

Masks based on spectral subtraction

The purpose of the experiment was to check whether treating the points where SS failed as missing would produce any improvements over SS alone.

Figures 6.1, 6.2 and 6.3 depict the results on factory noise. Figures 6.4, 6.5 and 6.6 depict the results on Lynx helicopter noise.

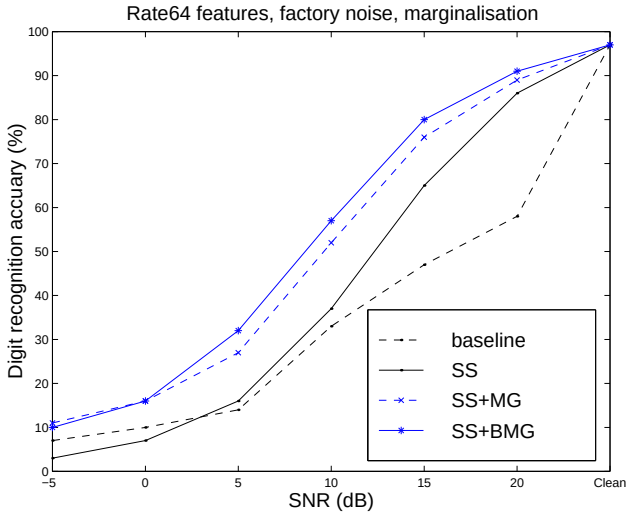


Figure 6.1: Marginalisation compared with spectral subtraction on factory noise (64–channel ratemap features)

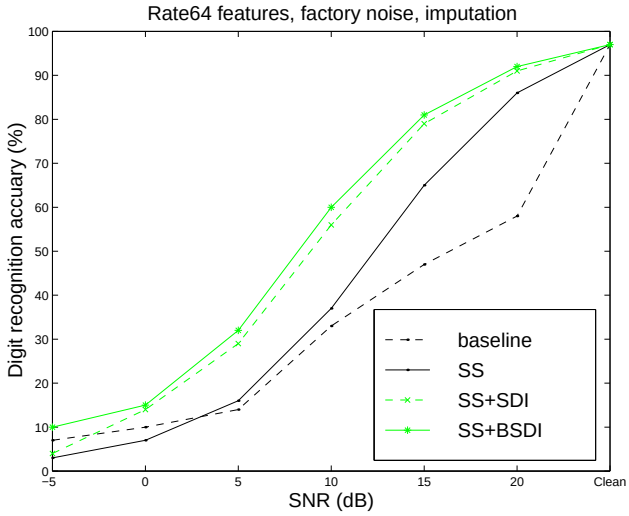
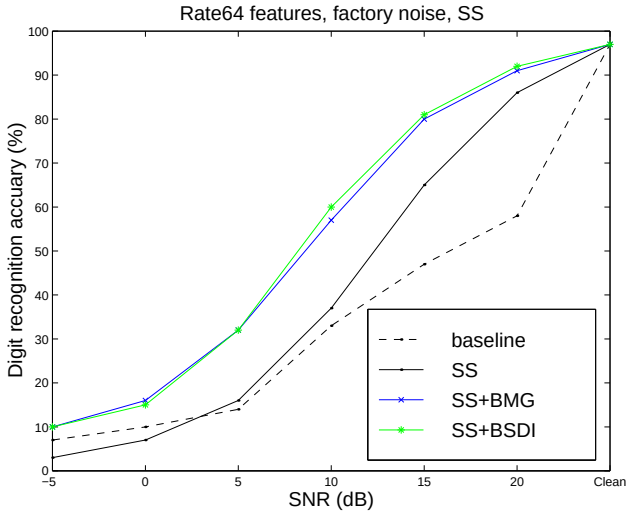


Figure 6.2: Data imputation compared with spectral subtraction on factory noise (64–channel ratemap features)



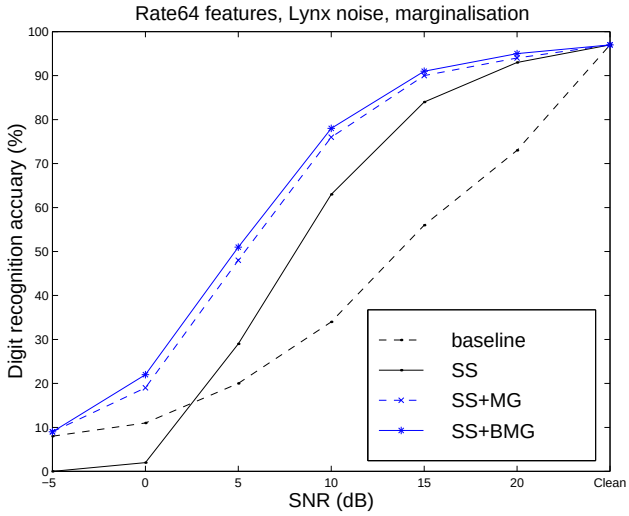


Figure 6.4: Marginalisation compared with spectral subtraction on Lynx noise (64–channel ratemap features)

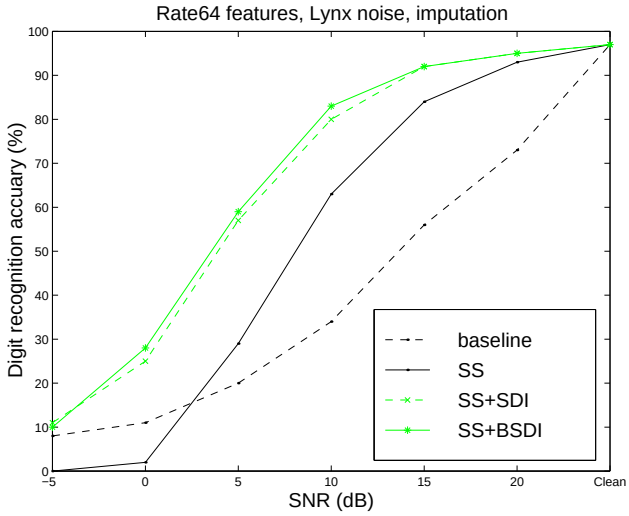
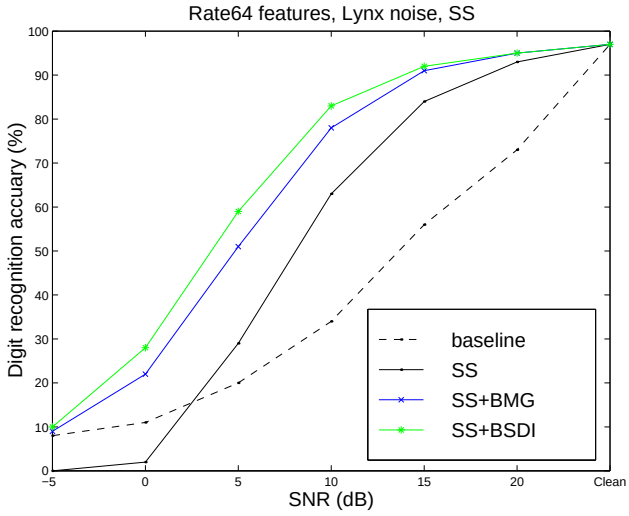


Figure 6.5: Data imputation compared with spectral subtraction on Lynx noise (64–channel ratemap features)



In all cases a discrete non–stationary SS mask was derived. The baseline and the SS curve are the accuracy of the recogniser without any compensation and with spectral subtraction respectively. The noise estimate for SS was the same one as for deriving the SS mask.

Figures 6.1 and 6.4 show the results of the marginalisation (MG) and the bounded marginalisation (BMG) technique. Figures 6.2 and 6.5 show the results of state based data imputation (SDI) and bounded SDI (BSDI). In all cases treating the points where SS failed as missing improves the results. Further, using the bounds constraint gives slight, but consistent advantage.

Figures 6.3 and 6.6 compare the improvements between the bounded marginalisation and state based data imputation. For factory noise they are mostly the same, with BSDI performing only slightly better at 10 dB. For Lynx helicopter noise, it seams BSDI outperforms BMG at all SNRs.

Masks based on local SNR estimation

Figures 6.7, 6.8 and 6.9 depict the results on factory noise. Figures 6.10, 6.11 and 6.12 depict the results on Lynx helicopter noise.

In all cases a discrete non–stationary SNR mask was derived with a 7 dB threshold. The baseline and the SS curve are the accuracy of the recogniser without any compensation and with spectral subtraction respectively. The noise estimate for SS was the same one as for deriving the SNR mask.

Figures 6.7 and 6.10 show the results of the marginalisation (MG) and the bounded marginalisation (BMG) technique. Figures 6.8 and 6.11 show the results of state based data imputation (SDI) and bounded SDI (BSDI). In both cases when no bounds are used (MG and SDI) the accuracy suffers at mid to high SNRs and both perform worse then SS. They do perform better then SS at low SNRs. It was noted that in both cases a major source of errors are random insertions in frames where there is little or no data at all (all features in the frame are noisy). Introducing the bounds both with marginalisation (BMG) and state based data imputation (BSDI) rectifies this, as bounds make the silence model win in the quiet frames where the speech was swamped by noise. The accuracy is improved, and both BMG and BSDI outperform SS significantly at all SNRs.

Figures 6.9 and 6.12 compare the improvements between the bounded marginalisation and state based data imputation. Both for factory noise and Lynx helicopter noise, BMG seems to outperform BSDI at all SNRs. It seems that the SNR masks, which let less but more reliable data in (compared to the SS masks) suit marginalisation better, while SS masks (letting more, but less reliable data) suit data imputation better (see Figure 5.2 for masks example).

Apriori masks

The apriori masks are derived from the clean and noisy speech (as described on pp. 85). They are indicative of the performance that may be achieved with very good separation. They are also useful in assessing the performance of different methods for computing the likelihood of the partial data independently of the separation front–end.

Figures 6.13, 6.14 and 6.15 depict the results on factory noise. Figures 6.16, 6.17 and 6.18 depict the results on Lynx helicopter noise.

In all cases a discrete non–stationary APR mask was derived, with an estimated SNR threshold of 18 dB. The baseline and the SS curve are the accuracy of the recogniser without any compensation and with spectral subtraction respectively, and are plotted as indication only.

Figures 6.13 and 6.16 show the results of the marginalisation (MG) and the bounded marginalisation (BMG) technique. Figures 6.14 and 6.17 show the results of state based data imputation (SDI) and bounded SDI (BSDI). The trends are mostly similar to the ones observed with SNR mask (previous section). In both cases when no bounds are used (MG and SDI) the accuracy suffers at mid to high SNRs. The drop is most dramatic with MG – accuracy decreasing sharply when going from clean speech to 20dB, then staying mostly flat down to low SNRs. As noted before, a major source of errors are random insertions in the frames where there is little or no data at all. This is even more pronounced with APR masks (compared to SNR masks). Introducing

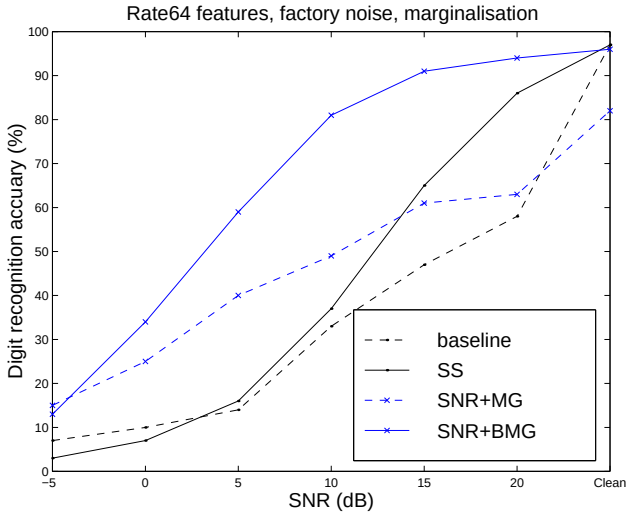


Figure 6.7: Marginalisation with SNR mask, spectral subtraction and the baseline on factory noise (64–channel ratemap features)

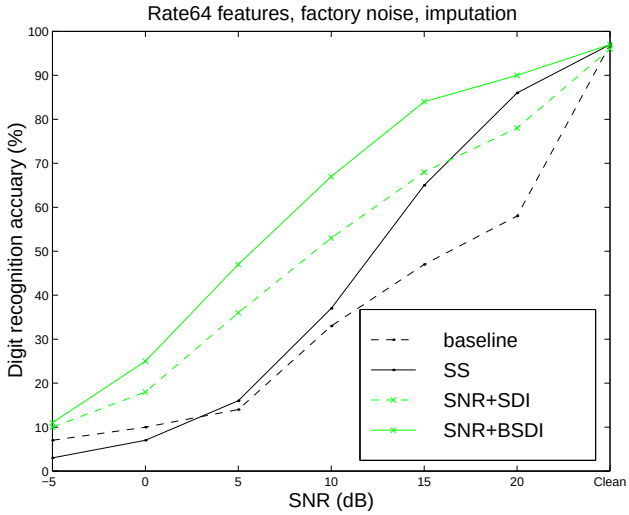
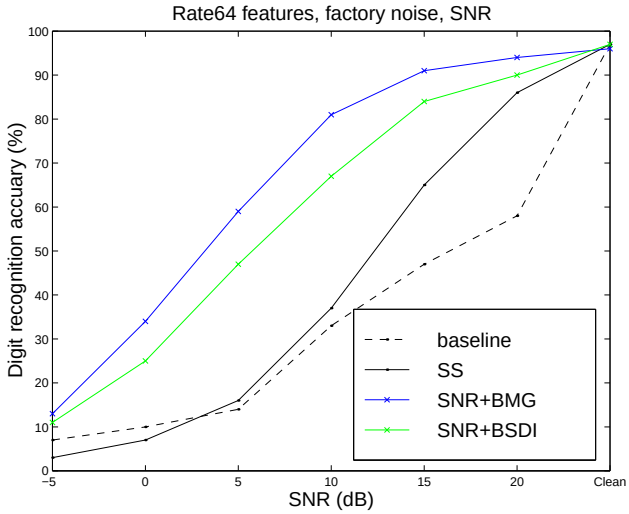


Figure 6.8: Data imputation with SNR mask, spectral subtraction and the baseline on factory noise (64–channel ratemap features)



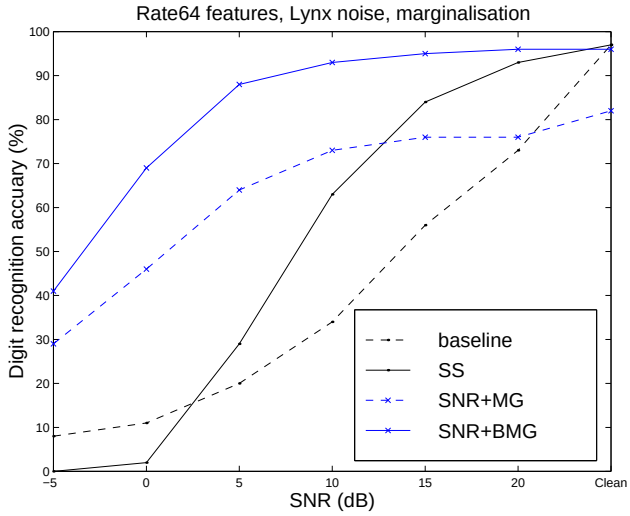


Figure 6.10: Marginalisation with SNR mask, spectral subtraction and the baseline on Lynx noise (64–channel ratemap features)

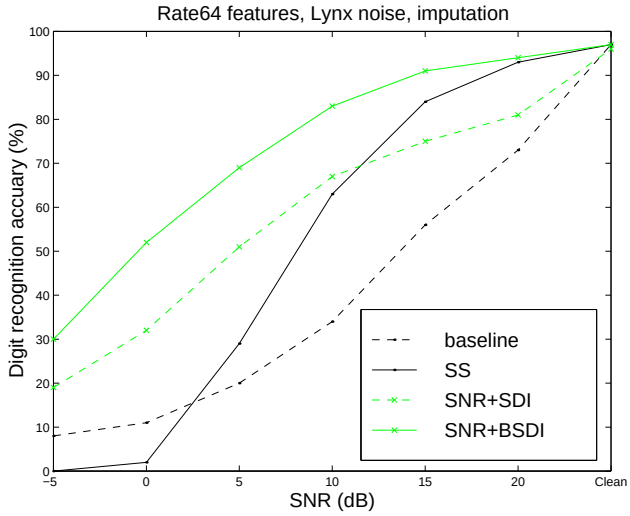
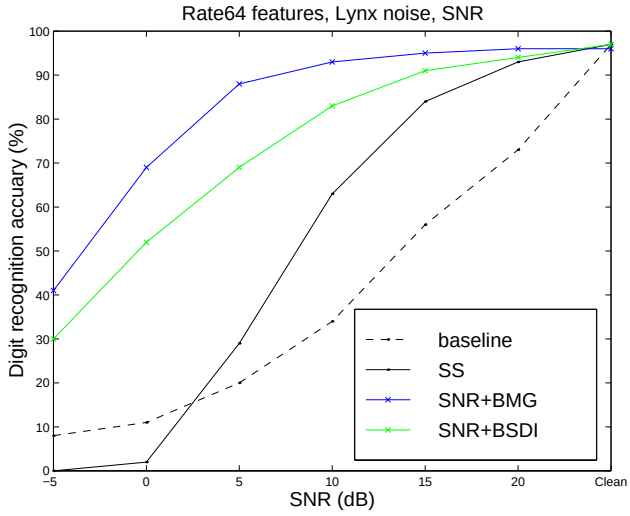


Figure 6.11: Data imputation with SNR mask, spectral subtraction and the baseline on Lynx noise (64–channel ratemap features)



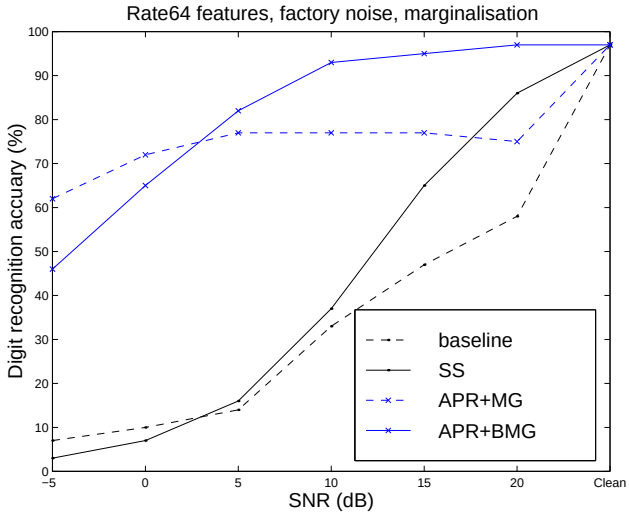


Figure 6.13: Marginalisation with APR mask on factory noise (64–channel ratemap features)

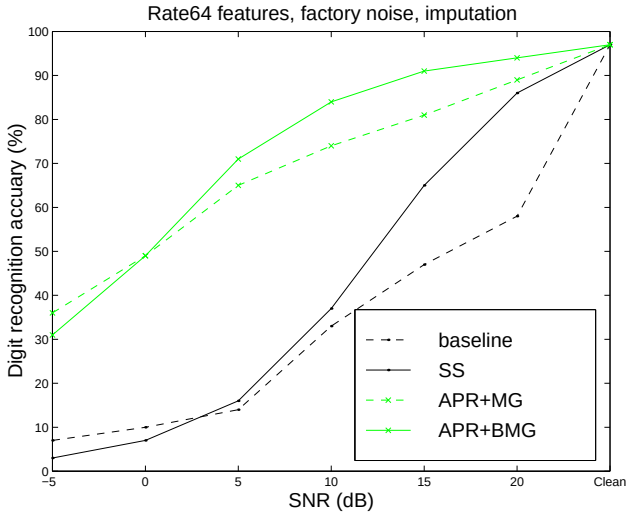


Figure 6.14: Data imputation with APR mask on factory noise (64–channel ratemap features)

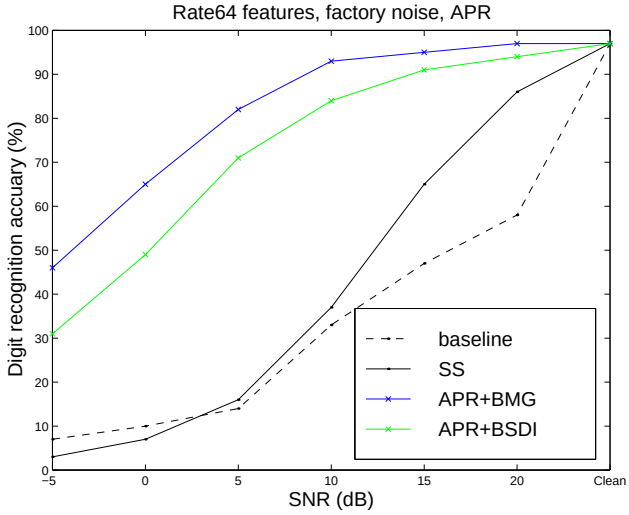


Figure 6.15: Bounded marginalisation and data imputation with APR mask on factory noise (64–channel ratemap features)

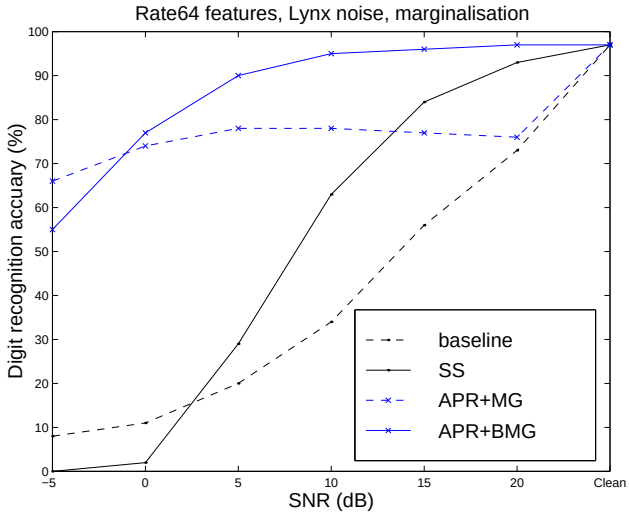


Figure 6.16: Marginalisation with APR mask on Lynx noise (64–channel ratemap features)

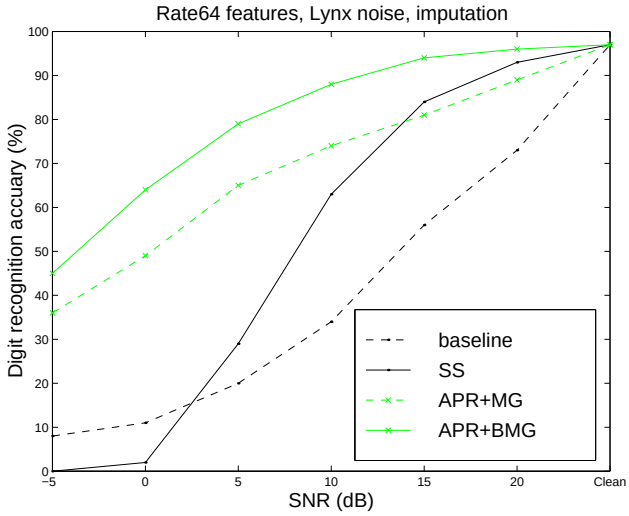


Figure 6.17: Data imputation with APR mask on Lynx noise (64–channel ratemap features)

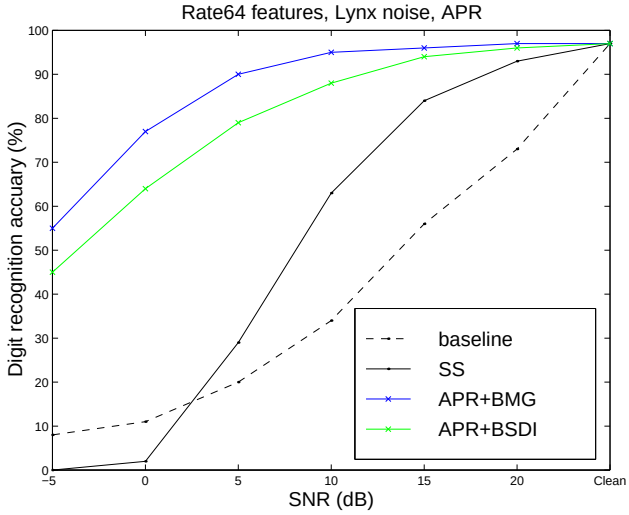


Figure 6.18: Bounded marginalisation and data imputation with APR mask on Lynx noise (64–channel ratemap features)

the bounds both with marginalisation (BMG) and state based data imputation (BSDI) rectifies this, as bounds make the silence model win in the quiet frames where the speech was swamped by noise. The accuracy improves at all SNRs and with both noises.

Figures 6.15 and 6.18 compare the improvements between the bounded marginalisation and state based data imputation. Both for factory noise and Lynx helicopter noise, BMG seems to outperform BSDI at all SNRs. As discussed in Section 5.5.6, imputation seems a harder task then marginalisation, as speech reconstruction is attempted, in addition to computing the likelihood of the partial data.

Apriori mask threshold

The apriori masks are derived from the clean and noisy speech, and are indicative of the performance that may be achieved with very good separation. Figures 6.19, 6.20, 6.21 and 6.22 show the sensitivity to the choice of a threshold value for ratemap features.

Figures 6.19 and 6.20 depict the results on factory noise. Figures 6.21 and 6.22 depict the results on Lynx helicopter noise.

In all cases a discrete non–stationary APR mask was derived by comparing the clean and the noisy speech. APR18 stands for SNR threshold of 18.27 dB (clean and noisy speech differ 1 dB or less). APR8 stands for SNR threshold of 7.69 dB (clean and noisy speech differ 3dB or less). APR0 stands for SNR threshold of 0.04dB (clean and noisy speech differ 6dB or less).

Figures 6.19 and 6.21 show that both with factory and Lynx noise, MG is slightly better off with a higher threshold of 18.27 dB then a lower one of 7.69 dB. The opposite, but to a greater degree, is true for BMG: BMG is better off with a lower threshold of 7.69 dB then a larger one of 18.27 dB, and the difference is more pronounced. This may be due to the imposition of the additional constraint – the bounds. Lacking this constraint, MG is more sensitive to noisy data getting through the mask. Whereas BMG is able to cope with more data, even it is of lesser quality.

Results for data imputation on Figures 6.20 and (6.22) are more consistent. Letting more data in (APR8 v.s. APR18) helps accuracy both with SDI and BSDI. Using bounds always improves accuracy at the same threshold. Data imputation, compared to marginalisation, seems more sensitive to lack of data then it is to its noisiness – using an even lower threshold of 0.04 dB (APR0 on Figure 6.20) increases the accuracy further.

However, the improvement seems to depend on the mask quality. Using similar thresholds with SNR ("real") masks does not lead to improved results there.

6.3.4 Results with 24 channel filterbank features

The experimental setup was similar as in the previous Section 6.3.3. The only difference was that the acoustic vectors consisted of 24 channel Mel-spaced triangular filterbank outputs (Young and Woodland, 1993) computed every 10ms.

Masks based on local SNR estimation

Figures 6.23, 6.24 and 6.25 depict the results on factory noise. Figures 6.26, 6.27 and 6.28 depict the results on Lynx helicopter noise.

In all cases a discrete non–stationary SNR mask was derived with a 7 dB threshold. The baseline and the SS curve are the accuracy of the recogniser without any compensation and with spectral subtraction respectively. The noise estimate for SS was the same one as for deriving the SNR mask.

All results largely mirror what has already been observed with SNR masks with 64–channel ratemap features.

Figures 6.23 and 6.26 show the results of the marginalisation (MG) and the bounded marginalisation (BMG) technique. Figures 6.24 and 6.27 show the results of state based data imputation

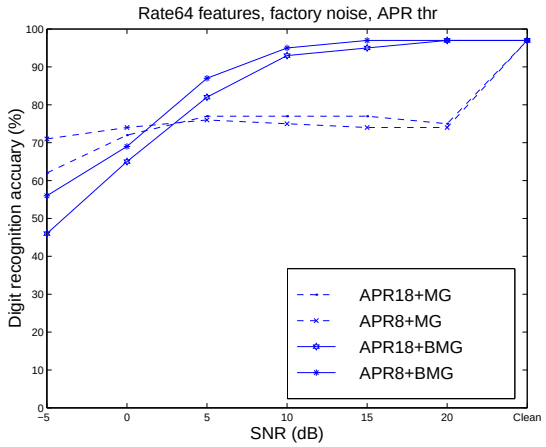


Figure 6.19: Marginalisation with APR mask with different thresholds on factory noise (64–channel ratemap features)

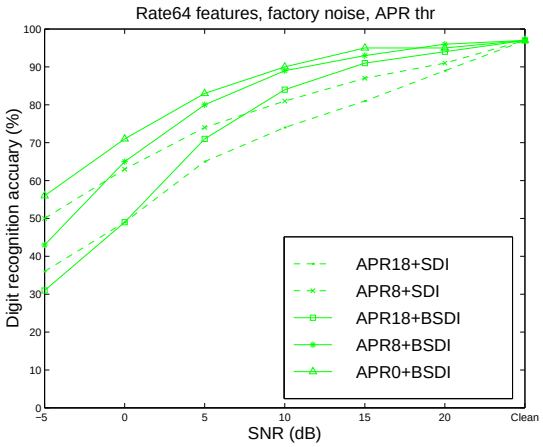


Figure 6.20: Data imputation with APR mask with different thresholds on factory noise (64–channel ratemap features)

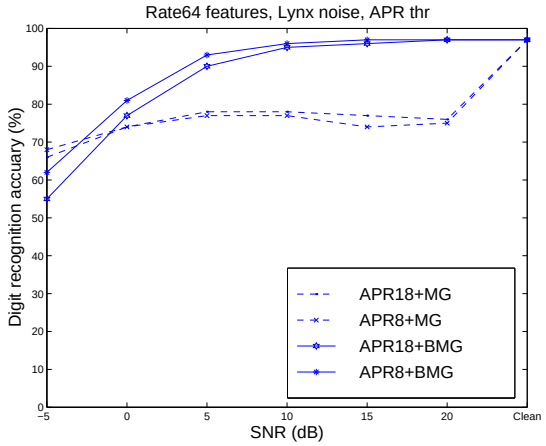


Figure 6.21: Marginalisation with APR mask with different thresholds on Lynx noise (64–channel ratemap features)

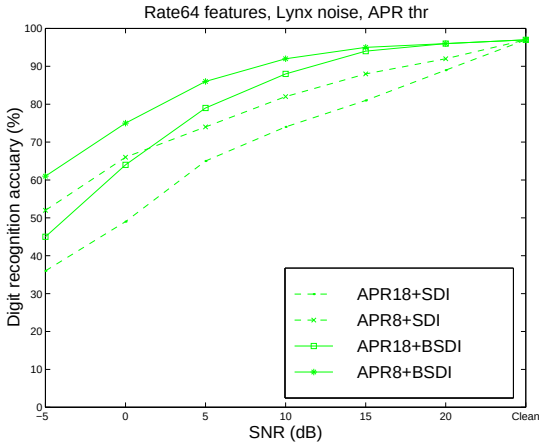


Figure 6.22: Data imputation with APR mask with different thresholds on Lynx noise (64–channel ratemap features)

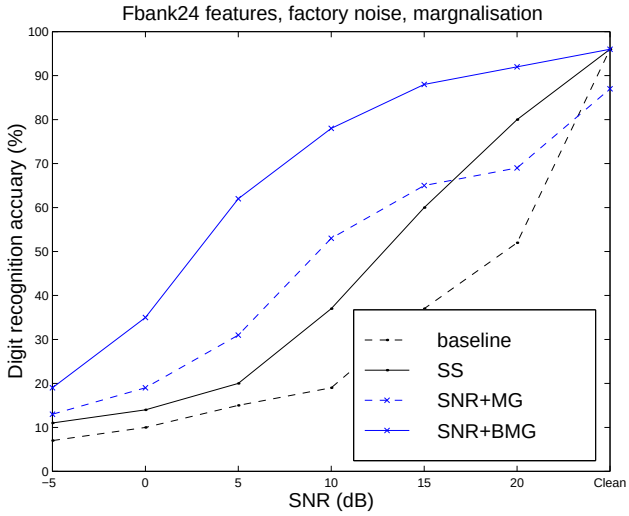


Figure 6.23: Marginalisation with SNR mask, spectral subtraction and the baseline on factory noise (24–channel filterbank features)

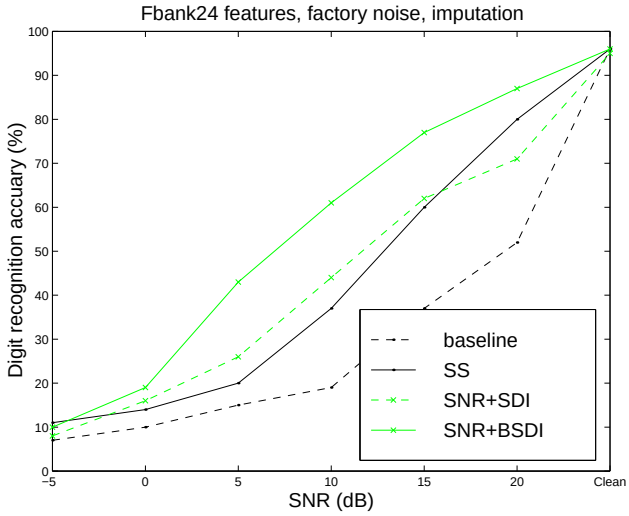
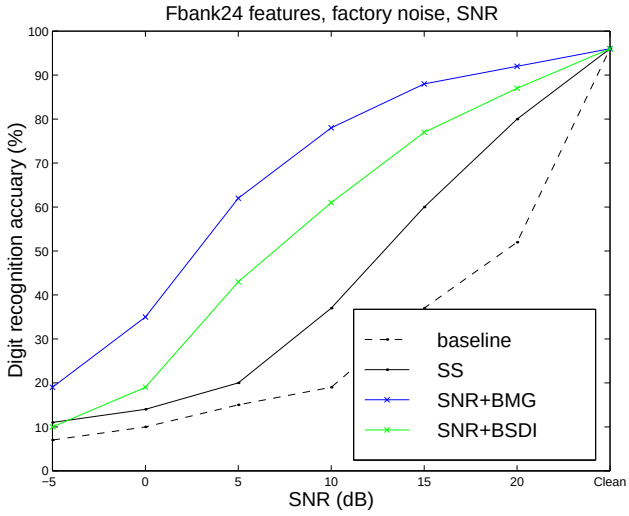


Figure 6.24: Data imputation with SNR mask, spectral subtraction and the baseline on factory noise (24–channel filterbank features)



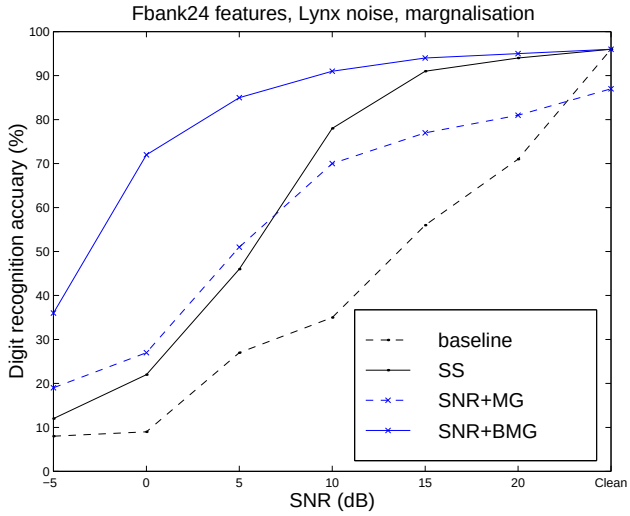


Figure 6.26: Marginalisation with SNR mask, spectral subtraction and the baseline on Lynx noise (24–channel filterbank features)

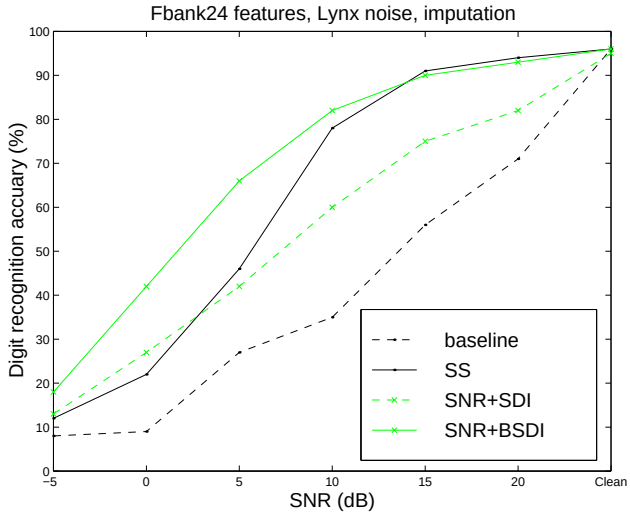
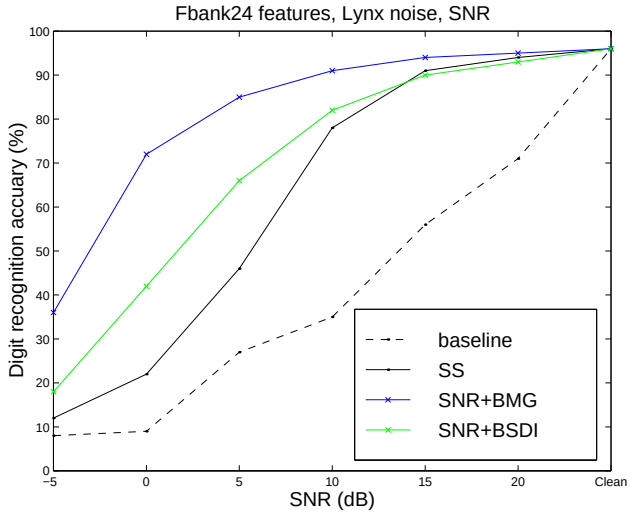


Figure 6.27: Data imputation with SNR mask, spectral subtraction and the baseline on Lynx noise (24–channel filterbank features)



(SDI) and bounded SDI (BSDI). In both cases when no bounds are used (MG and SDI) the accuracy suffers at mid to high SNRs and both perform worse then SS. The do perform better then SS at low SNRs. As previously noted, this is mostly due to random insertions in the frames where there is little or no data. Introducing the bounds both with marginalisation (BMG) and state based data imputation (BSDI) rectifies this, as bounds make the silence model win in the quiet frames where the speech was swamped by noise. The accuracy is improved, and both BMG and BSDI outperform SS significantly at all SNRs.

Figures 6.25 and 6.28 compare the improvements between the bounded marginalisation and state based data imputation. Both for factory noise and Lynx helicopter noise, BMG seems to outperform BSDI at all SNRs.

Apriori masks

As previously noted, the apriori masks are derived from the clean and noisy speech. They are indicative of the performance that may be achieved with very good separation. They are also useful in assessing the performance of different methods for computing the likelihood of the partial data independently of the separation front–end. The baseline and the SS curve are the accuracy of the recogniser without any compensation and with spectral subtraction respectively, and are plotted as indication only.

Figures 6.29, 6.30 and 6.31 depict the results on factory noise. Figures 6.32, 6.33 and 6.34 depict the results on Lynx helicopter noise.

Comparing the results with the APR masks on 64–channel ratemap features, it is notable that the threshold of 18.27 dB is not used anymore. With only 24 features (instead of 64), there isn't enough data left for ASR with such a stringent criterion for data quality. Thresholds of 7.69 dB (APR8) and 0.04 dB (APR0) were compared in various conditions and using the different MD techniques.

Figures 6.29 and 6.32 show the results of the marginalisation (MG) and the bounded marginalisation (BMG) technique. With both noises, the performance of the both techniques is better with the more stringent criterion. It seems that for marginalisation it's better to let less, but more reliable data in. Figures 6.30 and 6.33 show the results of state based data imputation (SDI) and bounded SDI (BSDI). Here, the opposite (compared to MG and BMG) seems to hold: letting more, but less reliable data in helps improving the accuracy (with the exception of SDI on Lynx noise).

Figures 6.31 and 6.34 compare the four techniques (MG, BMG, SDI, BSDI) at their best SNR threshold. As expected, using the bounds constraint improves the accuracy significantly. Also, BMG seems to consistently outperform BSDI, while MG is worse than SDI at higher SNRs and better at lower ones.

The trends in the results are mostly in line with the previously observed results on 64–channel ratemap features, with the notable exception of using a lower threshold (in general) with 24–channel filterbank features.

Using "cleaned" (clean) models

Figure 6.35 depicts the results on factory, and Figure 6.36 on Lynx helicopter noise, with models trained on clean speech that has been processed with SS. The stationary noise estimate was obtained as the mean of the first 10 frames of each sentence, and was subsequently subtracted. Although clean speech was used for training, the benefit of the process is that the models "learn" some of the artifacts introduced by the "cleaning process" (that is going to be used during testing latter) during training. These models are referred to as "cleaned models", whereas the models obtained by training on unprocessed clean speech are "clean models".

On both noises, SS with "cleaned models" (SScl) was compared with SS (with "clean models"), and so was bounded marginalisation (BMGcl) with SNR mask. In both noises SScl performs better then SS at all but the lowest SNRs. BMGcl outperforms BMG by a smaller margin, but from a higher baseline. It seems that this commonly used technique for improving the performance

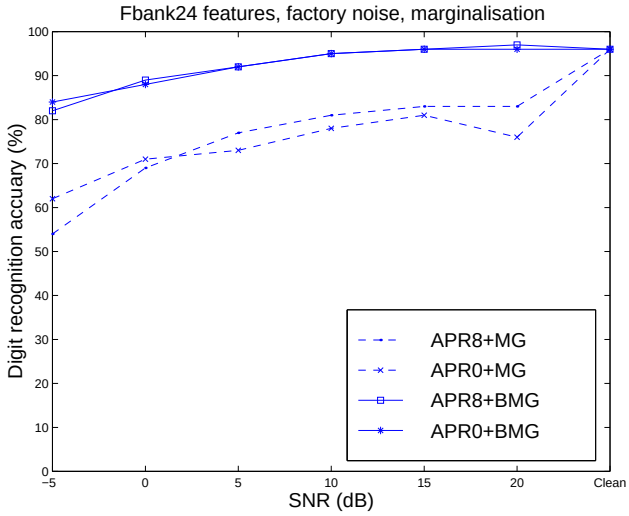


Figure 6.29: Marginalisation with APR mask on factory noise (24–channel filterbank features)

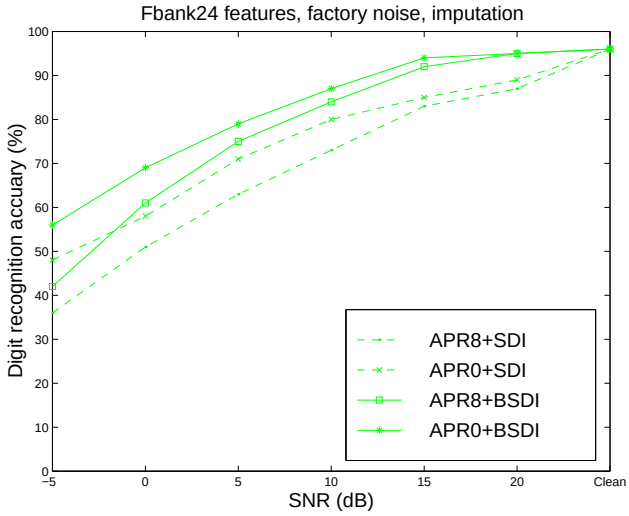


Figure 6.30: Data imputation with APR mask on factory noise (24–channel filterbank features)

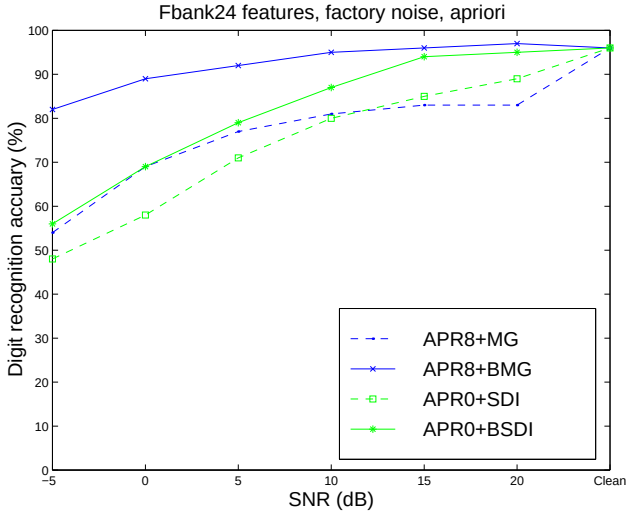


Figure 6.31: Bounded marginalisation and data imputation with APR mask on factory noise (24–channel filterbank features)

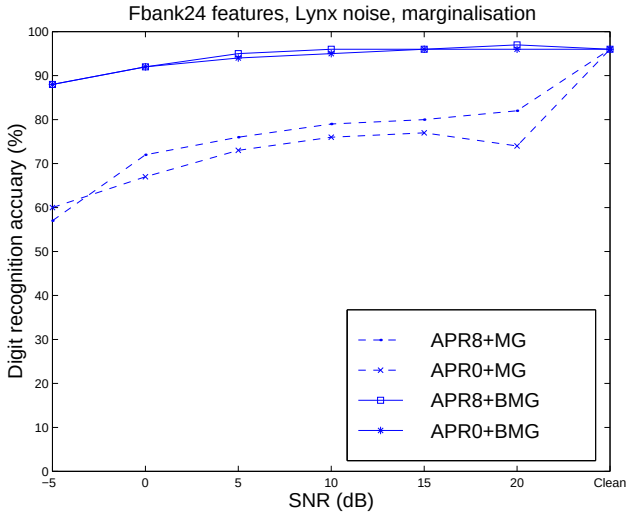


Figure 6.32: Marginalisation with APR mask on Lynx noise (24–channel filterbank features)

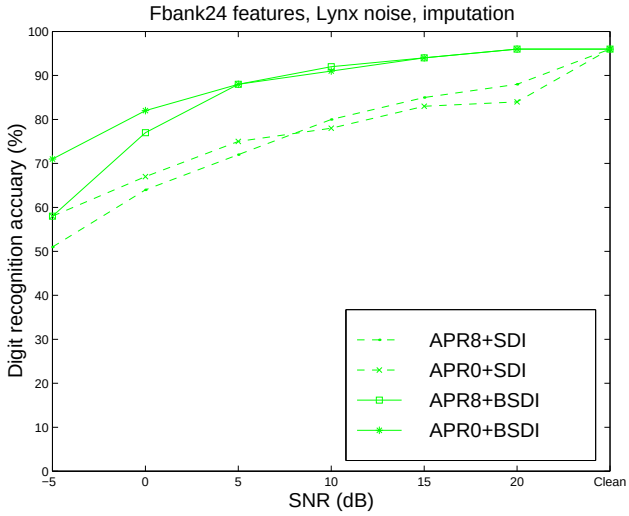


Figure 6.33: Data imputation with APR mask on Lynx noise (24–channel filterbank features)

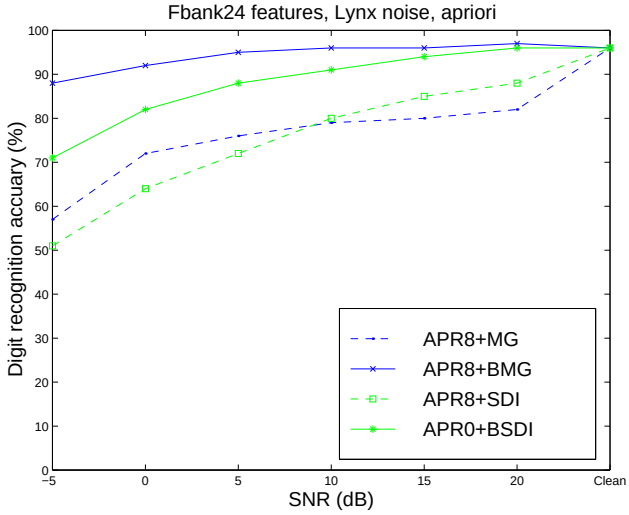


Figure 6.34: Bounded marginalisation and data imputation with APR mask on Lynx noise (24–channel filterbank features)

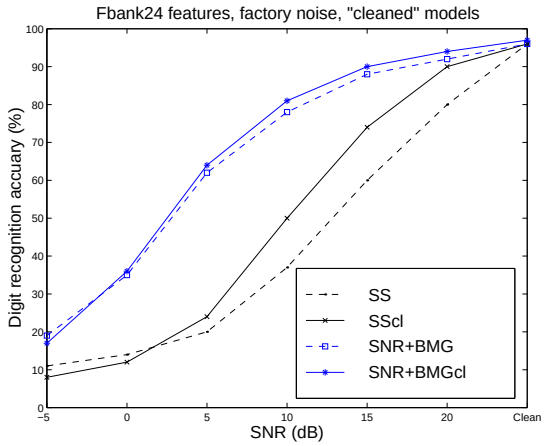


Figure 6.35: Marginalisation and spectral subtraction with "cleaned" models on factory noise (24–channel filterbank features)

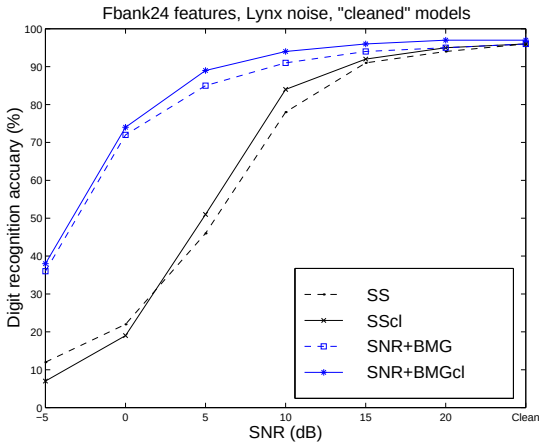


Figure 6.36: Marginalisation and spectral subtraction with "cleaned" models on Lynx noise (24–channel filterbank features)

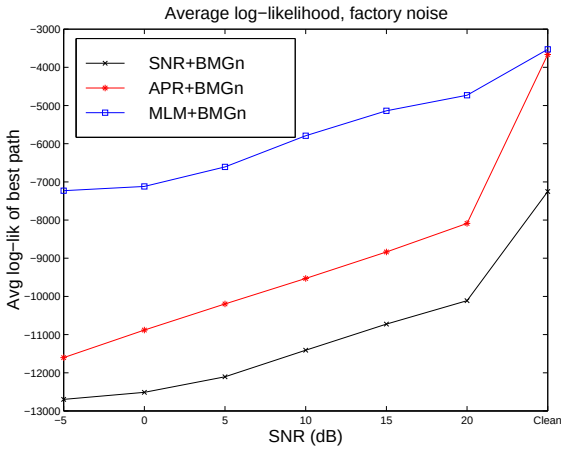


Figure 6.37: The average log–likelihood of the best path on factory noise (24–channel filterbank features)

without any additional cost carries the improvements over even when the MD techniques are used.

Average log-likelihood of the best path

Considering the large difference in accuracy when using APR masks compared to using "real", SNR masks, it was of interest to get an insight to the likelihood of the best path in both cases.

Figure 6.37 depicts the average log–likelihood (averaged over the 240 test sentences) of the best path (the ASR's best result) with three different masks. Along the Y–axis is the log–likelihood, while the SNR decreasing from clean speech, to 20 dB, to -5 dB in 5 dB steps is on the X–axis. Bounded marginalisation on the noisy data (BMGn), with no noise estimate subtracted from the noisy speech, was used. The contributions of the missing features to the likelihood were divided by the range $\mathbf{o}_m(t) - 0 = \mathbf{o}_m(t)$, to yield the "average likelihood" (see Section 5.5.7 and Figure 5.4).

APR and SNR masks are computed as before. The Maximum Likelihood Mask (MLM) is computed by comparing the contribution to the likelihood (of the the vector) by each feature: how much does a feature contribute when it is present, and how much when it is missing. The

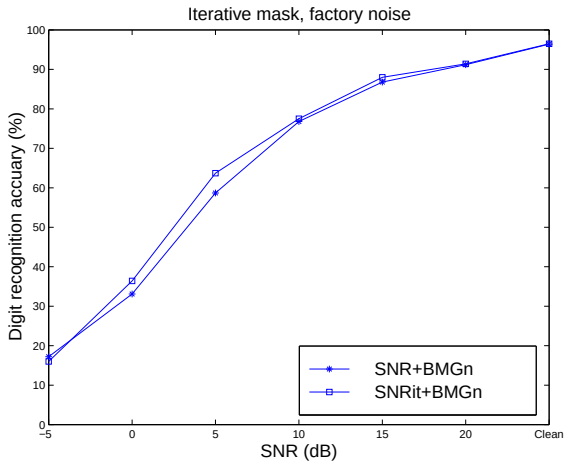


Figure 6.38: Accuracy with iterative mask refinement on factory noise (24–channel filterbank features)

larger of the two values is taken and that determines whether the feature is present or missing. This is done on a per–state basis, as each state has different p.d.f. used to compute the likelihood of the observation vector. After the recognition has finished, the best path is backtracked and finally determines the MLM mask used. In a sense, the MLM mask maximises the local acoustic evidence for each individual state.

The difference in accuracy between the APR and SNR masks carries over in the log–likelihood domain: the average log–likelihood of the best path with APR mask is considerably greater then with SNR mask at all SNRs. This raises the possibility of using the acoustic likelihoods as guides during mask creation (speech/noise separation). The average log–likelihood of the best path with MLM mask is indeed much larger then with SNR mask. But it is also larger then when APR masks are used. And the accuracy of the best path with MLM mask is much worse then with SNR (and APR) masks (not shown). Investigation of the MLM masks showed that although giving rise to best paths with high likelihoods, the masks themselves were very unlikely. The present and missing features in the mask were finely dispersed over the whole T–F plane, without any of the grouping apparent in the APR masks.

Using MLM mask effectively locally maximises the partial likelihood $P(O|M, Q^*, W)$ from Eq. (5.2), without taking into account the likelihood of the mask itself $P(M|Q^*, W)$ when computing the "best path"[3] W^*. If a suitable mask model $P(M|Q^*, W)$ penalised the very unlikely MLM mask, the accuracy of the best path with MLM would probably correlate to its log–likelihood, as is the case with SNR and APR masks.

Iterative mask refinement

Starting with a SNR mask, the mask was iteratively refined by cycling through recognition (alignment) and (most likely) mask reestimation. In each iteration the state alignment of the best path was obtained. Having one state corresponding to each frame, the p.d.f. of that particular state was used to infer the most likely mask for that frame alone. As in the previous section, the feature was considered to be present if its likelihood was greater then its "average likelihood" (see Section 5.5.7). Otherwise, it was considered missing. Once a new mask was obtained, the best path with that mask was computed. This path was used in the next iteration, etc. The iterative process was aborted if the likelihood of the best path did not increase sufficiently.

Figure 6.38 depicts the accuracy with this method (SNRit) used together with bounded

[3]the equation refers to isolated word recognition, but it generalises to the connected word recognition task (like the experiment above)

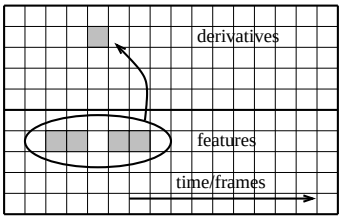


Figure 6.39: Computing the "strict" mask for the derivatives

marginalisation on noisy data (BMGn). The accuracy is compared to the one of using the initial SNR mask alone (SNR+BMGn). There is some improvement at mid SNRs, but it is questionable if it is enough to justify manifold increase in computational cost (compared to improvements achievable with other means).

6.3.5 Results with 24 channel filterbank features with their first derivatives

In this set of experiments, the 24–channel filterbank features were supplemented by their temporal derivatives approximations, yielding 48 feature vectors. The temporal derivatives were approximated with the "standard" expression (Furui, 1986):

$$\Delta x_i(t) = \frac{\sum_{j=-N}^{N} j \cdot x_i(t+j)}{\sum_{j=-N}^{N} j^2} \tag{6.9}$$

with $N = 2$.

The problem with MD ASR is that some of the $x_i(t+j)$ for $j = -N \ldots N$ features may be missing. One solution is to treat the derivative $\Delta x_i(t)$ as missing if any of the features $x_i(t+j)$, $j = -N \ldots N$ needed to compute $\Delta x_i(t)$ are missing (the *strict mask*), as depicted in Figure 6.39. If the missing mask pattern was random, this would create a very sparse mask for the derivatives. However, in the experiments with speech and noise this is not the case. The reliable features tend to be clustered into T–F blocks, so the sparsity of the derivative mask is not much greater than that of the features mask.

It was also noted that when strict masks were used with bounded marginalisation and data imputation, the bounds on the derivatives were so wide that they made little difference (at a great computational cost). Hence in all experiments (unless noted otherwise) the contribution of the missing derivatives to the likelihood were disregarded, effectively turning BMG into MG and BSDI into SDI as far as the derivatives were concerned.

Strict SNR and APR masks

Figures 6.40 and 6.41 depict the results with strict SNR masks (SNRst) on factory and Lynx noise respectively. Figures 6.42 and 6.43 depict the results with strict APR masks (APRst) on factory and Lynx noise respectively.

All figures contain results with spectral subtraction (SS) and Mel–cepstral features (MFCC), without or with Cepstral Mean Normalisation (MFCC+CMN) as well. For the latter 13 cepstral features were extracted from the 24 filterbank outputs via Discrete Cosine Transform (DCT) (Young and Woodland, 1993). For MFCC+CMN, the mean (on a per–sentence basis) of each feature was subtracted from it. Then, their first and second derivatives were appended to the feature vector yielding 39 features for each vector.

MD techniques tested were bounded marginalisation (BMG) and bounded state based imputation on noisy data (BSDIn). It was found that with first derivatives, the data imputation technique is very sensitive to the disturbances introduced to the derivatives due to the subtraction of the noise estimate. Hence the noisy data was used for imputation.

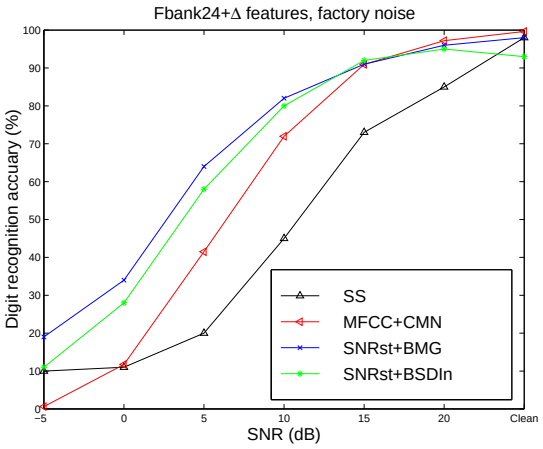


Figure 6.40: Bounded marginalisation and data imputation with SNRst mask on factory noise (24–channel filterbank features with first derivatives)

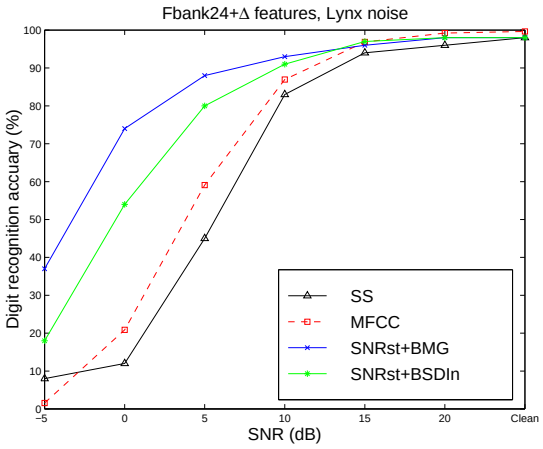


Figure 6.41: Bounded marginalisation and data imputation with SNRst mask on Lynx noise (24–channel filterbank features with first derivatives)

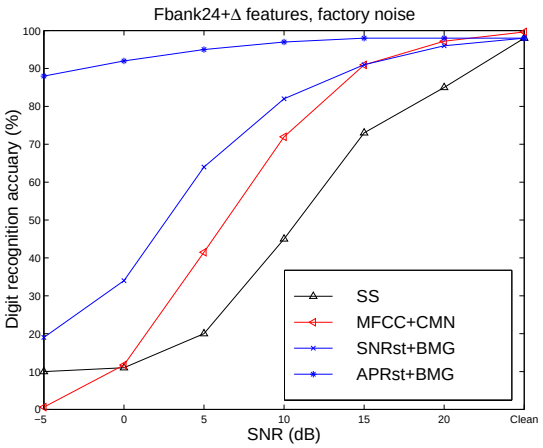


Figure 6.42: Bounded marginalisation with APRst mask on factory noise (24–channel filterbank features with first derivatives)

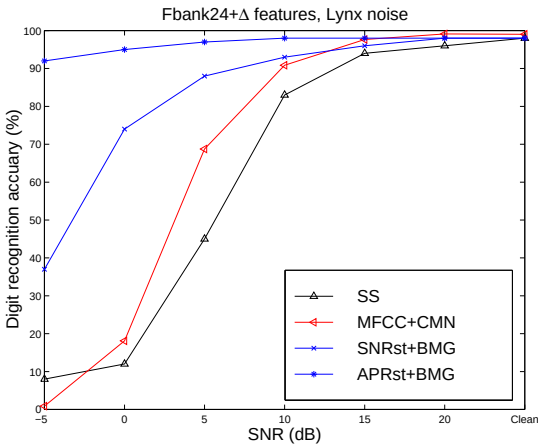


Figure 6.43: Bounded marginalisation with APRst mask on Lynx noise (24–channel filterbank features with first derivatives)

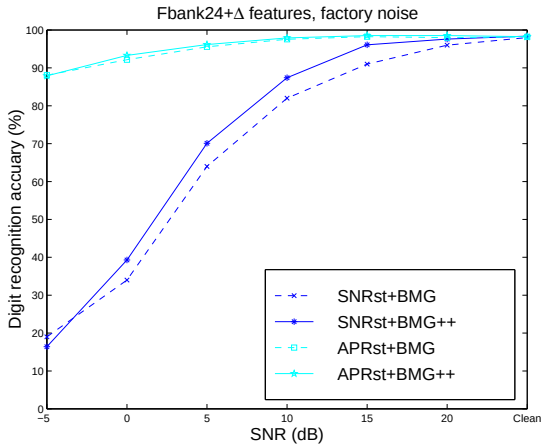


Figure 6.44: Bounded marginalisation with SNRst and APRst masks on factory noise with few small recogniser improvements (24–channel filterbank features with first derivatives)

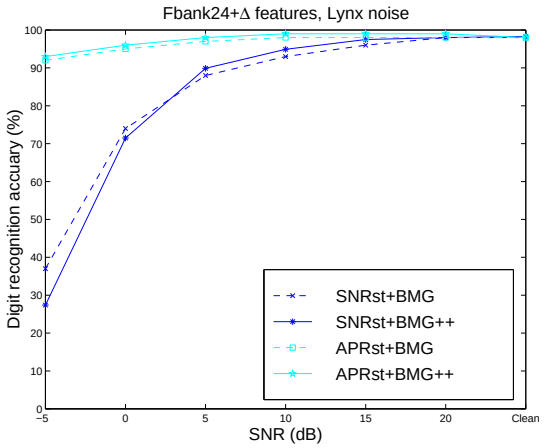


Figure 6.45: Bounded marginalisation with SNRst and APRst masks on Lynx noise with few small recogniser improvements (24–channel filterbank features with first derivatives)

With both noises both MD techniques outperform both the spectral subtraction and the cepstral features (with and without mean normalisation) at almost all SNRs. The only exception is BSDI performing worse then SS and MFCC at high SNRs (clean and 20 dB) on factory noise. Again, BMG performs better then BSDI at all SNRs and on both noises.

Therefore, with APRst masks (Figures 6.42 and 6.43) only the results with bounded marginalisation are shown. The apriori masks are not "true" masks, as they are derived by knowing the clean (in addition to the noisy) speech. However, they are indicative of what can be achieved with very good speech/noise separation. The results are very tempting: accuracy of around 90% (a bit more for Lynx, a but less for factory noise) at -5 dB. It seems that these results point firmly to poor mask quality of SNRst masks as the major reason for poor performance (compared to the usage of APRst masks). They also may indicate that major improvements are unlikely to come from using some new technique for estimation of the partial likelihood. Rather, major improvements maybe expected with better masks estimation.

"Common" ASR system tuning techniques

Figures 6.44 and 6.45 show the effects on using some common ASR tuning techniques on the recognition accuracy on factory and Lynx helicopter noise.

The baseline MD ASR system (BMG) was tested with strict SNR (SNRst) and apriori (APRst) masks. The improved system (BMG++) consisted of the following:

- tuned word insertion penalty
- additional, short interword silence
- rudimentary language modelling – it was noted that no sentence contained both the "zero" and "oh" models, hence transcriptions mixing both in the same sentence were rejected during testing

Figure 6.44 shows that on factory noise there is small but notable improvement, which is more pronounced with SNRst mask then with APRst mask. On Lynx helicopter noise (Figure 6.45), the results with SNRst mask are mixed, and the improvement with APRst mask is smaller.

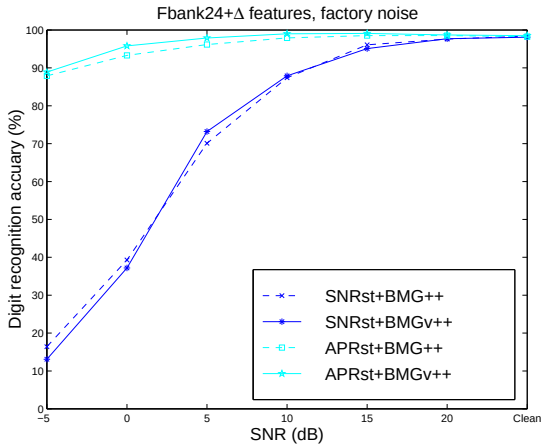


Figure 6.46: Bounded marginalisation with and without bounds on the derivatives with SNRst and APRst masks on factory noise (24–channel filterbank features with first derivatives)

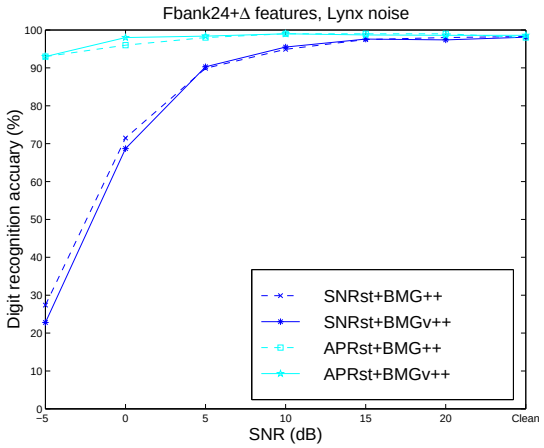


Figure 6.47: Bounded marginalisation with and without bounds on the derivatives with SNRst and APRst masks on Lynx noise (24–channel filterbank features with first derivatives)

Bounding the missing derivatives

Figure 6.46 shows the effects of using bounds on the missing derivatives in the strict mask on factory noise, while Figure 6.47 depicts the same for Lynx helicopter noise. In both cases the "improved system" from the previous section is the baseline.

The bounds on the missing derivatives were computed by knowing the bounds on the missing "static" features (used to compute the derivative) and the way the "static" features are used to compute the derivative (Eq. (6.9)).

The results for both noises are consistent. If the mask is accurate (e.g. APRst mask), then bounding the derivatives does help at lower SNRs (only slightly, but the baseline is already quite high). Whereas when the mask is not accurate (e.g. SNRst mask), the benefit of using bounds is wiped out by them being not accurate in significant number of cases. Therefore, there is little incentive in using the bounds on he derivatives when the bounds are unreliable.

"Standard" techniques for improving ASR system's robustness

As an indication of what is possible with some "standard" robustness technique, Figures 6.48 and 6.49 depict the accuracy with Mel cepstral features (MFCC) with and without Cepstral mean normalisation (subtraction) (CMN) on factory and Lynx helicopter noise respectively. The results with spectral subtraction on 24–channel filterbanks with their first derivatives are also shown. These two techniques were chosen as they are the most representative of the currently used techniques for improving the robustness of an ASR system.

The system was exactly the same, and trained and tested on exactly the same data as the one using the MD techniques.

It seems that MFCC features are inherently more robust and they outperform SS both with and without CMN on both noises.

6.4 Experiments on the Aurora 2 database

As described in Section 6.2, the Aurora 2 database (Hirsch and Pearce, 2000) is a noisy and downsampled version of the TIdigits database (Leonard, 1984).

For this set of experiments, 32–channel ratemaps (Cooke, 1991) were used as features in the MD ASR system. After the experiments with 64–channel ratemaps and 24–channel filterbanks (in

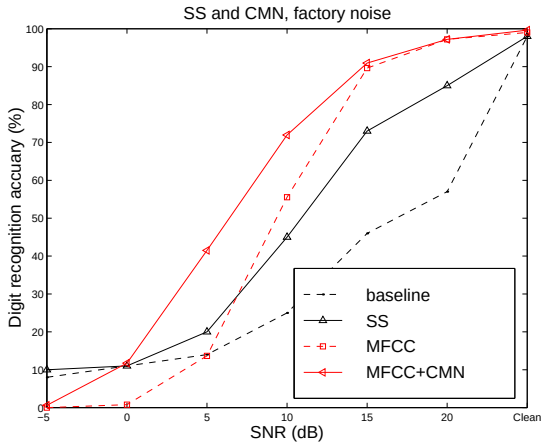


Figure 6.48: MFCC features with and without CMN, 24–channel filterbank features with first derivatives with SS on factory noise

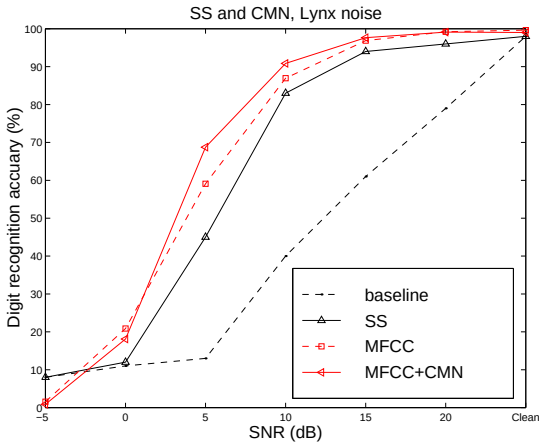


Figure 6.49: MFCC features with and without CMN, 24–channel filterbank features with first derivatives with SS on Lynx noise

the previous sections), the features and their number was chosen as a compromise between two opposing aims that frequency domain based features need to provide to an MD ASR system:

- fine enough frequency resolution for the purpose of mask estimation
- coarse enough frequency resolution so that speaker/pitch dependent harmonics are not resolved (but only the envelope of the short–term spectrum is sampled)

Ratemaps also provide some smoothing (that Mel–scaled filterbanks lack) making the assumption that speech and noise are additive in power spectral domain (used to derive the estimated SNR after the noise estimation) more viable. Their first order derivatives (Eq. (6.9)) were appended to the ratemaps to form a 64–element feature vector. An appropriate strict mask was used for the derivatives.

Previously used methods for mask estimation assumed a mask with probability of unity, i.e. no attempt was made to estimate the factor $P(M|Q^*, W)$ from Eq. (5.2) when estimating the mask. The probability of all possible masks other then the estimated one was set to 0. Hence the sum For the experiments on Aurora 2, two methods that associate certain probability with each mask were tested. The following two sections will outline the techniques employed.

6.4.1 Soft/fuzzy SNR mask (SNRSoft)

With the soft (fuzzy) SNR masks, the probability of each point being speech was expressed by suitable mapping of the estimated SNR. Hence the term soft/fuzzy – instead of thresholding the SNR (as with the SNR masks), each point in the mask gets an associated probability. The mapping was accomplished via a sigmoid function:

$$P(m_i(t) = 1) = \frac{1}{1 + e^{-\alpha(\hat{SNR}_i(t) - \beta)}} \tag{6.10}$$

The local SNR estimate $\hat{SNR}$ was computed as for the SNR masks. The centre β and the slope α of the sigmoid were chosen empirically to achieve best result on the Aurora 2 testa, noise1 subset (Subway noise). The same parameters were used across both test subsets (testa and testb) and across all noises.

6.4.2 Adaptive noise tracking (SNRA)

The adaptive noise tracking scheme consists of tracking both the noise mean and variance. It is assumed that the distribution of the noise is Normal, and that the noise features are independent. The first 10 frames of each utterance were used for initial estimates of the noise mean and variance. Then, each feature $o_i(t)$ in the incoming frame $\mathbf{o}(t)$ was scored on how likely is that it was generated from the noise distribution estimated in that channel so far. This was accomplished by thresholding the probability of an SNR being less then a SNR threshold. In the experiments reported in the next section, if $P(\hat{SNR}_i(t) < -7dB) > 0.6$ the feature $o_i(t)$ was considered to have been generated by the noise source. Therefore it was used to recursively adapt the noise mean and variance estimates. The $P(\hat{SNR}_i(t) < X)$ with threshold X in dB was computed as

$$P(\hat{SNR}_i(t) < X) = 0.5 - 0.5 \cdot erf\left(\frac{\frac{o_i(t)}{1+10^{X/20}} - \mu_i(t)}{\sqrt{2\Sigma_i(t)}}\right) \tag{6.11}$$

for the spectral data. The recursive update of the mean $\mu_i(t)$ and the variance $\Sigma_i(t)$ (independently for each feature/channel) was:

$$\begin{aligned} \mu_i(t+1) &= \alpha\mu_i(t) + (1-\alpha)o_i \\ s_i(t+1) &= \alpha s_i(t) + (1-\alpha)o_i^2 \\ \Sigma_i(t+1) &= s_i(t+1) - \mu_i(t+1) \end{aligned} \tag{6.12}$$

with $\alpha = 0.995$.

Once a noise estimate was obtained, an SNR estimate was obtained by assuming that the speech and the noise are additive in the spectral magnitude domain (resulting in the noisy speech magnitude) at all times.

The probability of belonging to the speech/noise source was subsequently computed from the local SNR. For clean models the SNR threshold was fixed at 7 dB, i.e. the probability of each point in the T–F plane being speech was computed as:

$$P(m_i(t) = 1) = P(\hat{SNR}_i(t) > 7dB) \tag{6.13}$$

This results in a mask that has very few, but reliable points. For noisy models it was found that it is preferable to use a lower threshold of 0 dB instead of 7 dB. This less stringent assessment of the speech quality results in more points considered more likely to be speech.

This probability can be used in the "soft" MD computation instead of the sigmoid from the previous subsection. The advantage of SNRA seems to be that while the centre β and the slope α of the sigmoid (mapping the SNR estimate into a reliability estimate) are somewhat noise dependent, we haven't observed the same with the threshold X and the "forgetting factor" α of SNRA. The disadvantage is the computational cost. Assuming that computing $P(\hat{SNR}_i(t) < X)$ is comparable to sigmoid evaluation, the computational cost of SNRA is at least twice the one of the sigmoid mapping.

If a discrete mask is needed, the probability $P(m_i(t) = 1) = P(\hat{SNR}_i(t) > XdB)$ is simply thresholded to provide binary speech/noise discrimination.

6.4.3 Computing the state likelihood with fuzzy masks

Since each point in the mask is treated being independent of the rest, and the state p.d.f.s are sums of factorisable p.d.f.s, Eq. (5.8) can be used to efficiently calculate the sum over all possible masks in Eq. (5.2):

$$\begin{aligned} P(\mathbf{o}(t)|q(t),W) = \sum_k P(k) \prod_i [&p_i(o_i(t)|k, m_i(t)=0, q^*(t), W) p(m_i(t)=0|q^*(t), W) \\ &+ p_i(o_i(t)|k, m_i(t)=1, q^*(t), W) p(m_i(t)=0|q^*(t), W)] \end{aligned} \tag{6.14}$$

The contribution of the present and missing features to the likelihood is weighted by the probability of them being present or missing, which is provided by the fuzzy mask.

6.4.4 Results with discrete and fuzzy strict SNR masks

Figures 6.50, 6.51, 6.52 and 6.53 depict the results on the four noises from Aurora 2 testa test set. Figures 6.54, 6.55, 6.56 and 6.57 depict the results on the four noises from Aurora 2 testb test set.

The "clean baseline" results are the baseline results with MFCC features and models trained on the clean portion of the database. It is set by the rules of the Aurora 2 competition. Similarly, the "multi baseline" results are the baseline results with MFCC features and models trained on the noisy portion of the database.

The bounded marginalisation (BMG) was tested with strict discrete SNR masks (SNRst) as well as with strict fuzzy SNR masks (SNRstSoft). In both cases it outperforms the "clean baseline" in all conditions. Estimating the mask probability, even crudely as with SNRstSoft, considerably improves the BMG results. In all cases the "multiconditional baseline" (models trained on noisy speech) performs best.

Figure 6.50 contains an additional result with discrete apriori mask (APRst). It is interesting to compare that to the "multi baseline" result, as:

- using data contaminated with noise for training is often considered the upper limit on what can be achieved with various noise robustness techniques (e.g. various models adaptation techniques)
- using apriori masks can be indicative of the upper limit of what can be achieved with the MD approach to robust ASR

Both methods provide comparable performance at high SNRs. But for mid and low SNRs, the "multiconditional training" rapidly deteriorates, while BMG with APRst masks steadily holds onto the performance.

6.4.5 Results with adaptive noise tracking

Figures 6.58, 6.59, 6.60 and 6.61 depict the results on the four noises from Aurora 2 testa test set. Figures 6.62, 6.63, 6.64 and 6.65 depict the results on the four noises from Aurora 2 testb test set.

As before, the "clean baseline" results are the baseline results with MFCC features and models trained on the clean portion of the database, and fixed the rules of the Aurora 2 competition. Similarly, the "multi baseline" results are the baseline results with MFCC features and models trained on the noisy portion of the database.

The models used for the experiments were trained on the noisy portion of the Aurora 2 database. The adaptive noise estimation had a threshold of 0 dB for assessing the probability of the mask.

Bounded marginalisation (BMG) was tested with strict fuzzy SNRA masks (SNRAstSoft) and strict discrete APR masks (APRst). Using SNRAstSoft masks and noisy models makes the MD results comparable with the multiconditional baseline. Bounded marginalisation in this case performs the same with the multiconditional baseline at high SNRs, slightly better for some noises and worse for others on mid SNRs, and always significantly better at low SNRs.

The clean baseline and marginalisation with apriori masks are also plotted for indication. Bounded marginalisation with apriori masks (APRst) performs surprisingly well even at the lowest SNR, where very few reliable speech points are available.

The main advantage of the fuzzy SNRA masks over discrete SNR masks seems to be the fuzziness, rather then the adaptation. SNRA masks, when used as discrete masks (by thresholding the probability estimate) behave only marginally better then discrete SNR masks.

6.4.6 Token dependent noise estimation

Figure 6.66 depicts the results with "token dependent noise estimate" (TDNE) on the Aurora 2 testa noise1 subset (Subway noise). 24–Channel filterbank features together with their first derivatives were used.

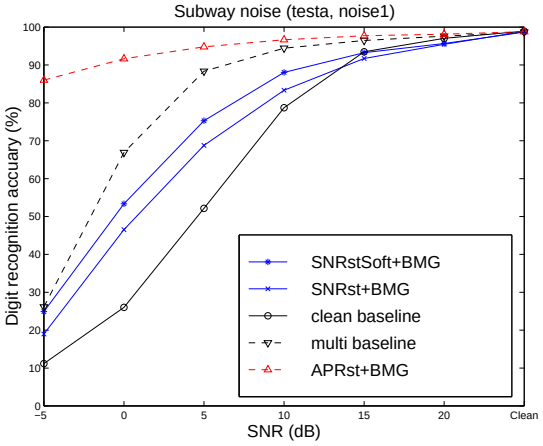


Figure 6.50: Bounded marginalisation with discrete SNRst, fuzzy SNRstSoft and discrete apriori APR masks on the Aurora 2 Subway noise (testa, N1).

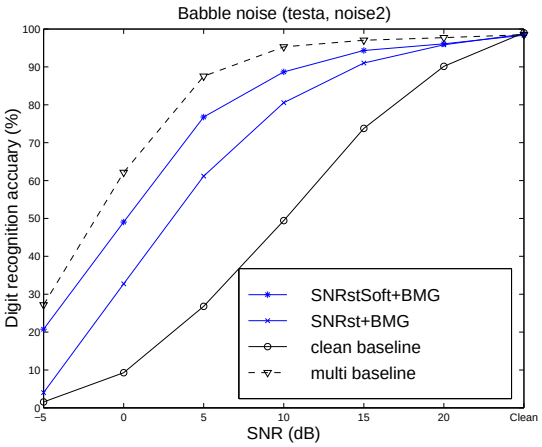


Figure 6.51: Bounded marginalisation with discrete SNRst and fuzzy SNRstSoft masks on the Aurora 2 Babble noise (testa, N2).

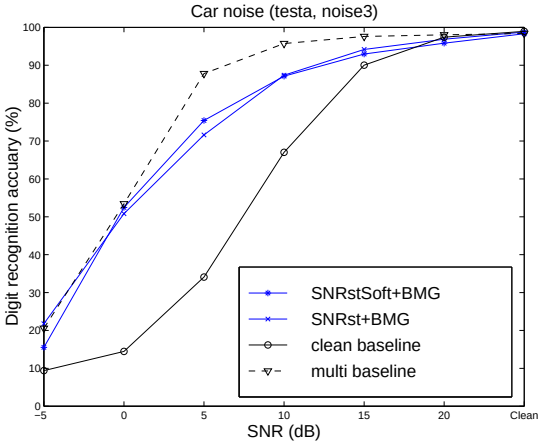


Figure 6.52: Bounded marginalisation with discrete SNRst and fuzzy SNRstSoft masks on the Aurora 2 Car noise (testa, N3).

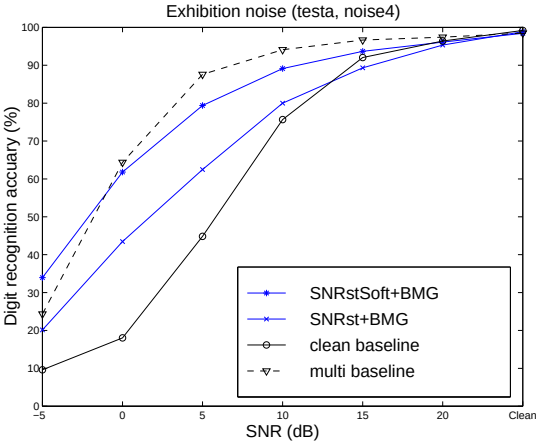


Figure 6.53: Bounded marginalisation with discrete SNRst and fuzzy SNRstSoft masks on the Aurora 2 Exhibition noise (testa, N4).

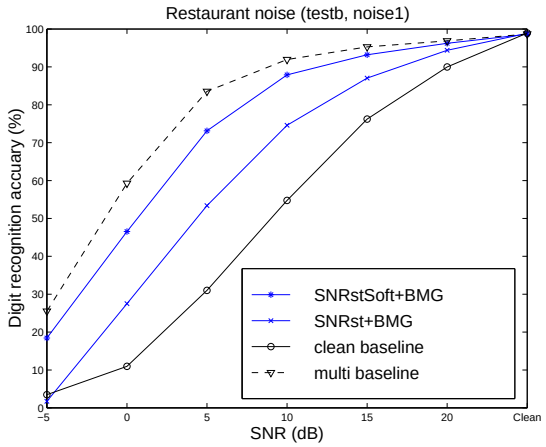


Figure 6.54: Bounded marginalisation with discrete SNRst and fuzzy SNRstSoft masks on the Aurora 2 Restaurant noise (testb, N1).

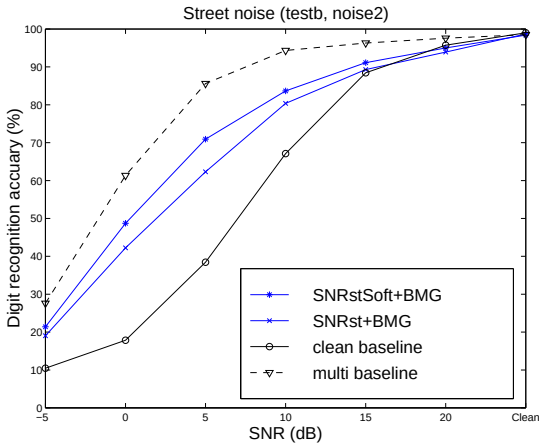


Figure 6.55: Bounded marginalisation with discrete SNRst and fuzzy SNRstSoft masks on the Aurora 2 Street noise (testb, N2).

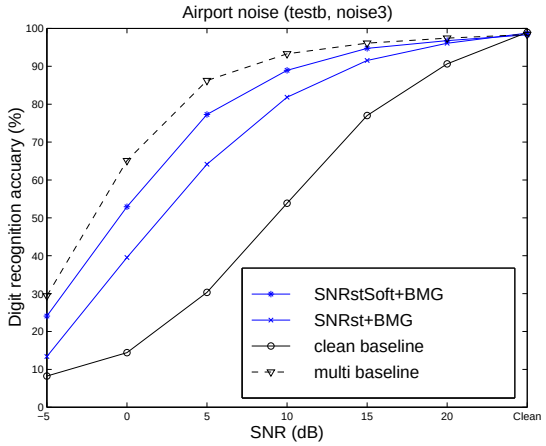


Figure 6.56: Bounded marginalisation with discrete SNRst and fuzzy SNRstSoft masks on the Aurora 2 Airport noise (testb, N3).

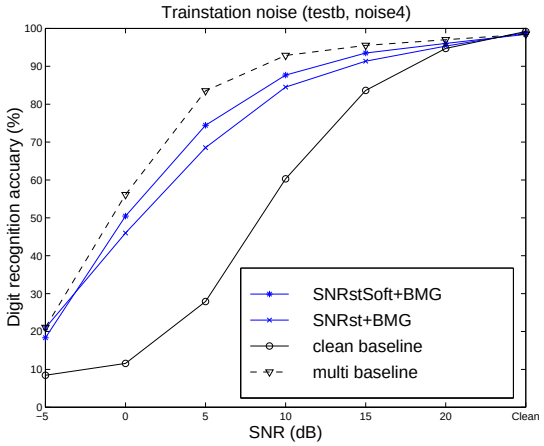


Figure 6.57: Bounded marginalisation with discrete SNRst and fuzzy SNRstSoft masks on the Aurora 2 Train station noise (testb, N4).

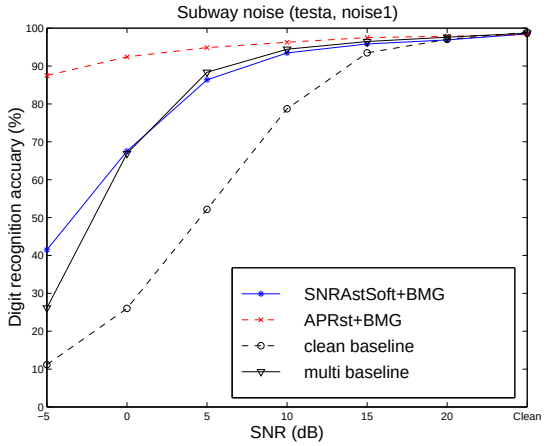


Figure 6.58: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Subway noise (testa, N1).

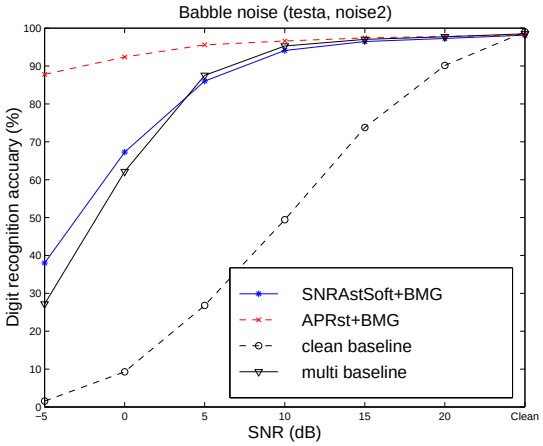


Figure 6.59: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Babble noise (testa, N2).

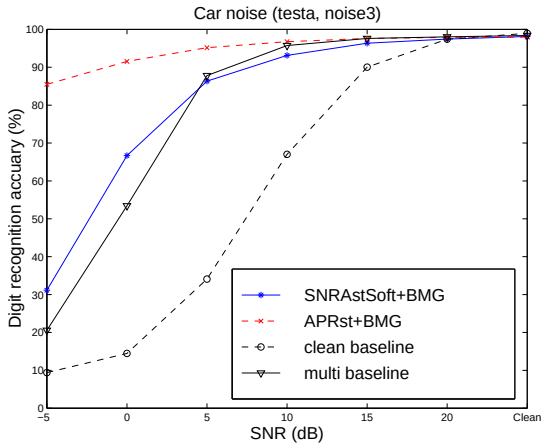


Figure 6.60: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Car noise (testa, N3).

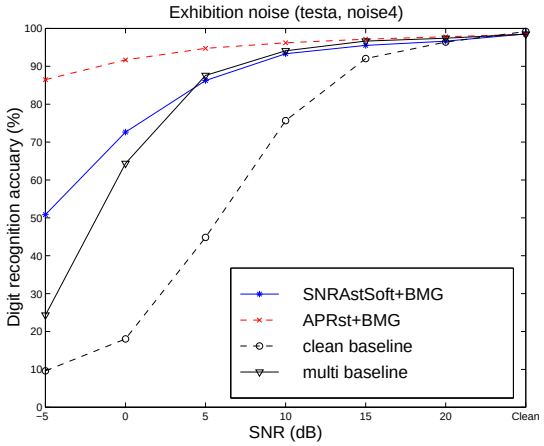


Figure 6.61: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Exhibition noise (testa, N4).

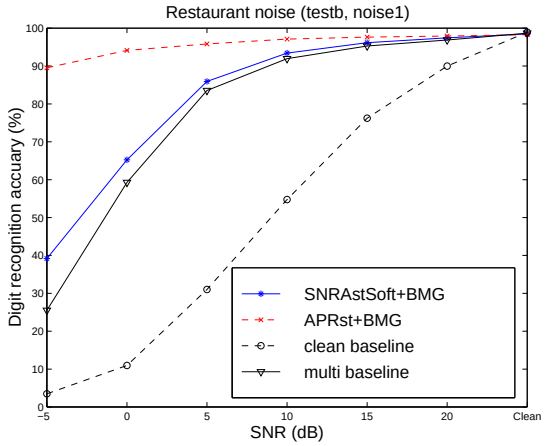


Figure 6.62: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Restaurant noise (testb, N1).

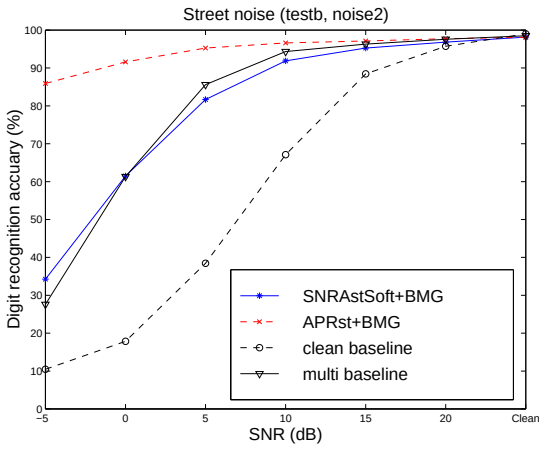


Figure 6.63: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Street noise (testb, N2).

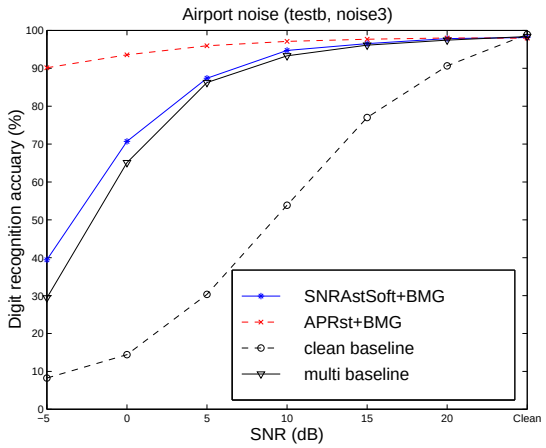


Figure 6.64: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Airport noise (testb, N3).

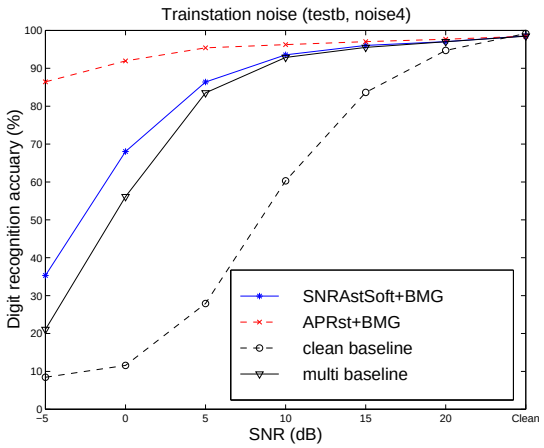


Figure 6.65: Bounded marginalisation with fuzzy SNRAstSoft and apriori discrete APRst masks on the Aurora 2 Train station noise (testb, N4).

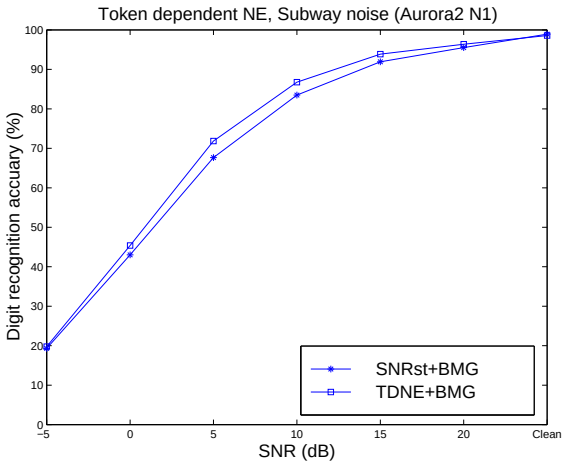


Figure 6.66: Bounded marginalisation with token dependent noise estimation SNRst mask on the Aurora 2 Subway noise (24–channel filterbank with the first derivatives).

The TDNE scheme is connected to the "token passing" scheme (Young et al., 1989) for performing dynamic Viterbi search during ASR. Briefly, the scheme assumes an imagined "token" exists in every state of the speech models. Each token carries its path and the likelihood of the data it has "seen" so far. Every time frame, the token from every state is propagated into all admissible subsequent states and has its path and likelihood score updated. Next, for every state, only the token with the highest score is kept while the rest are discarded. The process is repeated until there is data available. Tokens in the "last" states of every model are compared, and the most likely among them is the winner.

In TDNE, each token carries a noise estimate in addition to the other data (path, likelihood score). Tokens propagated through silence model(s) have their noise estimates updated. When computing the frame likelihood, each token, having separate noise estimate, gets different mask (and maybe speech) estimate. The scheme was employed with the standard bounded marginalisation technique. For the dynamic features the static mask was used as well, as computing the strict mask as before requires knowledge not only of the static masks in the past frames, but the future frames as well, which are not available.[4]

As shown on Figure 6.66, using SNRst mask derived with TDNE does improve the accuracy over the standard SNR mask (which uses stationary noise estimate) at all but the highest and the lowest SNRs.

6.5 Summary of the experimental results

Two series of experiments were carried out on two noisy versions of the same data (TIdigits). The MD techniques were used in conjunction with masks provided by separation that relies on noise estimation. The MD techniques themselves do not call for noise estimation. However, at the time the experiments were done, no functioning CASA system was available for masks creation.

In the first series, the data was contaminated with two noises: factory (non–stationary) and Lynx helicopter (stationary). The main conclusions are:

- Both BMG and BSDI perform better then SS with the same noise estimate.
- MG and SDI both suffer from random insertions when there is little data present in a frame. SDI is more suspect to this problem, as it has to recover the missing data from the present

[4]this can be overcome by shifting the derivatives N frames back, i.e. using $\Delta x_i(t) = \sum_{j=-N}^{N} j \cdot x_i(t+j-N) / \sum_{j=-N}^{N} j^2$ instead of Eq. (6.9)

data and the speech models. Using bounds (BMG and BSDI) solves that problem.

- Marginalisation (MG and BMG) is more accurate then imputation (SDI and BSDI). The data can be imputed after it is Viterbi aligned with the models using BMG.
- Using cleaned models gives only slight improvement (both for MD and SS).
- Using standard deltas with strict mask provides improvement; the missing deltas need not be bounded, but simply marginalised.
- Using soft masks brings big improvement.

In the second series of experiments, the Aurora 2 noisy database was used. Several points can be made regarding the Aurora 2 graphs:

- MD with noisy models and adaptive noise tracking performs as good as or better then the baseline with the noisy models - same or slightly worse at high SNRs and better at low SNRs. It should be noted that the filterbank models used in MD have baseline worse then the Aurora 2 MFCC baseline. So the adaptive noise estimation with MD recovers some of that loss.
- MD with clean models and with stationary noise estimate performs significantly better then the clean MFCCs based baseline. This is expected as the baseline doesn't take into account the noise at all, neither during training nor during testing.
- MD with apriori masks shows the potential of the technique with an extremely good/perfect mask. The results are surprisingly good even at low SNRs and the accuracy doesn't plunge catastrophically.
- The apriori masks results clearly point that the deficiencies in the MD approach at present are not of poor speech modelling, but of poor identification of the speech components in the noisy speech. Therefore the huge performance gap between the realistic mask estimates and the apriori mask at low SNRs can be addressed by improving the identification of the speech in noise and therefore the mask.[5]
- Adaptive noise estimation and similar techniques for noise tracking are somewhat more effective then stationary noise estimation. However, not only do they introduce more complex processing (their computational cost is negligible compared to the rest of the system), but also rely on tunable parameters which need to be optimised by trial and error.

6.6 Summary

Results of the recognition experiments using an MD HMM system were presented in this chapter. A "standard"/textbook HMM system, adapted suitably to handle the MD, was employed on a connected digits task. The speech was artificially contaminated with noise at SNRs from -5 dB to 20 dB. Clean speech was also used for testing. MD ASR was tried with different features (although all constrained to be frequency domain features) and with and without their first derivatives. For mask estimation, couple of techniques based on local SNR estimation were employed. They rely on noise estimation. We mostly made use of simplest forms of noise estimation, and with purpose. MD techniques don't need noise estimation if there is another process (see Chapter 4) that can produce the mask. However, in absence of widely available, computationally cheap and easily reproducible methods for speech/noise identification (separation) we had to make use of noise

[5]This is consistent with the results with and without voice activity detection, and in general with the approaches taken in most of the noise reduction algorithms. These algorithms are knowledge driven and rely on the gross speech features (harmonics, continuity, simultaneous transitions, onsets and offsets across frequency bands) which have already been singled out in the (C)ASA community as features significant for speech separation (or identification of the speech portions in a noise mixture).

estimation in order to do the experiments. Further, we saw how adding even a crude probabilistic treatment to the mask brings improvements over a simple assumption that the particular mask used is the only possible one. Both marginalisation and data imputation were tested as methods for computing the likelihood of the partial data vectors. Toward the end (i.e. for the experiments on the Aurora 2 database) only marginalisation was used as it was always computationally cheaper and usually more accurate then data imputation. The results on the Aurora 2 database are directly comparable to other published results on the same, standardised database of noisy data. We think that even with the simple mask estimation techniques employed here, the MD ASR system does perform competitively. Further, the results with apriori masks point that improving the separation part in the MD chain is bound to bring most benefit to improving the accuracy.

Chapter 7

Discussion

7.1 Introduction

In this chapter the relation of MD techniques with other approaches to robust ASR which touch either or both the separation and the recognition parts of a robust system will be discussed first. Some Frequently Asked Questions about MD techniques will be answered next. Several possibilities open for further research, the unresolved problems and unknowns will also get a mention. The chapter (and the thesis) closes with the main conclusions that we have drawn from this work.

7.2 Relation to other approaches to robust ASR

The missing data techniques discussed throughout this text are related to different extents to several known techniques. Some take the "missing data" idea further (e.g. multisource decoding), others have arrived at the same techniques spurred by different motivation (e.g. masking), and still others have utilised all-or-nothing separation motivated by nothing more then sound signal processing principles (e.g. MAX approximation of the compressive nonlinearity) and were attracted by the simplicity that makes the computations trackable (HMM decomposition).

7.2.1 Multisource decoder by Barker, Cooke, and Ellis (2000, 2001a)

The multisource decoder (Barker et al., 2000, 2001a) takes MD techniques a step further, insofar that it relaxes the requirements on the separation frontend. Not only does it allow for arbitrary groupings of features in the time frequency (T–F) plane, but it doesn't need to label which ones are speech, and which ones are noise. It allows for the both possibilities, forking two decoders when it encounters a start of a new patch (a group of T–F points coming from the same source). Both decoders continue to decode in parallel the same patch of data. One of the forked decoders assumes that the patch belonged to the noise source and decodes using the BMG technique (Section 6.3.2). The other decoder assumes that the patch was generated by the speech source. All points that do not belong to that patch are treated the same by both decoders. At the end of the patch, the decoders are merged. For each state of every model the higher of the two scores (resulting from the data assumed to be speech and noise) is kept. If during the decoding of the patch the start of a new patch is encountered, each of the decoders forks two new ones, etc.

This amounts to a simultaneous search both for the ML path Q^* and the ML mask M^* (from Chapter 5). In (Barker et al., 2000) the mask is initially based on local SNR estimate (Section 6.3.1). Then it is subsequently broken in coherent fragments – patches – such that:

- the neighbouring present points belong to the same patch
- the patches are broken at the edges of the four arbitrarily chosen frequency bands

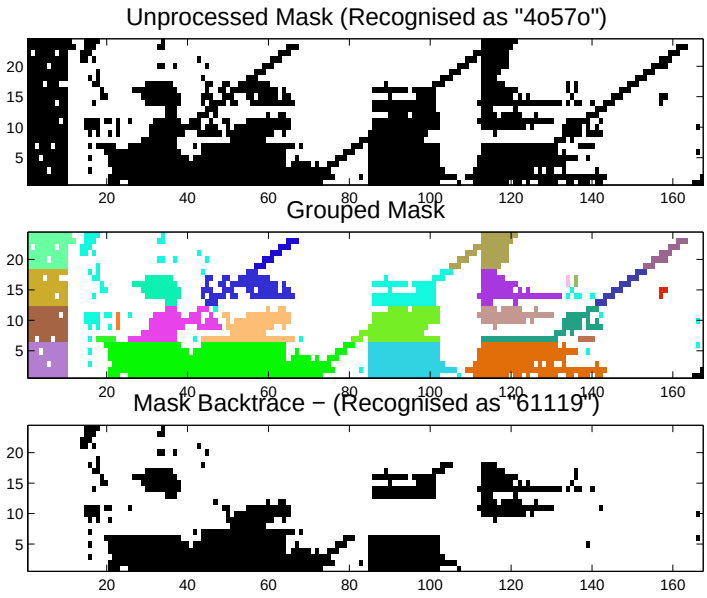


Figure 7.1: Local SNR derived mask (top); The mask decomposed into coloured coherent fragments – whole patches containing only speech or only noise (middle); The mask giving the ML word sequence after the decoding (bottom). The figure was kindly supplied by Jon Barker.

This a precursor to "schema driven grouping". Barker et al. (2001a) make a step in that direction: a voicing detector is used to split the patches further into parts containing harmonic energy and the others containing in–harmonic energy. Further, an adaptive noise estimate which assigns probability to each point is used to create the initial "soft" SNR masks (Section 6.4.1).

It is worth stressing again that the grouping process that creates the patches does not label them as speech or noise. After the decoding of the data has finished, the best path is backtracked and only then the mask giving rise to that path is established. The grouping is completed together with the decoding.

At present no patches overlap (each point in the T–F plane belongs to one patch only), and the "grouping process" does not associate probability estimates with the patches ("how likely it is that a group of points in the T–F plane were generated by a same source").[1]

Figure 7.1 depicts the original mask on the top panel. It was derived with stationary noise estimation followed by local SNR estimation and thresholding (Section 6.3.1). The noise is artificial diagonal chirps. The mask is broken into patches shown in the same colour on the middle panel. Each patch is either speech or noise exclusively. During the decoding, the decoders examine both hypotheses, and only the more likely is retained when the decoders merge at the end of a patch. After the decoding, the patches belonging to the recognised sequence are backtracked and shown on the bottom panel. The multisource decoding improves the accuracy of the MD ASR with non–stationary noise and at low SNRs (Barker et al., 2001a). The improvement seems proportional to the non–stationarity of the noise.

The four–bands break-up of the initial mask was chosen empirically. It seems that making the fragments smaller[2] hinders the performance.[3] This is reminiscent of MLM (previous chapter, pp. 101): choosing the feature to be present or missing solely to maximise local state likelihood alone leads to highly likely (Q^*, W^*) (higher then the Q^* of the correct model W) but poor accu-

[1] although the "soft" mask assigns some probability to whether the point is speech or noise, this is unrelated to the bottom–up (BU) constraints suspected of restricting which points can locally be grouped together as coming from the same source

[2] ex: dividing the mask in tiles instead of whole bands

[3] J. Barker, personal communication

racy. The reason may be that although the resulting mask M' gives rise to high $P(O|M', Q^*, W)$, the probability of the mask itself $P(M'|Q^*, W)$ is very low. A particular collection of points may give rise to high likelihood (of some model), but it is unlikely that the points in the collection alone were generated by a single source.

Developing models that will capture the BU constraints of which points can be grouped together seems to promise improvement of the ASR performance in non–stationary noise.

7.2.2 Multistream and multiband approaches to ASR

The multistream approach to ASR is based on combining several streams of evidence (Bourlard et al., 1996). The streams of evidence can be recombined at certain points, synchronously (the easier case) or asynchronously (harder, and with tenuous advantage over the synchronous case). With this architecture, it's easy to envisage integration of evidence coming from different types of features on different time scales, and/or from different sources and/or modalities. For example, Dupont and Bourlard (1997) combined evidence from short (10ms, phoneme) and longer (200ms, syllable level) term features in a context of a hybrid system.

A special case of the multistream approach is the multiband approach (Bourlard et al., 1996; Hagen et al., 1998; Cerisara et al., 1998; McCourt et al., 1998; Sarikaya and Gowdy, 1998; Hagen et al., 2000). The separate streams are (critical) bands of a filterbank. All acoustic processing is performed independently in each subband. This results in as many streams of features as there are subbands. They or their scores can be further combined in a final classifier. For example, Tibrewala and Hermansky (1998) extracted all pole PLP features independently (across bands) from the filterbank energies of seven separate bands (each spanning across two critical bands), and then recombined the (context independent) phone posteriors synchronously via a recombination network. Okawa et al. (1998) merged all the features in a single vector and classified using that huge feature vector alone.

There are several reasons in favour of the multistream approach:

- Allen (1994)'s review of the work of Fletcher about the intelligibility of low and high–pass filtered speech seems to suggest that humans process speech in more or less independent bands. The so called "product of error" rule (a conjecture that probability of human hearing making error is a product of the error rates in the individual bands) can be accounted for by this model.

- Per–band feature modelling (typically all–pole) should give more accurate models simply due to less variation present when each band is treated in isolation.

- Different recognition strategies may be more effective in different bands. For example, different windows and time/frequency trade–off for different bands.

- Allowing asynchrony between the bands may lessen some of the constraints of the current models. Mirghafori and Morgan (1998) tested the assumption that transitions between the sounds in the natural speech occur asynchronously across the bands. The findings seem to suggest that a significant number of transitions (about one third) occur more then 50ms apart from each other in different bands, with high frequency bands timings spread dependent on the speaking rate.

- It seems no significant phonetic information is lost due to independent processing in each band (Mirghafori and Morgan, 1998).

With regards to the problem of robustness in ASR, the multiband approach is particularly suited to band limited noise. As long as the noise leaves some bands unaffected, there is enough information in the remaining bands for ASR.

The multiband approach has interesting connection to missing data–once the garbled bands are identified, in the recombination stage all evidence coming from that band is completely disregarded. This is akin to marginalisation of the unreliable features. However, the hybrid ASR systems (most

often used in multistream/multiband ASR) trained on clean speech and with no data missing can not be easily adapted to handle partial evidence (see Section 3.4.2). Therefore as many recognisers need to be trained (with parts of the spectrum missing) as there are possible combinations of missing/present bands (Hagen et al., 1998, 2000). This clearly limits the number of possible streams of evidence that can be used and this number is rarely greater then seven.

7.2.3 "Bounded masking" by Holmes and Sedgwick (1986)

Holmes and Sedgwick (1986) work is an early precursor of one of the techniques presented in this thesis. Starting with different motivation (modelling the effects of masking in noisy speech), a technique which is essentially the bounded marginalisation (Chapter 6.3.2) is developed and applied both to a Dynamic Time Warping (DTW) as well as an HMM recogniser. This work is pioneering in several other respects. For example, it stresses the importance of keeping the noise which is localised in the T–F plane local in the further stages of processing (i.e. using spectral filterbanks rather then LPC coefficients).[4] After some study, de Veth et al. (1999) conclude the same a decade latter. Holmes and Sedgwick (1986) further give a recipe of training with noisy data, and advise on the perils of relying on quiet parts of the spectrum that may be swamped with noise later.

However, the work is in the context of noise masking and does not go any further than that. There is no notion of a general binary mask that (economically) separates the speech and the noise. Nor that it may be formed by enforcing the BU constraints to group patches of points in the T–F plane that come from the same source.

7.2.4 HMM decomposition by Varga and Moore (1990)

Varga and Moore (1990) introduced the HMM decomposition method (reviewed in Section 2.8.2) as a means of decoding multiple sources from a single sequence of observations. Like the work presented here, they also use spectral features and assume that the MAX operator combines the speech and the noise observations into noisy ones. Most often the number of sources is limited to 2 (as the size of the search space increases exponentially with the number of sources) with one of the sources being the speech source, while the other is the noise source.

This is a general method for combining any sources of variability in an observed signal, provided that the way they combine to yield the observation is known. An N–dimensional Viterbi search (N being the number of sources) is used to find the most likely sequence of states in the joint state–space.

The main differences of the approach compared to the MD work presented here are:

- HMM decompositions decodes all speech sources even if one of them only is of interest. It provides a complete solution. But usually the sequence of states some of the sources went through (e.g. the noise source) is of little interest.[5] In decomposition's terms, MD ignores the noise by assuming that in the missing points all values of the noise (and hence the speech) are equally likely between 0 and the noisy observation.

- When applied to recognising of speech in noise[6], it is assumed that the noise model brings valuable constraints to the decoding process. However, the sources are independent. Their state transition probabilities are independent. Their only interaction is through the environmental function (MAX in this case). MD's assumptions in composition's terms would be equivalent to a noise model which:

[4] even experienced researchers sometimes get surprised how similar the spectrograms of e.g. 0dB SNR look like to the spectrogram of the corresponding clean speech (with all the important speech structures clearly visible), while the difference in WERs the contemporary ASR systems achieve is stark

[5] it is possible to SUM (as opposed to MAX) the paths in the dimensions of the sources that aren't of interest; averaging over the possible states for the not–of–interest–sources may bring increased accuracy

[6] model decomposition is a general technique, and it has also been applied to recognition and separation of simultaneous multi–speech

- has as many states as frames, with the probability of the transition between successive states 1 while the rest of the transition probabilities are 0
- state dependent continuous probability density function which: for the noise points assigns equal probability of $1/o$ to the noise values between 0 and the observed value o, and 0 out of this interval; for the speech points it assumes the noise was 0 with absolute certainty (the distributions is a Dirac delta function)

Whether one considers a prior noise model to be stronger (more constrained) then the one equivalent to the MD assumptions, may depend on what is deemed "noise". To the "speech technologist" noise is mostly synonymous with "not speech". In the "hearing community" the term seems more specific: a sound with no discernible structure is noise – the rest are sounds.[7]

When considered in its stricter sense (the latter case), noise should bring large variance, and hence weak constraints. When considered in the wider sense (the former case) the noise model may bring constraints not captured by the MD model.[8]

The advantages of decomposition seem proportional to the amount of structure in the interfering sounds. For sounds of weak structure the properties of the speech alone may be enough to guide the separation. For example, Hirsch and Pearce (2000) reported on noisy training with speech mixed with 4 noises, and testing on the same 4 noises as well as 4 other, unseen noises. The differences in performance on the seen and the unseen noises was very small.

To bridge the gap of orders of magnitude between HSR and ASR in a real life environment, both approaches (with MD augmented with models capturing the BU constraints) may need to be applied in concert.

Integrated models of signal and background by Rose, Hofstetter, and Reynolds (1994)

Rose, Hofstetter, and Reynolds (1994) extend Varga and Moore (1990)'s work and treat the problem of merging any two (or more) discrete state–space models in a combined model in its full generality. Environmental functions other then MAX, and different feature domains (and how they influence the choice of the mixing function) are considered. The specifics of the MD model (MAX function, compressed spectral features) correspond to the special case of HMM decomposition discussed above.

7.3 Frequently Asked Questions

7.3.1 Is mask estimation just another name for noise estimation?

Mask estimation can be achieved through noise estimation, followed by local SNR estimation, as presented in this work. But it is not limited to it.

Mask estimation can also be accomplished via CASA. However limited in their utility, the CASA systems built over the past couple of decades are proof of a concept that separation using the properties of the speech alone is achievable. This is the important assumption for any practical application.

It is also worth repeating that the motivation for using a binary mask is not just an assumption of convenience. The principle of exclusive allocation (a point in a T–F plane is either speech or noise, without any go–betweens) is not advocated for any domain other than compressed spectral features. Choosing the mask as an interface between the separation and the recognition part of the system is not only supported by what we know about human hearing, but has already been utilised in computational models both for recognition (HMM decomposition) and separation (Wu et al., 1998a).

[7] e.g. a roaring bus passing nearby is noise to the former, or another sound in the auditory scene to the latter

[8] it seems the two communities treat the sounds that have prominent structure along the frequency axis only (significantly change along the frequency axis, but not in time) of the T-F plane (e.g. "coloured noise")

The CASA systems built so far have been mostly knowledge based, rather then statistical. Their drawbacks are typical of that approach, have prevented their further development, and are reminiscent of the drawbacks of the ASR systems that predated the current statistical ones:

- complex models, not really trainable on a large corpus of data
- difficulty reconciling multiple, and sometimes conflicting, sources of evidence

with the added

- impossible to integrate with ASR systems, apart from being an entirely independent frontend

However, these drawbacks can be tackled successfully (as was the case with the ASR systems) in a data driven system. Work toward learning what is speech from the data has been reported recently. Ris (2000) trained a neural network to estimate the posterior probability of the mask given the noisy data. Seltzer, Raj, and Stern (2000) trained a multivariate Gaussian classifier to assess the probability of a feature being present or missing. Brown, Wang, and Barker (2001) integrated a connectionist CASA system with missing data ASR. All attempts reported significant performance improvements.

7.3.2 Can acoustic evidence alone guide the separation?

When the non–speech features are matched with speech models, they typically fall in very low density regions. These points are known as *outliers*. When some speech state during the Viterbi search produces an extremely low score, every path passing through it will have an extremely low score. The acoustic back–off technique (de Veth et al., 1998) allocates a minimal constant probability mass for every observation, limiting how low a score can become. The UNION model (Ming et al., 2000) models the speech vectors p.d.f. with a form that is a sum of products. Consequently, the products containing outliers are very small and contribute little to the sum. Similarly, the full combination multiband approach (Hagen et al., 1998, 2000) sums the acoustic evidence over all possible combinations of present/missing data in every frame and relies on outliers to cancel themselves out in the sum.

These methods seem to be more beneficial for artificial then real noises. They suffer from common problems:

- patches of realistic noise will match some speech states; hence noise may not always result in an outlier[9]
- all possible combinations of the reliable features are weighted equally, as if all possible masks M have the same probability $P(M|Q,W)$

The experiment with the MLM (pp. 101) points to the limitations of using the ASR acoustic models to select features as speech or non–speech. It seems that the speech models capture a set of constraints different then the ones pertinent to the BU grouping. So, while all the above referenced work uses some of the ideas that motivated the MD work presented here (exclusive allocation, localised disturbance of the features) in various ways, other ideas, maybe important for increased robustness (mainly regarding the separation part of an integrated ASR–separation system), are omitted from consideration.

7.3.3 What about convolutional noise?

The MD work presented here is entirely concerned with additive noise only. We have no reasons to believe that the principle of disjoint allocation of energy (central to the MD techniques) is true for convolutional noise. On the contrary, a limited test carried out on test C of the Aurora 2 database (containing artificially induced convolutional noise) suggest that MD techniques as presented here

[9] stated p.d.f. $p(\mathbf{x}|q)$ models the speech source only; we have no reason to assume anything about the joint speech–noise p.d.f. $p(\mathbf{x}_p, \mathbf{x}_m|q)$

are not suitable for convolutional noise. Using the apriori masks, it was found that the speech models, rather then mask estimation, are to blame. Spectral domain features are susceptible to the spectral tilt introduced by the convolutional noise. The models expect energy in the wrong frequency bins and perform poorly even at high SNRs and apriori mask.

7.4 Problems with the MD model for ASR

7.4.1 Mask estimation

The identification of the speech regions is the main problem at present. The apriori oracle masks indicate that the principal gap between the achievable and achieved performance is due to poor mask estimation. In this work a stationary (in the former) and adaptive (in the latter set of experiments) noise estimation was used. Recently it was augmented by harmonicity based CASA as well as apriori 4–band mask grouping in the multisource decoder (Section 7.2.1). The problem of separating the speech from the rest of the auditory scene even in simple acoustic environments remains challenging.

7.4.2 Merging the likelihoods during MD Viterbi search

It is not completely clear what is the best way to merge the likelihoods resulting from paths with different masks. This problem does not arise in the "single source" Viterbi search, as all paths "see" the same data. However, it does arise in the context of the multisource decoder. Dividing the probability mass from the missing features by the range of integration to yield an "average likelihood" (for the assumed noise model) has been used so far. But the problem re–emerges when dealing with any features that need to be marginalised completely as no bounds are known, and carry no information whatsoever about the source (e.g. when missing delta features without bounds).

7.4.3 Choice of features for separation and recognition

Separation (and noise estimation) usually require frequency domain features. For example, good frequency resolution is necessary for on–line noise estimation. In contrast, ASR models are generally better–off with features that contain information about the gross vocal tract shape only (spectral envelope) – not the fine spectrum. At present, the former set of features is converted to the latter by frequency axis warping, compression, projection on cosine basis and truncation (Mel–frequency cepstral coefficients, MFCCs). The nature of the transform path makes the merger of the separation and recognition techniques difficult. The noise which is localised in the T–F plane, in which the principle of disjoint allocation of energy holds (after the compression), is spread over every cepstral coefficient.[10] Spectral features are correlated and need more Gaussians in the models to capture those correlations (compared to MFCCs). Subtracting the channel dependent long term component of the spectrum also seems somewhat less effective then cepstral subtraction for channel normalisation. The truncation of the cepstral coefficients (typically, only the first 12 out of 24 are retained) also provides a certain resilience to noise (especially at high SNRs). More fundamentally, separation makes use of speech features (like F_0) which are speaker dependent, while the recognition subsystem needs as speaker independent features as possible.

[10] it was also argued that the whole conversion path is altogether prone to side effects that diminish the discriminability between the vowels when the pitch increases (de Cheveigne and Kawahara, 1999); it seems that sampling the short spectrum at the harmonics of the fundamental F_0 provides information about the gross spectral shape and is not affected by the pitch change

7.5 Future work

7.5.1 Data driven masks models

It was already noted in Section 7.3.1 that it seems probable that further gains in robust ASR will come from data driven separation models, which are computationally trackable and trainable on a large corpus of data. The data needed can be easily generated as apriori masks. It would be advantageous if such techniques not only single out the most probable mask, but also produce a probability distribution for all possible masks.

For example, one possibility is to infer a state dependent model for the apriori (before any data is seen)[11] distribution $P(M|Q,W)$ of the apriori masks. Training data can be produced from state force aligned speech and apriori masks. This is akin to Viterbi (rather then Baum–Welch) training, as every state is assigned a pool of data vectors. For binary feature vectors a Bernoulli mixture p.d.f. $P(\mathbf{m}|q,W) = \sum_k P(k) \prod_i \mu_{i,k}^{m_i}(1-\mu_{i,k})^{1-m_i}$ may be used for modelling (Carreira-Perpiñán and Renals, 2000). While this above may capture some of the apriori probability of a mask, ultimately features $\mathbf{o}_{sep}$ pertinent to the separation process will need to be used in a mask model $P(M|O_{sep},Q)$.

7.5.2 Coupling separation and recognition for better models

Speech separation/enhancement researchers tend to use a simple speech model, while ASR researchers tend to assume a simple environmental model. Usually, the former model the speech source as static, without time structure. While the latter assume that the convolutional noise is constant and that additive noise is slowly changing. It maybe beneficial to reconsider the trade–offs in both cases. For example, Acero, Altschuler, and Wu (2000) reported on using both a more complex environmental and speech model. The usual obstacle is that the ASR features are in domain where separation model becomes complicated, without closed form solutions. The autoregressive (AR) speech models (Logan, 1998) may be one possible compromise.

7.5.3 A speculation on an integrated speech separation and recognition model

Model (de)composition, with the speech along X, the mask along Y, and the time along Z axis is used to tie the separation and recognition parts of the system together. For the purposes of mask estimation, the source can be in several states. For each "separation state" q_Y either:

- a separate algorithm for identification of a particular cue thought to be important for BU grouping is applied (e.g. identification of harmonics, onsets, offsets, common modulation, etc)[12]

or

- the states are purely a modelling tool to achieve a better mask model, with no explicit relation to the auditory grouping cues[13]

Models and topology

Both speech and mask models are HMMs. The speech model is a straight–through standard speech HMM with states q_X with a distribution of $p(\mathbf{o}|\mathbf{m},q_X)$, where $\mathbf{m}$ is a binary mask. The mask HMM is a "noise–like" HMM with interconnected states q_Y, each with a probability distribution $p(\mathbf{o}|\mathbf{m},q_Y)$. For example, $p(\mathbf{o},\mathbf{m}|q_Y) = \prod_i \mu_i^{m_i}\mathcal{N}(o_i;\mu_{1,i},\sigma_{1,i}^2)^{m_i}(1-\mu_i)^{1-m_i}\mathcal{N}(o_i;\mu_{2,i},\sigma_{2,i}^2)^{1-m_i}$ with $p(\mathbf{o}|\mathbf{m},q_Y) = \prod_i \mathcal{N}(o_i;\mu_{1,i},\sigma_{1,i}^2)^{m_i}\mathcal{N}(o;\mu_{2,i},\sigma_{2,i}^2)^{1-m_i}$ and $p(\mathbf{m}|q_Y) = \prod_i \mu_i^{m_i}(1-\mu_i)^{1-m_i}$

[11] not to be confused with apriori mask, or even probability of apriori mask – which is 1
[12] similarly to an early CASA system by Weintraub (1985)
[13] but we would expect that during training cues pertinent to BU grouping will be "discovered" from the data

for a joint Gaussian distribution for the continuous observations and Bernoulli distribution for the discrete mask. This distribution is different from $p(\mathbf{o}|\mathbf{m}, q_X)$ in that it expresses the probability of any speech, not a particular sound in the language.

Features

Features **o** are compressed FFT magnitude or auditory filterbank features with fine spectral resolution. For the speech states the p.d.f. is $p(\mathbf{o}|\mathbf{m}, q_X)$ (not mere $p(\mathbf{o}|q_X)$) and most of the points are usually missing. The feature vector has many more dimensions then usual for recognition, hence:

- MG (pp. 85), instead of BMG (pp. 85) is used to match the evidence, with the counterevidence ignored

or

- the evaluation is computationally intensive

Using the features this way solves the problems of: (a) choice of different features for separation and recognition; (b) decorrelation; and (c) sampling the envelope of the spectrum instead of its fine structure.[14]

Training

The parameters of the speech model are inferred separately from the mask model using clean speech. The mask model is inferred from noisy speech and apriori masks. Or both are inferred simultaneously, with a joint–estimation scheme similar to the ones presented by Kadirkamanathan and Varga (1991), Roweis (2000) and Graciarena (2000).

Testing – recognition, separation, or both

Decoding along the X axis only (with "don't care" along Y axis, i.e. paths along Y axis are SUMmed together, not MAXimised) to find the best path in that direction amounts to recognition:[15]

$$\begin{aligned}
Q_X^* &= \underset{Q_X}{argmax} P(Q_X|O) = \underset{Q_X}{argmax} P(O|Q_X)P(Q_X) \\
&= \underset{Q_X}{argmax} \sum_{Q_Y} \sum_{M} P(O|M, Q_X, Q_Y)P(M|Q_Y)P(Q_Y)P(Q_X) \\
&= \underset{Q_X}{argmax} \sum_{Q_Y} \sum_{M} \frac{P(O|M, Q_X)P(O|M, Q_Y)}{P(O|M)} P(M|Q_Y)P(Q_Y)P(Q_X) \\
&= \underset{Q_X}{argmax} \sum_{Q_Y} \sum_{M} \frac{P(O|M, Q_X)P(O|M, Q_Y)P(M|Q_Y)P(Q_Y)P(Q_X)}{\frac{\sum_{Q'_Y} P(O|M,Q'_Y)P(M|Q'_Y)P(Y')}{\sum_{Q_Y"} P(M|Q_Y")P(Q_Y")}}
\end{aligned} \tag{7.1}$$

Separation amounts to finding the most likely mask M^*:

$$M^* = \underset{M}{argmax} P(M|O) = \underset{M}{argmax} \sum_{Q_Y} P(O|M, Q_Y)P(M|Q_Y)P(Q_Y) \tag{7.2}$$

[14] de Cheveigne and Kawahara (1999) demonstrated that this works for synthetic vowels; for the purpose of this speculation it is assumed that the same or similar is true for real sounds and all phones; it is also reminiscent of the "spectral peaks" MD work of Barker (1998) (Section 2.7.7), with the mask playing the role of the peaks detector

[15] conditioning of the word W for Q_X is dropped; $P(a|b, c, d) = P(a|b, d)P(a|c, d)/P(a|d)$ iif b and c are independent; Q_X and Q_Y are independent; M does not depend on Q_X

Decoding along the X axis and finding the most likely mask in the same time amounts to simultaneous recognition and separation:

$$\begin{aligned}(Q_X^*, M^*) &= \underset{(Q_X,M)}{argmax} P(Q_X|O) = \underset{(Q_X,M)}{argmax} P(O|Q_X)P(Q_X) \\ &= \underset{(Q_X,M)}{argmax} \sum_{Q_Y} \frac{P(O|M,Q_X)P(O|M,Q_Y)P(M|Q_Y)P(Q_Y)P(Q_X)}{\frac{\sum_{Q'_Y} P(O|M,Q'_Y)P(M|Q'_Y)P(Y')}{\sum_{Q_Y"} P(M|Q_Y")P(Q_Y")}}\end{aligned} \tag{7.3}$$

This is Eq. (7.1) with the $\sum_M$ replaced with $argmax_M$.

7.6 Conclusions

Current ASR systems, although still lagging far behind HSR, perform well enough for many applications in a controlled environment. However, in a less restricted environment desirable for many applications, their performance degrades dramatically rendering them unusable. The lack of robustness has been identified as the most significant limitation of the current technology (Sagayama and Kiyoami, 1997).

The present systems assume first and foremost a single source. Attempts to accommodate the basic design for multiple sources during the Aurora 2 competition seem to suggest that:

- Even training with noisy data (multiconditional baseline) when the noise is known in advance does not achieve the performance needed for many applications.[16]
- The performance gains achieved are mainly due to:
  - frame deletion – frames considered too noisy are discarded
  - cleaning of the noisy speech via algorithms that use the gross speech features (e.g. harmonicity) to discriminate between the speech and the noise

Regarding the first point, the techniques presented here offer possible gains which exceed what is currently possible with matched training, as demonstrated with the apriori oracle masks (pp. 112–113).

With regards to the second point, using the "ideal" filter (the apriori masks) demonstrates that it is not the faulty models that cause the performance loss. But rather feeding the models non–speech when they expect speech is the main reason for degradation.

The problem of lack of robustness seems to be one of separation, rather than of poor speech modelling. The only untapped source of constraints at the moment appear to be the BU constraints pertinent to "grouping", as discussed in Chapter 4. On its part, the ASR backend can make the task of the separation frontend easier by foregoing full speech spectrum reconstruction. Merely identifying the speech parts in a sounds mixture should suffice for the ASR backend.

In this work we have experimented with some aspects of a possible system that may implement these ideas. We believe to have demonstrated that:

- The hard problem of ASR of speech in noise can be separated into two subproblems – speech identification and recognition of the partial speech. Treating the mask that binds both subsystems as a random variable opens opportunities to tackle the mask estimation problem.
- Using the partial spectrum for recognition is good enough for ASR not only with artificial, but also with realistic noises. Hence an enhancement frontend which reconstructs the whole spectrum is not necessary.

[16] a probable application of the connected digits task is phone dialling by voice; the target sentence error rate (SER) is about 3%, which requires WER of about 0.3%

- However, if needed for a particular application (e.g. speech enhancement), the missing parts of the spectrum can be reconstructed during the decoding in a principled way and employing a strong speech model (as provided by the recogniser) for this purpose.
- Further gains in robust ASR at mid and low SNRs are achievable through better modelling of the identification/separation frontend of the system.
- Tapping into the BU constraints by devising appropriate features for the "cues" and modelling them accordingly is likely to lead to better recognition accuracy in less controlled environments.

In the end, several commonly held assumptions (Peters et al., 1999)[17] about how humans and machines hear may need to be revisited if machines are to approach human hearing in its resilience to noise.

[17]Peters et al. (1999) reversed the tables and played to trained human subjects speech that has been processed by a standard ASR frontend and then reconstructed; HSR deteriorated with each step along the ASR frontend processing chain; the decrease is ever larger as the SNR decreases

Appendix A

Comparative performance of techniques for noise robust ASR

An incomplete list of improvements in the accuracy of various ASR systems with proposed techniques for robust ASR published in the surveyed literature:

Technique and/or reference	Vocabulary	Speaker(s)	Noise	Baseline	Compensated	Matched	Clean
SS (Lockwood and Boudy, 1991)	43 isolated words	multi	Car	67%	94.9%		
NSS (Lockwood and Boudy, 1991)	43 isolated words	multi	Car	67%	98%		
SS (Xie and Campernolle, 1993)		multi	10dB Telephone	44.76%	54.42%		
NSS (Xie and Campernolle, 1993)		multi	10dB Telephone	44.76%	67.48%		
NSS+noise variance (Xie and Campernolle, 1993)		multi	10dB Telephone	44.76%	71.56%		
Adaptive Wiener filter (Vaseghi and Milner, 1993)	26 connected letters	multi	10dB	4.5%	56.2%	61.4%	
Warped Wiener filter (Agarwal and Cheng, 1999)	12 connected digits	multi	10dB, 4 noises average	79.43%	82.70%	92.75%	
Klatt (1976) algorithm (Varga et al., 1988)	isolated digits	multi	9dB pink		99%		
Bridle et al. (1984) algorithm (Varga et al., 1988)	isolated digits	multi	9dB pink		90%		
Holmes and Sedgwick (1986) algorithm (Varga et al., 1988)	isolated digits	multi	9dB pink		86%		

continued on next page

Table A.1: Summary table of performance of various techniques for robust ASR published in the literature

continued from previous page

Technique and/or reference	Vocabulary	Speaker(s)	Noise	Baseline	Compensated	Matched	Clean
Feature normalisation (Tibrewala and Hermansky, 1998)	13 isolated digits	multi	destroyer–engine	73.4%	94.8%		
Feature normalisation (Tibrewala and Hermansky, 1998)	13 isolated digits	multi	factory	73.8%	95.8%		
Feature normalisation (Tibrewala and Hermansky, 1998)	13 isolated digits	multi	pink	75.7%	96.3%		
Feature normalisation (Tibrewala and Hermansky, 1998)	13 isolated digits	multi	babble	75.4%	94.3%		
Feature normalisation (Tibrewala and Hermansky, 1998)	13 isolated digits	multi	Volvo	75.4%	96.5%		
Feature normalisation (Tibrewala and Hermansky, 1998)	13 isolated digits	multi	white	75.2%	90.3%		
Feature normalisation (Tibrewala and Hermansky, 1998)	13 isolated digits	multi	high–frequency radio	74.2%	90.8%		
Noisy–to–clean mapping (Gao and Haton, 1993)	3–digit strings	multi	12dB babble		98.7%		
Noisy–to–clean mapping (Gao and Haton, 1993)	3–digit strings	multi	12dB Lynx helicopter		98.7%		
Morphological constraints (Hansen, 1994)	35 words	3 speakers	10dB white, Lombard	8.1%	62.1%		
Morphological constraints (Hansen, 1994)	35 words	3 speakers	10dB aircraft cockpit, Lombard	34.3%	72.1%		
Morphological constraints (Hansen, 1994)	35 words	3 speakers	10dB computer, Lombard	23.3%	72.3%		
Autoregressive HMM (Logan and Robinson, 1998)	Resource management task	multi	12dB Lynx helicopter	18.9%	59.2%	63.4%	
Autocorrelation features (Yuo and Wang, 1999)	Mandarin isolated digits	multi	10dB white	62.1%	93.3%		
PLP (Hermansky and Morgan, 1994)	isolated digits	multi	10dB car+wireless phone		63.0%	90%	95%

continued on next page

Table A.1: Summary table of performance of various techniques for robust ASR published in the literature

continued from previous page

Technique and/or reference	Vocabulary	Speaker(s)	Noise	Baseline	Compensated	Matched	Clean
RASTA (Hermansky and Morgan, 1994)	isolated digits	multi	10dB car+wireless phone		50.0%		96.7%
PLP+CMN (Hermansky and Morgan, 1994)	isolated digits	multi	10dB car+wireless phone		58.0%		95.7%
Lin–log RASTA (Hermansky and Morgan, 1994)	isolated digits	multi	10dB car+wireless phone		86.3%		96.3%
"textbook" auditory (Tian et al., 1998)	isolated word	multi	average, clean to -10dB, Volkswagen car at 100km/h	88.31%	89.72%		99.02%
two stream auditory (Tian et al., 1998)	isolated word	multi	average, clean to -10dB, Volkswagen car at 100km/h	88.31%	90.30%		99.13%

Table A.1: Summary table of performance of various techniques for robust ASR published in the literature

Appendix B

Multidimensional integral of the sigmoid function - an analytic solution

Lets consider a single unit with two inputs x_1 and x_2, corresponding weights w_1 and w_2 and sigmoid transfer function $net(x_1, x_2) = \frac{1}{1+e^{-w_1x_1-w_2x_2}}$. We integrate over the input x_1, limits being a_1 and b_1:

$$\begin{aligned}
\int_{a_1}^{b_1} \frac{dx_1}{1+e^{-w_1x_1-w_2x_2}} &= \int_{a_1}^{b_1} \frac{dx_1}{1+e^{-w_1x_1-w_2x_2}} \cdot \frac{e^{w_1x_1}}{e^{w_1x_1}} = \int_{a_1}^{b_1} \frac{e^{w_1x_1}dx_1}{e^{w_1x_1}+e^{-w_2x_2}}, \\
&= \frac{1}{w_1} \int_{a_1}^{b_1} \frac{d(e^{w_1x_1})}{e^{w_1x_1}+e^{-w_2x_2}} = \frac{1}{w_1} \int_{a_1}^{b_1} \frac{d(e^{w_1x_1}+e^{-w_2x_2})}{(e^{w_1x_1}+e^{-w_2x_2})}, \\
&= \frac{1}{w_1} \ln(e^{w_1x_1}+e^{-w_2x_2}) \Big|_{a_1}^{b_1} = \frac{1}{w_1} \ln \frac{e^{w_1b_1}+e^{-w_2x_2}}{e^{w_1a_1}+e^{-w_2x_2}}, \\
&= \frac{1}{w_1} \ln \frac{e^{w_1b_1}+e^{-w_2x_2}}{e^{w_1a_1}+e^{-w_2x_2}} \cdot \frac{e^{w_2x_2}}{e^{w_2x_2}} = \frac{1}{w_1} \ln \frac{1+e^{w_1b_1+w_2x_2}}{1+e^{w_1a_1+w_2x_2}}.
\end{aligned} \tag{B.1}$$

The result can also be expressed in terms of the transfer function $net(x_1, x_2)$:

$$\begin{aligned}
\int_{a_1}^{b_1} \frac{dx_1}{1+e^{-w_1x_1-w_2x_2}} &= \frac{1}{w_1} \ln \frac{1+e^{w_1a_1+w_2x_2}}{1+e^{w_1b_1+w_2x_2}}, \\
&= \frac{1}{w_1} \ln \frac{\frac{1}{1+e^{w_1a_1+w_2x_2}}}{\frac{1}{1+e^{w_1b_1+w_2x_2}}} = \frac{1}{w_1} \ln \frac{net(-a_1, -x_2)}{net(-b_1, -x_2)}.
\end{aligned} \tag{B.2}$$

But the result is not going to be used later.

We can easily generalise the result for n inputs - instead w_2x_2 of we would have $\sum_{i=2}^{n} w_ix_i$ and the integral would be:

$$\int_{a_1}^{b_1} \frac{dx_1}{1+e^{-\sum_{i=1}^{n} w_ix_i}} = \frac{1}{w_1} \ln \frac{1+e^{w_1b_1+\sum_{i=2}^{n} w_ix_i}}{1+e^{w_1a_1+\sum_{i=2}^{n} w_ix_i}} \tag{B.3}$$

Now let's consider double integral. The unit has three inputs and one output. The inputs are x_1, x_2 and x_3 and the corresponding weights w_1, w_2 and w_3. The unknown inputs are x_1 and x_2

with distribution bounded in the corresponding intervals $[a_1, b_1]$ and $[a_2, b_2]$. The sigmoid transfer function is $\frac{1}{1+e^{-w_1x_1-w_2x_2-w_3x_3}}$. Using Eq. (B.1) we have for the marginal:

$$\begin{aligned}
&\int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \cdot \frac{1}{1+e^{-w_1x_1-w_2x_2-w_3x_3}} = \\
&\qquad = \frac{1}{w_2} \int_{a_1}^{b_1} \ln \frac{1+e^{w_1x_1+w_2b_2+w_3x_3}}{1+e^{w_1x_1+w_2a_2+w_3x_3}} dx_1, \\
&\qquad = \frac{1}{w_2} \left\{ \int_{a_1}^{b_1} \ln(1+e^{w_1x_1+w_2b_2+w_3x_3}) dx_1 - \int_{a_1}^{b_1} \ln(1+e^{w_1x_1+w_2a_2+w_3x_3}) dx_1 \right\}.
\end{aligned} \tag{B.4}$$

For forms of type $\int_a^b \ln(1+e^{wx+C})dx$ we have:

$$\begin{aligned}
\int_a^b \ln(1+e^{wx+C})dx &= \int_{-e^{wa+C}}^{-e^{wb+C}} \ln(1-u) \cdot \frac{1}{w} \cdot \frac{du}{u}, \\
&= \frac{1}{w} \left\{ \int_{-e^{wa+C}}^{0} \ln(1-u)\frac{du}{u} \quad + \int_{0}^{-e^{wb+C}} \ln(1-u)\frac{du}{u} \right\}, \\
&= \frac{1}{w} \left\{ \int_{0}^{-e^{wa+C}} -\ln(1-u)\frac{du}{u} \quad - \int_{0}^{-e^{wb+C}} -\ln(1-u)\frac{du}{u} \right\}, \\
&= \frac{1}{w} \left\{ \int_{0}^{-e^{wa+C}} Li_1(u)\frac{du}{u} \quad - \int_{0}^{-e^{wb+C}} Li_1(u)\frac{du}{u} \right\}, \\
&= \frac{1}{w} \left\{ Li_2(-e^{wa+C}) - Li_2(-e^{wb+C}) \right\},
\end{aligned} \tag{B.5}$$

with the substitution $-e^{wx+C} = u$; $-e^{wx+C}wdx = du$; $dx = \frac{1}{w}\frac{du}{u}$ and using the Eq. (B.8).

The function $Li_1(u)$ is polilogarithm function of order one (Lewin, 1958). It can be expressed as $Li_1(z) = -\ln(1-z)$ for $|z| < 1$, or as an infinite sum:

$$Li_1(z) = z + \frac{z^2}{2} + \frac{z^3}{3} + \ldots = \sum_{n=1}^{\infty} \frac{z^n}{n}. \tag{B.6}$$

The polilogarithm function of order m, $Li_m(z)$, expressed through it's infinite sum is:

$$Li_m(z) = z + \frac{z^2}{2^m} + \frac{z^3}{3^m} + \ldots = \sum_{n=1}^{\infty} \frac{z^n}{n^m}, \tag{B.7}$$

for $|z| < 1$. For z outside of this interval, the expression:

$$Li_{m+1}(z) = \int_0^z Li_m(t)\frac{dt}{t} \tag{B.8}$$

can be used (Lewin, 1958, pp. 169).

So, for the double integral (B.4) we have:

$$\begin{aligned}
&\int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \cdot \frac{1}{1+e^{-w_1x_1-w_2x_2-w_3x_3}} = \\
&\quad = \frac{1}{w_1w_2}\Big\{ Li_2(-e^{w_1a_1+w_2b_2+w_3x_3}) - Li_2(-e^{w_1b_1+w_2b_2+w_3x_3}) \\
&\qquad - Li_2(-e^{w_1a_1+w_2a_2+w_3x_3}) + Li_2(-e^{w_1b_1+w_2a_2+w_3x_3})\Big\}, \\
&\quad = \frac{1}{w_1w_2}\Big\{ -Li_2(-e^{w_1a_1+w_2a_2+w_3x_3}) + Li_2(-e^{w_1a_1+w_2b_2+w_3x_3}) \\
&\qquad + Li_2(-e^{w_1b_1+w_2a_2+w_3x_3}) - Li_2(-e^{w_1b_1+w_2b_2+w_3x_3})\Big\}.
\end{aligned} \tag{B.9}$$

Using substitution $-e^{wx+C}$ (as in Eq. (B.5)) and Eq. (B.8) the definite integral of polilogarithm function of any argument of form $-e^{wx+C}$ can be expressed as subtraction of two terms – polilogarithm functions of higher (by one) order of arguments of the same type ($-e^{wx+C}$):

$$\begin{aligned}
\int_a^b Li_m(-e^{wx+C})dx &= \frac{1}{w}\int_{-e^{wa+C}}^{-e^{wb+C}} Li_m(u)\frac{du}{u}, \\
&= \frac{1}{w}\Big\{ \int_{-e^{wa+C}}^{0} Li_m(u)\frac{du}{u} + \int_{0}^{-e^{wb+C}} Li_m(u)\frac{du}{u}\Big\}, \\
&= \frac{1}{w}\Big\{ -\int_{0}^{-e^{wa+C}} Li_m(u)\frac{du}{u} + \int_{0}^{-e^{wb+C}} Li_m(u)\frac{du}{u}\Big\}, \\
&= \frac{1}{w}\Big\{ -Li_{m+1}(-e^{wa+C}) + Li_{m+1}(-e^{wb+C})\Big\}.
\end{aligned} \tag{B.10}$$

The multidimensional integral of a sigmoid transfer function can be analytically expressed as a sum of polilogarithms of the same order. Every additional unknown input x_i integrated over the bounds $[a_i, b_i]$, raises the dimensionality of the integral by one, which consequently doubles the number of terms in the sum of polilogarithm functions and also raises by one the order of the functions in the sum. For example, for three dimensional integral we get:

$$\begin{aligned}
\int_{a_1}^{b_1} b_1dx_1 \int_{a_2}^{b_2} b_2dx_2 \int_{a_3}^{b_3} b_3dx_3 \cdot \frac{1}{1+e^{-w_1x_1-w_2x_2-w_3x_3-w_4x_4}} = \frac{1}{w_1w_2w_3}\Big\{ \\
Li_3(-e^{w_1a_1+w_2a_2+w_3a_3+w_4x_4}) - Li_3(-e^{w_1a_1+w_2a_2+w_3b_3+w_4x_4}) \\
- Li_3(-e^{w_1a_1+w_2b_2+w_3a_3+w_4x_4}) + Li_3(-e^{w_1a_1+w_2b_2+w_3b_3+w_4x_4}) \\
- Li_3(-e^{w_1b_1+w_2a_2+w_3a_3+w_4x_4}) + Li_3(-e^{w_1b_1+w_2a_2+w_3b_3+w_4x_4}) \\
+ Li_3(-e^{w_1b_1+w_2b_2+w_3a_3+w_4x_4}) - Li_3(-e^{w_1b_1+w_2b_2+w_3b_3+w_4x_4})\Big\}.
\end{aligned} \tag{B.11}$$

Considering that one–dimensional integral B.1 can be expressed as:

$$\begin{aligned}
\int_{a_1}^{b_1} \frac{dx_1}{1+e^{-w_1x_1-w_2x_2}} &= \frac{1}{w_1}\Big\{ \ln(1-(-e^{w_1b_1+w_2x_2})) - \ln(1-(-e^{w_1a_1+w_2x_2}))\Big\}, \\
&= \frac{1}{w_1}\Big\{ Li_1(-e^{w_1a_1+w_2x_2}) - Li_1(-e^{w_1b_1+w_2x_2})\Big\},
\end{aligned} \tag{B.12}$$

it can be shown that the signs of the terms are:
for 1-D integral: + -
for 2-D integral: - + + -
for 3-D integral: + - - + - + + -
for 4-D integral: - + + - + - - + + - - + - + + -
⋮
and can be computed by the following pseudo–code:

```
int Sign(int NoIntegral, NoTerm) /* returns ±1 */
if (NoIntegral == 1)
  if (NoTerm == 0)
    return 1;
  else
    return -1;
else
  if (NoTerm < 2^(NoIntegral-1))
    return -Sign(NoIntegral-1, NoTerm);
  else
    return Sign(NoIntegral-1, NoTerm - 2^(NoIntegral-1));
```

Finally, for a single sigmoid unit with n inputs x_i, and weights w_i for $i = 1, 2, \ldots, n$ and transfer function

$$\frac{1}{1 + e^{-\sum_{i=1}^{n} w_i x_i}}, \tag{B.13}$$

and integrating over m of the inputs in the intervals $[a_i, b_i]$ for $i = 1, 2, \ldots, m$ and for the rest $n-m$ of the inputs the exact values c_i for $i = m+1, m+2, \ldots, n$ are known, we have the expression:

$$\left(\int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \ldots \int_{a_m}^{b_m} dx_m \cdot \frac{1}{1 + e^{-\sum_{i=1}^{n} w_i x_i}} \right) \Bigg|_{(x_{m+1}, x_{m+2}, \ldots, x_n) = (c_{m+1}, c_{m+2}, \ldots, c_n)}$$
$$= \frac{1}{\prod_{i=1}^{m} w_i} \sum_{i=0}^{2^m - 1} Sign(m, i) \cdot Li_m \left(-e^{\sum_{j=1}^{m} w_j \langle a_j \text{ or } b_j \rangle + \sum_{j=m+1}^{n} w_j c_j} \right) \tag{B.14}$$

where $\langle a_j \text{ or } b_j \rangle$ means "a_j or b_j depending on the term number (i.e. the value of i)". For $i = 0$ the sum is $w_1 a_1 + \ldots + w_m a_m$; for $i = 1$ it is $w_1 a_1 + \ldots + w_{m-1} a_{m-1} + w_m b_m$; for $i = 2$ it is $w_1 a_1 + \ldots + w_{m-2} a_{m-2} + w_{m-1} b_{m-1} + w_m a_m$; $\ldots$; for $i = 2^m - 2$ it is $w_1 b_1 + \ldots + w_{m-1} b_{m-1} + w_m a_m$; for $i = 2^m - 1$ it is $w_1 b_1 + \ldots + w_m b_m$.

Appendix C

Linear transformation of the missing features

The idea that parts of the spectrum are unaffected by noise and can be used alone, ignoring ones contaminated with noise, assumes that the subsequent pattern matching is performed in the same domain as the identification of the reliable features. However, it is of interest to consider what happens if the features undergo a linear transform before the pattern matching phase. This is often the case in the contemporary ASR systems. The systems utilise the envelope of some spectral representation, rather then spectral representation itself. The spectral envelope is much more speaker and pitch independent, while allowing speech discrimination. It is also advantageous to use a transform that approximately decorrelates the features, thus reducing the number of Gaussians in the mixture needed to accurately model the correlations between the features in the state p.d.f.s[1]. The Discrete Cosine Transform (Rao and Yip, 1990) is routinely used with the speech signal to achieve both.

Lets assume that the spectral feature vector $\mathbf{x}$ undergoes a linear transform $C\mathbf{x}$ before being used in the state likelihood computation Eq. (5.12). The contribution of each individual Gaussian to the likelihood (with conditioning on the state and mixture dropped) is:

$$p(\mathbf{x}) = \mathcal{N}(C\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \; e^{-\frac{1}{2}(C\mathbf{x}-\mu)^T\Sigma^{-1}(C\mathbf{x}-\mu)} \tag{C.1}$$

where Σ is diagonal, and C is a linear transform matrix with dimensionality $\mathbf{k}$ x $(\mathbf{p}+\mathbf{m})$ (p is the number of present, and m the number of missing features). The aim is to compute the marginal $p(\mathbf{x}_p) = \int p(\mathbf{x}) d\mathbf{x}_m$.

With suitable reordering of the rows of $\mathbf{x}$ and columns of C we have:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_p \\ \mathbf{x}_m \end{bmatrix}, \quad C = \begin{bmatrix} C_p & C_m \end{bmatrix}, \quad C\mathbf{x} - \mu = \begin{bmatrix} C_p\mathbf{x}_p + C_m\mathbf{x}_m - \mu \end{bmatrix}, \tag{C.2}$$

where $\mathbf{x}_p$ and $\mathbf{x}_m$ are vectors with dimensions $\mathbf{p}$ and $\mathbf{m}$ respectively and C_p and C_m are matrices of $\mathbf{k}$ x $\mathbf{p}$ and $\mathbf{k}$ x $\mathbf{m}$ elements respectively. The quadratic form in the exponent of Eq. (C.1) becomes:

$$\begin{aligned} (C\mathbf{x}-\mu)^T\Sigma^{-1}(C\mathbf{x}-\mu) &= (C_m\mathbf{x}_m + C_p\mathbf{x}_p - \mu)^T\Sigma^{-1}(C_m\mathbf{x}_m + C_p\mathbf{x}_p - \mu) \\ &= \{C_m[\mathbf{x}_m + C_m^{\#}(C_p\mathbf{x}_p - \mu)]\}^T\Sigma^{-1}\{C_m[\mathbf{x}_m + C_m^{\#}(C_p\mathbf{x}_p - \mu)]\} \\ &= [\mathbf{x}_m + \underbrace{C_m^{\#}(C_p\mathbf{x}_p - \mu)}_{-\mu_1}]^T[\underbrace{C_m^T\Sigma^{-1}C_m}_{\Sigma_1^{-1}}][\mathbf{x}_m + \underbrace{C_m^{\#}(C_p\mathbf{x}_p - \mu)}_{-\mu_1}] \end{aligned} \tag{C.3}$$

[1] but a mixture may still be needed to model a potential multimodality of the distributions, especially on a large and varied speech corpus

where $C_m^{\#}$ is a generalised inverse of a non-quadratic matrix C_m with the property of $C_m C_m^{\#} C_m = C_m$ and is not unique.

The marginal becomes:

$$p(\mathbf{x}_p) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}|C_m^T \Sigma^{-1} C_m|^{1/2}} \int \mathcal{N}(\mathbf{x}_m; \mu_1, \Sigma_1)\ d\mathbf{x}_m \tag{C.4}$$

If $\mathbf{x}_m$ is unbounded the integral above is indefinite and vanishes (evaluates to 1). Note that the contributions of the known features $\mathbf{x}_p$ are also absent from the remaining form (being absorbed in $\mu_1 = -C_m^{\#}(C_p \mathbf{x}_p - \mu)$). This is unsatisfactory, but expected – a linear combination of a known and unknown values can take any value in $(-\infty, \infty)$.

If the integral is definite with a rectangular hyper-volume of integration, it can not be evaluated conveniently. The covariance matrix Σ_1 is not diagonal (as C_m in $C_m^T \Sigma^{-1} C_m$ is a non diagonal matrix), preventing factorisation of the p.d.f. and evaluation of the multidimensional integral as a product of one dimensional integrals. Numerical methods that evaluate the integral by sampling the space do exist (see Genz (1993) and references therein; the lower and higher bounds on the integral from Eq. (C.4) decompose to a difference of multivariate normal probabilities). But they can hardly be considered suitable in the context of an HMM system where the probability in Eq. (C.4) is calculated for every frame and for every state.

Appendix D

Efficient calculation of the likelihood of the MD model with factorisable probability functions

The missing data model for speech recognition (Section 5.3) depends on summation of the partial likelihood over all possible masks, weighed by their respective probability (Eqs. 5.2 and 5.3, pp. 69):

$$W^* \approx \underset{W}{argmax} \sum_{all\ M} P(O|M, Q^*, W)P(M|Q^*, W)P(Q^*|W)P(W) \tag{D.1}$$

or

$$W^* \approx \underset{W}{argmax} \sum_{all\ M} P(O|M, Q^*, W)P(Q^*|W)P(M|W)P(W) \tag{D.2}$$

Under the i.i.d. assumptions the probabilities of the sequence of vectors decompose into products of the individual vectors probabilities. Still, it is not clear how to efficiently calculate the weighted sum over all possible masks in every time frame.

Fortunately, if both $p(\mathbf{o}(t)|\mathbf{m}(t), q^*(t), W)$ and $P(\mathbf{m}(t)|q^*(t), W)$ (or $P(\mathbf{m}(t)|W)$) are factorisable or weighted sums of factorisable distributions, an efficient calculation of the sum over all masks is possible. Lets consider the case when the state p.d.f. $p(\mathbf{o}(t)|\mathbf{m}(t), q^*(t), W)$ is a weighted sum of product of factors (ex: a mixture of Gaussians with diagonal covariance matrices) and $P(\mathbf{m}(t)|q^*(t), W)$ (or $P(\mathbf{m}(t)|W)$) is factorisable probability distribution function:

$$p(\mathbf{o}(t)|\mathbf{m}(t), q^*(t), W) = \sum_k P(k)p(\mathbf{o}(t)|k, \mathbf{m}(t), q^*(t), W) = \sum_k P(k) \prod_i p_i(o_i(t)|k, m_i(t), q^*(t), W)$$

where k indexes the mixture components, while i indexes the features in the feature vector. The mask can take only two values, 0 and 1 (0 meaning the respective feature is missing, 1 the feature is present):

$$P(\mathbf{m}(t)|q^*(t), W) = \prod_i P_i(m_i(t)|q^*(t), W)$$

(the second variant with the mask M in $P(M|W)$ dependent on the model, but not the state on the path is analogous)

Bernoulli (Carreira-Perpiñán and Renals, 2000) distributed mask probability:

$$P(\mathbf{m}(t)|q^*(t), W) = \prod_i \mu_{i|q^*(t),W}^{m_i(t)} (1 - \mu_{i|q^*(t),W})^{m_i(t)}$$

is one possible distribution satisfying the assumptions above.

Lets denote (dropping the conditioning on state and word, and disregarding the time index t):

$$\begin{aligned} \alpha_k &= P(k), \\ a_{ki} &= p_i(o_i(t)|k, m_i(t)=1, q^*(t), W), \\ b_{ki} &= p_i(o_i(t)|k, m_i(t)=0, q^*(t), W) \quad \text{and} \\ r_i &= p_i(m_i(t)=1|q^*(t), W) \\ \textbf{(consequently} \quad 1-r_i &= p_i(m_i(t)=0|q^*(t), W) \ \textbf{)} \end{aligned} \tag{D.3}$$

The core of the likelihood calculation in Eq. (D.1) (and Eq. (D.2)) is form of type:

$$\begin{aligned} F &= \sum_{all\ \mathbf{m}(t)} \left\{ \sum_k P(k) \prod_i p_i(o_i(t)|k, m_i(t), q^*(t), W) \right\} \left\{ \prod_i P_i(m_i(t)|q^*(t), W) \right\} \\ &= \sum_{all\ \mathbf{m}(t)} \sum_k \left\{ P(k) \prod_i p_i(o_i(t)|k, m_i(t), q^*(t), W) \prod_i P_i(m_i(t)|q^*(t), W) \right\} \\ &= \sum_k \sum_{all\ \mathbf{m}(t)} \left\{ P(k) \prod_i p_i(o_i(t)|k, m_i(t), q^*(t), W) \prod_i P_i(m_i(t)|q^*(t), W) \right\} \\ &= \sum_k P(k) \underbrace{\sum_{all\ \mathbf{m}(t)} \left\{ \prod_i p_i(o_i(t)|k, m_i(t), q^*(t), W) P_i(m_i(t)|q^*(t), W) \right\}}_{S} \end{aligned} \tag{D.4}$$

We will prove by induction that for the inner sum over all possible masks $\mathbf{m}(t)$ the following is true:

$$\begin{aligned} S &= \sum_{all\ \mathbf{m}(t)} \left\{ \prod_i p_i(o_i(t)|k, m_i(t), q^*(t), W) P_i(m_i(t)|q^*(t), W) \right\} \\ &= \prod_i \sum_{m_i(t)=\{0,1\}} p_i(o_i(t)|k, m_i(t), q^*(t), W) P_i(m_i(t)|q^*(t), W) \end{aligned} \tag{D.5}$$

Let N be the number of features in the feature vector. Using the notation from (D.3), previous Eq. (D.5) can be rewritten as:

$$S_N = \underbrace{\sum_{j=0}^{2^N-1}}_{all\ \mathbf{m}} \prod_{i=1}^{N} c_{kij} s_{ij} = \prod_{i=1}^{N} [b_{ki}(1-r_i) + a_{ki} r_i] \tag{D.6}$$

where $c_{kij} = a_{ki}$ and $s_{ij} = r_i$ iif j has 1 at the j-th bit (counting from left – leftmost bit is 1st, while the rightmost bit is N-th) in its N bits long binary representation., otherwise $c_{kij} = b_{ki}$ and $s_{ij} = 1 - r_i$.

The case of $N = 1$ is obvious:

$$\begin{aligned} S_1 &= \sum_{j=0}^{1} c_{k1j} s_{1j} = \underbrace{b_{k1}(1-r_1)}_{j=0} + \underbrace{a_{k1} r_1}_{j=1} \\ &= b_{k1}(1-r_1) + a_{k1} r_1 \end{aligned} \tag{D.7}$$

For $N = 2$ we have:

$$\begin{aligned}
S_2 &= \sum_{j=0}^{3} \{c_{k1j} c_{k2j} s_{1j} s_{2j}\} \\
&= \underbrace{b_{k1} b_{k2} (1 - r_1)(1 - r_2)}_{j=00} + \underbrace{b_{k1} a_{k2} (1 - r_1) r_2}_{j=01} + \underbrace{a_{k1} b_{k2} r_1 (1 - r_2)}_{j=10} + \underbrace{a_{k1} a_{k2} r_1 r_2}_{j=11} \\
&= b_{k1}(1 - r_1)[b_{k2}(1 - r_2) + a_{k2} r_2] + a_{k1} r_1 [b_{k2}(1 - r_2) + a_{k2} r_1] \\
&= [b_{k1}(1 - r_1) + a_{k1} r_1][b_{k2}(1 - r_2) + a_{k2} r_2] \\
&= \prod_{i=1}^{2} [b_{ki}(1 - r_i) + a_{ki} r_i]
\end{aligned} \tag{D.8}$$

Lets assume that our claim (Eq. (D.6)) is true for $N = n$:

$$S_n = \sum_{j=0}^{2^n - 1} \left\{ \prod_{i=1}^{n} c_{kij} s_{ij} \right\} = \prod_{i=1}^{n} [b_{ki}(1 - r_i) + a_{ki} r_i] \tag{D.9}$$

For $N = n + 1$ we have:

$$\begin{aligned}
S_{n+1} &= \sum_{j=0}^{2^{(n+1)} - 1} \left\{ \prod_{i=1}^{n+1} c_{kij} s_{ij} \right\} \\
&= \sum_{j=0}^{2^n - 1} \left\{ \prod_{i=1}^{n} c_{kij} s_{ij} \right\} b_{k(n+1)} (1 - r_{(n+1)}) + \sum_{j=0}^{2^n - 1} \left\{ \prod_{i=1}^{n} c_{kij} s_{ij} \right\} a_{k(n+1)} r_{(n+1)} \\
&\text{(since for half of the elements in the sum } j \text{ has highest bit 0, and for the other half 1)} \\
&= b_{k(n+1)} (1 - r_{(n+1)}) \sum_{j=0}^{2^n - 1} \left\{ \prod_{i=1}^{n} c_{kij} s_{ij} \right\} + a_{k(n+1)} r_{(n+1)} \sum_{j=0}^{2^n - 1} \left\{ \prod_{i=1}^{n} c_{kij} s_{ij} \right\} \\
&= [b_{k(n+1)} (1 - r_{(n+1)}) + a_{k(n+1)} r_{(n+1)}] \sum_{j=0}^{2^n - 1} \left\{ \prod_{i=1}^{n} c_{kij} s_{ij} \right\} \\
&\text{(using the induction assumption for } N = n) \\
&= [b_{k(n+1)} (1 - r_{(n+1)}) + a_{k(n+1)} r_{(n+1)}] \prod_{i=1}^{n} [b_{ki}(1 - r_i) + a_{ki} r_i] \\
&= \prod_{i=1}^{n+1} [b_{ki}(1 - r_i) + a_{ki} r_i]
\end{aligned} \tag{D.10}$$

which is exactly Eq. (D.6) for $N = n + 1$.

The sum is as convenient as it is intuitive: the sum of the data likelihoods over all possible masks weighted by the probability of each mask can be computed using the sum of the individual features likelihoods, each weighted by the probability of being present or missing. The assumptions used are that both the data likelihood and the mask likelihood are weighted sums of factorisable distributions. They are observed in a typical HMM system.

Appendix E

Attributions

The work reported in this thesis has been done in collaboration with several members of the Speech and Hearing Group (SPandH) in the Department of Computer Science at the Sheffield University, UK, during the course of several years. A. Morris did a lot of the earlier work on missing data ASR. A. Vizinho was instrumental in the work on separation using local SNR based estimation. M. Cooke was the first to realise that the large number of insertions appearing in the frames with little reliable data can be solved by using bounded marginalisation instead of mere marginalisation. He also suggested a mask reliability measure derived from the local SNR estimate via a sigmoid mapping. J. Barker not only provided the tools (the CTK) for the latter set of the experiments, but also verified many of the results via separate and independent experiments. P. Green was also fully involved in all aspects of the work reported in this thesis.

Portions of the work reported in this thesis have been published in the following papers:

- P.D. Green, J. Barker, M.P. Cooke, and L. Josifovski. Handling missing and unreliable information in speech recognition. In *In Proc. of 8th Int. Workshop on AI and Statistics*, pages 49–56, Key West, Florida, jan 2001.

- M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, 34 (3): 267–285, jun 2001.

- J. Barker, L. Josifovski, M.P. Cooke, and P.D. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proc. ICSLP*, pages 373–376, 2000.

- A.C. Morris, L. Josifovski, H. Bourlard, M.P. Cooke, and P.D. Green. A neural network for classification with incomplete data: application to robust ASR. In *Proc. ICSLP*, volume 1, pages 409–412, 2000.

- A. Vizinho, P. Green, M. Cooke, and L. Josifovski. Missing data theory, spectral subtraction and signal–to–noise estimation for robust ASR: An integrated study. In *Proc. Eurospeech*, pages 2407–2410, 1999.

- 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*, volume 6, pages 2837–2840, sep 1999.

- M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust ASR with unreliable data and minimal assumption. In *Robust Methods for Speech Recognition in Adverse Conditions*, pages 195–198, Tampere, Finland, may 1999.

- M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. Technical Report CS–99–05, Department of Computer Science, University of Sheffield, 1999.

Bibliography

A. Acero. *Acoustical and Environmental Robustness for Automatic Speech Recognition.* PhD thesis, ECE Department, CMU, 1990.

A. Acero, S. Altschuler, and L. Wu. Speech/noise separation using two microphones and a VQ model of speech signals. In *Proc. ICSLP*, volume 4, pages 532–535, 2000.

H. Agaiby, C. Fyte, S. McGlinchey, and T. J. Moir. Commercial speech recognisers performance under adverse conditions, a survey. In *Robust speech recognition using unknown communication channels*, pages 163–166. ESCA-NATO Tutorial and Research Workshop, apr 1997.

A. Agarwal and Y. M. Cheng. Two–stage M-el–warped Wiener filter for robust speech recognition. In *Automatic speech recognition and understanding workshop*, dec 1999.

S. Ahmad and V. Tresp. Some solutions to the missing feature problem in vision. In J. H. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advanes in Neural Information Processing Systems 5*, pages 393–400. Morgan Kaufmann, San Mateo, CA, 1993.

K. Aikawa, H. Singer, H. Kawahara, and Y. Tohkura. Cepstral representation of speech motivated by time–frequency masking: An application to speech recognition. *Journal of Acoustical Society of America*, 10(1):603–614, jul 1996.

J.B. Allen. How do humans process and recognize speech. *IEEE Transactions on Speech and Audio Processing*, 2:567–577, oct 1994.

J.B. Allen. From Lord Rayleigh to Shannon: How do we decode speech? In *Proc. ICASSP*, may 2002. URL `http://auditorymodels.org/jba/PAPERS/ICASSP`.

C. Avendano and H. Hermansky. On the properties of temporal processing for speech in adverse environments. In *Proceedings of WASPA'97*, 1997.

D. Azzopardi, S. Semnani, B. Milner, and R. Wiseman. Improving accuracy of telephone–based, speaker–independent speech recognition. In *Proc. ICSLP*, pages 301–304, 1998.

J. Baker, P. Bamberg, L. Gillick, L. Lamel, R. Roth, F. Scattone, and D. Sturtevant. Dragon systems resource management benchmark results–February 1991. In *DARPA speech and natural language workshop*, pages 59–64, feb 1991.

J. Barker. *The rleationship between speech peerception and auditory organisation: Studies with spectrally reduced speech.* PhD thesis, Department of Computer Science, University of Sheefield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK, 1998.

J. Barker. User's guide and reference manual for the RESPITE CASA toolkit project, 2000. URL `http://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/ctk/`.

J. Barker and M. Cooke. Modeling the recognition of spectrally reduced speech. In *Proc. Eurospeech*, pages 2127–2130, 1997.

J. Barker, M. Cooke, and D. Ellis. Decoding speech in the presence of other sound sources. In *Proc. ICSLP*, volume 4, pages 270–273, 2000.

J. Barker, M.P. Cooke, and D. Ellis. Combining bottom–up and top–down constraints for robust ASR: the multisource decoder. In *Proc. Eurospeech*, 2001a.

J. Barker, P.D. Green, and M.P. Cooke. Linking auditory scene analysis and robust ASR by missing data techniques. In *Proc. Institute of Acoustics*, 2001b.

D.C. Bateman, D.K. Bye, and M.J. Hunt. Spectral contrast normalization and other techniques for speech recognition in noise. In *Proc. ICASSP*, volume 1, pages 241–244, 1992.

V. L. Beattie and S. J. Young. Hidden Markov Model state–based noise cancellation. Technical Report TR92, Cambridge University Engineering Department, feb 1992.

A. J. Bell and T. J. Sejnowski. An information–maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1004–1034, 1995.

Y. Bengio and F. Gingras. Recurrent neural networks for missing and asynchronous data. In *Advanes in Neural Information Processing Systems 8*. MIT Press, 1996.

M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. ICASSP*, pages 208–211, 1979.

F. Berthomier, H. Glotin, E. Tessier, and H. Bourlard. Interfacing of CASA and partial recognition based on a multistream technique. In *Proc. ICSLP*, volume 4, pages 1415–1419, 1998.

C. M. Bishop. *Neural networks for pattern recognition*. Clarendon press, Oxford, 1995.

S. F. Boll. Supression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech and signal processing*, 27(2):113–120, apr 1979.

H. Bourlard, S. Dupont, and C. Ris. Multi–stream speech recognition. Technical Report IDIAP–RR 96–07, IDIAP, Martigny, Valais, Switzerland, dec 1996.

H.A. Bourlard and N. Morgan. *Connectionist speech recognition: A hybrid approach*. Kluwer Academic, Boston, London, 1993.

A. S. Bregman. *Auditory scene analysis*. MIT Press, 1990.

M. K. Brendborg and B. Lindberg. Noise robust recognition using feature selective modeling. In *Proc. Eurospeech*, pages 295–298, 1997.

J.S. Bridle, K.M. Ponting, M.D. Brown, and A.W. Borrett. A noise compensation spectrum distance measure applied to automatic speech recognition. In *Institute of Acoustics, Autumn Meeting, Windermere*, nov 1984.

G.J. Brown. *Computational auditory scene analysis: A representational approach*. PhD thesis, Department of Computer Science, University of Sheffield, 1992.

G.J. Brown and M. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8:297–336, 1994.

G.J. Brown, D.L. Wang, and J. Barker. A neural oscillator sound separator for missing data speech recognition. In *Proc. IJCNN*, 2001.

J.-F. Cardoso. Multidimensional independant components analysis. In *Proc. ICASSP*, pages 1941–1944, 1998.

J.F. Cardoso. Estimating equations for source separation. In *Proc. ICASSP*, pages 3449–3452, apr 1997.

M.Á. Carreira-Perpiñán. Mode–finding for mixtures of gaussian distributions. Technical Report CS–99–03, Department of Computer Science, University of Sheffield, 1999.

M.Á. Carreira-Perpiñán and S. Renals. Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12(1):141–152, January 2000. URL `http://www.dcs.shef.ac.uk/~miguel/papers/mix_bernoulli.html`.

C. Cerisara, J.-P. Haton, and D. Fohr. A recombination model for multi–band speech recognition. In *Proc. ICASSP*, pages 717–720, 1998.

J.-T. Chien, H.-C. Wang, and L.-M. Lee. A novel projection–based likelihood measure for noisy speech recognition. *Speech communication*, 24, 1998.

S. Choi, Y. Lyu, F. Berthommier, H. Glotin, and A. Cichocki. Blind separation of delayed and superimposed acoustic sources: learning algorithm and experimental study. In *Proc. ICP*, pages 109–114, Seoul, sep 1999.

S.M. Chu and Y. Zhao. Robust speech recognition using discriminative stream weighting and parameter interpolation. In *Proc. ICSLP*, pages 1423–1426, 1998.

R. Cole, K. Roginski, and M. Fanty. A telephone speech database of spelled and spoken names. In *Proc. ICSLP*, volume 2, pages 891–895, 1992.

R. Comerford, J. Makhoul, and R. Shwartz. The voice of the computer is heard in the land (and it listens, too!). *IEEE Spectrum*, pages 34–47, December 1997.

D. Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer speech and language*, 3:151–167, 1989a.

D. Van Compernolle. Spectral estimation using a log–distance error criterion applied to speech recognition. In *Proc. ICASSP*, volume 1, pages 258–261, may 1989b.

M. Cooke, M. Crawford, and P. Green. Learning to recognize speech in noisy environments. In *ATR Workshop on "Biological foundations for speech perception and production"*, Osaka, sep 1994a.

M. Cooke and D.P.W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech communication*, mar 1999. submitted.

M. Cooke, P. Green, C. Anderson, and D. Abberley. Recognition of occluded speech by Hidden Markov models. Technical Report TR–94–05–01, Department of Computer Science, University of Sheffield, may 1994b.

M. Cooke, P. Green, and M. Crawford. Handling missing data in speech recognition. In *Proc. ICSLP*, 1994c.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, 34(3):267–285, jun 2001.

M. Cooke, A. Morris, and P. Green. Recognising occluded speech. In *Proc. ESCA ETR Workshop on the auditory basis of speech perception*, pages 297–300, Keele, jul 1996.

M. P. Cooke. *Modelling auditory processing and organisation*. PhD thesis, Department of Computer Science, University of Sheffield, 1991. Published by Oxford University Press, 1993.

M.P. Cooke, G.J. Brown, M.D. Crawford, and P.D. Green. Computational auditory scene analysis: Listening to several things at once. *Endeavour*, 17:186–190, 1993.

C. Couvreur and H. Van Hamme. Model–based feature enhancement for noisy speech recognition. In *Proc. ICASSP*, volume 3, pages 1719–1722, 2000.

S. Crafa, L. Fissore, and C. Vair. Data–driven pmc and bayesean learning integration for fast model adaptation in noisy conditions. In *Proc. ICSLP*, pages 471–474, 1998.

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

A. de Cheveigne and H. Kawahara. Missing data model of vowel identification. *Journal of Acoustic Society of America*, 106(6):3497–3508, jun 1999.

J. de Veth, B. Cranen, and L. Boves. Acoustic backing–off in the local distance computation for robust automatic speech recognition. In *Proc. ICSLP*, pages 1427–1430, 1998.

J. de Veth, F. de Veth, B. Cranen, and L. Boves. Missing feature theory in ASR: Make sure you miss the right type of features. In *Robust Methods for Speech Recognition in Adverse Conditions*, pages 231–234, Tampere, Finland, may 1999.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39:1–38, 1977.

C. Dobrin, P. Haavisto, K. Laurila, and J. Astola. Speech recognition experiments in a noisy environment using auditory system modeling. In *Proc. Eurospeech*, pages 131–134, 1995.

L. Docio-Frnandez and C. Garcia-Mateo. Noise model selection for robust speech recognition. In *Proc. ICSLP*, pages 1431–1434, 1998.

S. Downey. An analysis of Wiener adaptation for speech recognition in adverse conditions. *Proceedings of the institute of acoustics*, 18(9):225–233, 1996.

J. Droppo, A. Acero, and L. Deng. Uncertain decoding with SPLICE for noise robust speech recognition. In *Proc. ICASSP*, 2002.

A. Drygajlo and M. El-Maliki. Speaker verification in noisy environment with combined spectral subtraction and missing feature theory. In *Proc. ICASSP*, volume 1, pages 121–124, 1998a.

A. Drygajlo and M. El-Maliki. Spectral subtraction and missing feature modeling for speaker verification. In *Signal Processing IX, Theories and Applications, EURASIP, Rhodes, Greece*, pages 355–358, 1998b.

A. Drygajlo and M. El-Maliki. Use of generalized spectral subtraction and missing feature compensation for robust speaker verification. In *Proc. Workshop on speaker recognition and its commercial and forensic applications, Avignon, France*, pages 80–83, apr 1998c.

A. Drygajlo, N. Virag, and G. Cosendai. Robust speech recognition in noise using speech enhancement based on the masking properies of the auditory system and adaptive HMM. In *Proc. Eurospeech*, pages 473–476, sep 1995.

R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* John Wiley & Sons, New York, 1973.

S. Dupont. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In *Proc. ICSLP*, pages 1439–1442, 1998.

S. Dupont and H. Bourlard. Using multiple time scales in a multi–stream speech recognition system. In *Proc. Eurospeech*, pages 3–6, 1997.

M. El-Maliki. *Speaker verification with missing features in noisy environments.* PhD thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, EPFL, 2000.

M. El-Maliki and A. Drygajlo. Missing feature detection and compensation for GMM–based speaker verification in noise. In *Proc. COST 250 Workshop on speaker recognition in telephony, Rome, Italy*, nov 1999.

D. Ellis. Speech recognition as a component in computational auditory scene analysis. In *unpublished*, 1998.

D. P. W. Ellis. *Prediction–driven computational auditory scene analysis.* PhD thesis, Department of Electrtical Engineering and Computer Science, M.I.T., 1996.

W.D. Ellis, editor. *A source book of Gestalt Psychiology.* Routledge & Kegan Paul Ltd, Brodway House, 68–74 Carter Lane, London, EC4, 1955.

Y. Ephraim and D. Malah. Speech enhancement using a minimum mean–square error short–time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32:1109–1121, dec 1984.

A. Erell and M. Weintraub. Energy conditioned spectral estimation for recognition of noisy speech. *IEEE transactions of speech and audio processing*, 1(1):84–89, jan 1993a.

A. Erell and M. Weintraub. Filterbank–energy estimation using mixture and Markov models for recognition of noisy speech. *IEEE transactions of speech and audio processing*, 1(1):68–76, jan 1993b.

J. A. Flores and S. J. Young. Adapting a HMM-based recogniser for noisy speech enhanced by spectral subtraction. In *Proc. Eurospeech*, volume 2, pages 829–832, 1993.

S. Furui. Speaker–independent isolated word recognition using dynamic features of the speech spectrum. *IEEE Transactions on acoustics, speech, and signal processing*, ASSP–34(1):52–59, feb 1986.

S. Furui. Recent advances in robust speech recognition. In *Robust speech recognition using unknown communication channels*, pages 11–20. ESCA-NATO Tutorial and Research Workshop, apr 1997.

M. J. F. Gales. *Model–based techniques for noise robust speech recognition.* PhD thesis, Gonville and Caius College, University of Cambridge, 1995.

M. J. F. Gales. "NICE" model–based compensation schemes for robust speech recognition. In *Robust speech recognition using unknown communication channels*, pages 55–64. ESCA-NATO Tutorial and Research Workshop, apr 1997.

M. J. F. Gales and S. J. Young. An improved approach to the Hidden Markov model decomposition of speech and noise. In *Proc. ICASSP*, volume 1, pages 233–236, 1992.

Y. Gao and J.-P. Haton. Noise reduction and speech recognition in noise conditions tested on LPNN-based continuous speech recognition system. In *Proc. Eurospeech*, volume 2, pages 1035–1038, 1993.

Y. Gao, T. Huang, S. Chen, and J.-P. Haton. Auditory model based speech processing. In *Proc. ICSLP*, pages 73–76, 1992.

P. N. Garner and W. J. Holmes. On the robust incorporation of robust features into Hidden Markov models for automatic speech recognition. In *Proc. ICASSP*, pages 1–4, 1998.

J.S. Garofolo and D.S. Pallet. Use of the CD–ROM for speech database storage and exchange. In *Proc. Europaen Conference on Speech Communication and Technology*, pages 309–315, 1989.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesean data analysis.* Chapman & Hall, 2–6 Boundary Row, London SE1 8HN, UK, 1995.

A. Genz. Numerical computation of multivariate normal probabilities. *Jorunal of Comp. Gaph. Stat.*, 1:141–149, 1992.

A. Genz. Comparison of methods for the computation of multivariate normal probabilities. *Computing science and statistics*, 25:400–405, 1993.

Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an em approach. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advanes in Neural Information Processing Systems 6*, pages 120–129. Morgan Kaufmann, San Mateo, CA, 1994a.

Z. Ghahramani and M.I. Jordan. Learning from incomplete data. Technical Report A.I. Memo No. 1509 and C.B.C.L. Paper No. 108, Artificial Intellegence Labaratory and Center for bilogical and computational learning, Department of brain and cognitive sciences, MIT, dec 1994b. URL `http://www.ai.mit.edu/publications/pubsDB/pubsDB/onlinehtml`.

O. Ghitza. Auditory nerve representation as a front–end for speech recognition in a noisy environment. *Computer speech and language*, 1:109–130, 1986.

D. Godsmark and G.J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech communication*, 27:351–336, 1999.

Y. Gong. Speech recognition in noisy environments. *Speech communication*, 16:261–291, 1995.

M. Graciarena. Maximum likelihood noise HMM estimation in model–based robust speech recognition. In *Proc. ICSLP*, pages 598–601, 2000.

P. D. Green, M. P. Cooke, and M. D. Crawford. Auditory scene analysis and Hidden Markov Model recognition of speech in noise. In *Proc. ICASSP*, pages 401–404, 1995.

S. Greenberg. Auditory function. In Encyclopedia of acoustics, editor, *M.J. Crocker*, pages 1301–1323. John Wiley & Sons, 1997.

S. Greenberg and E.D. Kingsbury. the modulation spectrogram: in pursuit of an invariant representation of speech. In *Proc. ICASSP*, volume 3, pages 1647–1650, 1997.

A. Hagen, A. Morris, and H. Bourlard. Subband–based speech recognmition in noise conditions: The full combination approach. Technical Report IDIAP–RR 15, IDIAP, Martigny, Valais, Switzerland, 1998.

A. Hagen, A. Morris, and H. Bourlard. From multi–band full combination to multi–stream full combination processing in robust ASR. In *ISCA ITRW ASR2000*, sep 2000.

J. Hakkinen, S. Suontausta, R. Hariharan, M. Vasilache, and K. Laurila. Improved feature vector normalization for noise robust connected speech recognition. In *Proc. Eurospeech*, pages 2833–2836, sep 1999.

J. H. L. Hansen. Morphological constrained geature enhancement with adaptive ceptral compensation (MCE-ACC) for speech recognition in noise and Lombard effect. *IEEE Transactions on speech and audio processing*, 2(4):598–614, oct 1994.

J. H. L. Hansen and L. M. Arslan. Robust feature–estimation and objective quality assessment for noisy speech recognition using the credit card corpus. *IEEE Transactions on speech and audio processing*, 3(3):169–184, may 1995.

B. A. Hanson and T. H. Applebaum. Features for noise–robust speaker–independent word recognition. In *Proc. ICSLP*, volume 2, pages 1117–1120, 1990.

R. Hariharan, I. Kiss, O. Vikki, and J. Tian. Multi–resolution front–end for noise robust speech recognition. In *Proc. ICSLP*, volume 3, pages 550–553, 2000.

H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *JASA*, 87(4):1738–1752, apr 1990.

H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE transactions on speech and audio processing*, 2(4):578–589, oct 1994.

H. Hermansky, N. Morgan, A. Bayya, and P. Kholn. Compensation for the effect of the communication channel on auditory–like analysis of speech (RASTA–PLP). In *Proc. Eurospeech*, pages 1367–1370, 1991.

J. Hernando and C. Nadeu. A comparative study of parameters and distances for noisy speech recognition. In *Proc. Eurospeech*, volume 1, pages 91–94, 1991.

G. Hirsch and D. Pearce. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000*, pages 181–188, sep 2000.

H. G. Hirsch. Estimation of noise spectrum and it's application to SNR-estimation and speech enhancement. Technical Report TR-93-012, ICSI, Berkeley, CA, 1993.

H. G. Hirsch and C. Enrichter. Noise estimation for robust speech recognition. In *Proc. ICASSP*, pages 153–156, 1995.

H. G. Hirsch, P. Meyer, and H. W. Ruehl. Improved speech recognition using high–pass filtering of subband enevelopes. In *Proc. Eurospeech*, pages 413–416, 1991.

J.N. Holmes and N.C. Sedgwick. Noise compensation for speech recognition using probabilistic models. In *Proc. ICASSP*, pages 741–844, 1986.

M. Hunke, M. Hyun, S. Love, and T. Holton. Improving the noise and spectral robustness of an isolated–word recognizer using an auditory–model front end. In *ICSPL'98*, pages 475–478, 1998.

M. Hunt. Spectral signal processing for ASR. In *Proc. International Workshop on Automatic Speech Recognition and Understanding*, dec 1999.

A. Hyvarinen. Survey on independent componnent analysis. *Neural computing surveys*, 2:94–128, 1999. URL `http://www.icsi.berkeley.edu/~jagota/NCS/`.

S. Ikeda and N. Murata. A method of ICA in time–frequency domain. In *International workshop on independant components analysis and blind signal separation*, pages 365–371, jan 1999.

ISO/IEC 11172-3. *Coding of Moving pictures and associated audio for digital storage media at up to 1.5 Mbit/s - Audio Part*. International Standard, 1993.

ISO/IEC 13818-3. *Information Technology: Generic coding of Moving pictures and associated audio – Audio Part*. International Standard, 1995.

J.-C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *JASA*, 1:510–524, 1993.

J.-C. Junqua, S. Fincke, and K. Field. Influence of the speaking style and the noise spectral tilt on the lombard reflex and automatic speech recognition. In *Proc. ICSLP*, pages 467–470, 1998.

M. Kadirkamanathan. Hidden Markov Model decomposition recognition of speech in noise: a comprehensive experimental study. In *ESCA workshop on speech processing in adverse conditions*, pages 187–190, Cannes, France, 1992.

M. Kadirkamanathan and A. P. Varga. Simultaneous model re-estimation from contaminated data by "Composed Hidden Markov Modelling". In *Proc. ICASSP*, pages 897–900, 1991.

S. Kajarekar, N. Malayath, and H. Hermansky. Analysis of speaker and channel variability in speech. In *International Workshop on Automatic Speech Recognition and Understanding*, dec 1999.

N. Kanedera, T. Arai, H. Hermansky, and M. Pavel. On the importance of various modulation frequencies for speech recognition. In *Proc. Eurospeech*, pages 1079–1082, 1997.

N. Kanedera, H. Hermansky, and T. Arai. On properties of modulation spectrum for robust automatic speech recognition. In *Proc. ICASSP*, pages 613–616, 1998.

H.J. Kappen and M.J. Nijman. Radial basis Boltzman machines and learning with missing values. In *Proc. World Congress on Neural Networks, Washington DC, USA*, pages 72–75, 1995. URL `ftp://galba.mbfys.kun.nl/Kappen.RBBM.ps.Z`.

D. Katz. *Gestalt Psychology.* Methuen & Co. Ltd., 36 Essex Street, London WC2, 1951.

C. Kermorvant. A comparison of noise reduction techniques for robust speech recognition. Technical Report 99–10, IDIAP, Martigny, Valais, Switzerland, jul 1999.

D.Y. Kim, C.K. Un, and N.S. Kim. Speech recognition in noisy environments using first–order vector taylor series. *Speech Communication*, 24:39–49, 1998.

B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech communication*, 25(1–3):117–132, 1998.

T. Kitamura, S. Ando, and E. Hayahara. Speaker–independent spoken digit recognition in noisy environments using dynamic spectral features and neural networks. In *Proc. ICSLP*, volume 1, pages 699–702, 1992.

D. H. Klatt. A digital filterbank for spectral matching. In *Proc. ICASSP*, pages 573–578, 1976.

T. Kobayashi, T. Kanno, and S. Imai. Generalized cepstral modeling of speech degraded by additive noise. In *Proc. Eurospeech*, volume 1, pages 609–612, 1993.

K Koffka. *Principle of Gestalt Psychiology.* Harcourt, Brace and World, New York, 1935.

W. Kohler. *Gestalt Psychology.* Liveright, New York, 1947.

D. Kryze, L. Rigazio, T. Applebaum, and J.-C. Junqua. A new noise–robust subband fron–end and its comparison to plp. In *International Workshop on Automatic Speech Recognition and Understanding*, dec 1999.

F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Plaveway, and R. Schwartz. Byblos speech recognition benchmark results. In *DARPA speech and natural language workshop*, pages 77–82, feb 1991.

R. Kuhn, P. Nguyen, J.-C. Jinqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eignenvoices for speaker adaptation. In *Proc. ICSLP*, pages 1771–1774, 1998.

C.-H. Lee. On feature and model compensation approach to robust speech recognition. In *Robust speech recognition using unknown communication channels*, pages 45–54. ESCA-NATO Tutorial and Research Workshop, apr 1997.

T.-W. Lee. *Independent component analysis: Theory and applications.* Kluwer Academic Publishers, P.O. Box 17, 3300 AA Dordrecht, The Netherlands, 1998.

T.-W. Lee, A.J. Bell, and R. Orglmeister. Blind source separation of real world signals. In *IEEE International conference on neural networks*, pages 2129–2135, Houston, June 1997.

R.G. Leonard. A database for speaker–independent digit recognition. In *Proc. ICASSP*, pages 111–114, 1984.

L. Lewin. *Dilogarithms and associated functions.* MacDonald & Co., London, 1958.

K. Linhard and T. Haulick. Spectral noise subtraction with recursive gain curves. In *Proc. ICSLP*, pages 1479–1482, 1998.

K. Linhard and H. Klemm. Noise reduction with spectral subtraction and median filtering for suppression of musical tones. In *Robust speech recognition using unknown communication channels*, pages 159–162. ESCA-NATO Tutorial and Research Workshop, apr 1997.

R. Lippmann. Speech perception by humans and machines. In *ESCA Workshop on the Auditory Basis of Speech Perception*, pages 309–316, 1996.

R. P. Lippmann and B. A. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In *Proc. Eurospeech*, pages 37–40, 1997.

R.J.A. Little. Regression with missing X's: A review. *Journal of American Statistical Association*, 87(420):1227–1237, dec 1992.

R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Wiley, New York, 1997.

P. Lockwood and J. Boudy. Experiments with a non-linear spectral subtractor (NSS) Hidden Markov Models and the projection, for robust speech recognition in cars. In *Proc. Eurospeech*, volume 1, pages 79–82, 1991.

B. Logan. *Adaptive model based speech enhancement*. PhD thesis, Univeristy of Cambridge, 1998.

B. Logan and T. Robinson. A practical perceptual frequency autoregressive HMM enhancement system. In *Proc. ICSLP*, pages 2815–2818, 1998.

R. Martin. An efficient algorithm to estimate the instantaneous snr of speech signal. In *Proc. Eurospeech*, pages 1093–1096, 1993.

R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512, jul 2001.

D. Matrouf and J. L. Gauvain. Model compensation for additive and covolutive noises in training and test data. In *Robust speech recognition using unknown communication channels*, pages 207–210. ESCA-NATO Tutorial and Research Workshop, apr 1997.

R.J. McAulay and M.L. Malpass. Speech enhancement using a soft–decision noise suppression filter. *IEEE Transactions on acoustics, speech and signal processing*, 28(2):137–145, apr 1980.

P. McCourt, S. Vaseghi, and N. Harte. Multi–resolution cepstral features for phoneme recognition across speech sub–bands. In *Proc. ICASSP*, pages 557–560, 1998.

B.A. Mellor and A.P. Varga. Noise masking in transform domain. In *Proc. ICASSP*, volume 2, pages 87–90, 1993.

N. Merhav and C.-H. Lee. A minimax classification approach with application to robust speech recognition. *IEEE transactions on speech and audio processing*, 1(1):90–100, jan 1993.

J. M. Meyer, K. U. Simmer, and K. D. Kammeyer. Comparason of one– and two–channel noise–estimation techniques, sep 1999. URL `http://www.comm.uni--bremen.de/pub/speech`.

B. Milner. A generalized approach for inclusion of tempral information into features for speech recognition. *Proceedings of the institute of acoustics*, 18(9):217–224, 1996.

J. Ming, P. Jancovic, P. Hanna, D. Stewart, and F.J. Smith. Robust features selection using probabilistic UNION models. In *Proc. ICSLP*, volume 3, pages 546–549, 2000.

J. Ming and F.J. Smith. A probabilistic UNION model for sub–band based robust speech recognition. In *Proc. ICASSP*, pages 1787–1790, 2000.

J. Ming, D. Stewart, P. Hanna, and F.J. Smith. A probabilistic UNION model for partial and temporal corruption of speech. In *Automatic speech recognition and understanding workshop*, dec 1999.

N. Mirghafori and N. Morgan. Transmissions and transitions: a study of two common assumptions in multiband ASR. In *Proc. ICASSP*, volume 2, pages 713–716, 1998.

S. Mizuta and K. Nakajima. Optimal discriminative training for HMMs to recognize noisy speech. In *Proc. ICSLP*, volume 2, pages 1519–1522, 1992.

C. Mokbel, L. Barbier, Y. Kerlou, and G. Chollet. Word recognition in the car: adapting recognizers to the new environments. In *Proc. Eurospeech*, volume 1, pages 707–710, 1992.

C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monne, J. Simonin, and K. Bartkova. Towards improving asr robustness for psn and gsm telephone applications. *Speech communication*, 23: 141–159, 1997.

B. C. J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, 24/28 Oval Road, London NW1, 1982.

P. J. Moreno. *Speech recognition in noisy environments.* PhD thesis, ECE Department, CMU, 1996.

A. C. Morris, M. P. Cooke, and P. D. Green. Some solutions to the missing feature problem in data classification, with application to noise robust ASR. In *Proc. ICASSP*, pages 737–740, 1998.

H. Murveit, J. Butzberger, and M. Weintraub. Speech recognition in SRI's resource management and AIS systems. In *DARPA speech and natural language workshop*, pages 94–100, feb 1991.

Y.K. Muthasamy, R.A. Cole, and B.T. Oshika. The OGI multi-language telephone speech corpus. In *Proc. ICSLP*, volume 2, pages 895–898, 1992.

N. Iwahashi nad H. Pao, K. Minamino, and M. Omote. Stochastic features for noise robust speech recognition. In *Proc. ICASSP*, pages 633–636, 1998.

A. Nadas, D. Nahamoo, and M.A. Picheny. Speech recognition using noise adaptive prototypes. *IEEE Transactions on speech and audio processing*, 37(10):1495–1503, oct 1989.

C. Nadeu, P. Paches-Leal, and B.-H. Juang. Filtering of time sequences of spectral parameters for speech recognition. *Speech communication*, 22:315–332, 1997.

S. Nakamura, T. Akabane, and S. Hamaguchi. Robust word spotting in adverse car environments. In *Proc. Eurospeech*, volume 2, pages 1045–1048, 1993.

T. Nakatani, H.G. Okuno, M. Goto, and T. Ito. Multiagent based binaural sound stream segregation. In D.F. Rosenthal and H.G. Okuno, editors, *Computational auditory scene analysis*, pages 195–214. Lawrence Erlbaum Associates, Inc., New Jersey 07430, 1998.

S. Okawa, E. Bocchieri, and A. Potamianos. Multi–band speech recognition in noisy environments. In *Proc. ICASSP*, pages 641–644, 1998.

H.G. Okuno, S. Ikeda, and T. Nakatani. Combining independent component analysis and sound stream segragation. In *IJCAI CASA'99*, pages 92–98, 1999.

J. P. Openshaw and J. S. Mason. Noise robust estimate of speech dynamics for speaker recognition. In *Proc. ICSLP*, volume 2, pages 925–928, 1996.

M. Padmanabhan and M. Picheny. Towards super–human speech recognition. In *Proc. ISCA Tutorial and Research Workshop ASR2000: Challenges for the new Millennium*, pages 188–194, sep 2000.

K. K. Paliwal. Spectral subband centroid features for speech recognition. In *Proc. ICASSP*, pages 617–620, 1998.

D. S. Pallet, J. G. Fiscus, A. Martin, and M. A. Przybocki. 1997 broadcast news benchmark test results: english and non–english. In *DARPA broadcast news transcription and understanding workshop*, 1998. URL `http://www.nist.gov/speech/publications/darpa98`.

K.-Y. Park and H.-S. Kim. Narrowband to wideband conversion of speech using GMM based transformation. In *Proc. ICASSP*, volume 3, pages 1842–1846, 2000.

R. D. Patterson, T. R. Anderson, and M. Allerhand. The auditory image model as a preprocessor for spoken language. In *Proc. ICSLP*, pages 1395–1398, 1994.

D.B. Paul and J.M. Baker. The design for the Wall Street Journal–based CSR corpus. In *Proc. ICSLP*, volume 2, pages 899–902, 1992.

F.S. Perdigao and L.V. Sa. Auditory models as front–ends for speech recognition. In *NATO ASI on computational hearing*, pages 179–182, jul 1998.

S.D. Peters, P. Stubley, and J.-M. Valin. On the limits of speech recognition in noise. In *Proc. ICASSP*, volume 1, pages 365–368, 1999.

M. Phillips, J. Glass, J. Polifroni, and V. Zue. Collection and analyses of WSJ–CSR corpus at MIT. In *Proc. ICSLP*, volume 2, pages 907–910, 1992.

J. W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9): 1215–1247, sep 1993.

P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallet. The DARPA 1000-word resource management database for continouous speech recognition. In *Proc. ICASSP*, pages 651–654, 1988.

L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice Hall, Englewood Cliffs, New Jersey 07632, 1993.

B. Raj, E. Gouvea, and R. M. Stern. Cepstral compensation using statistical linearization. In *Robust speech recognition using unknown communication channels*, pages 131–138. ESCA-NATO Tutorial and Research Workshop, apr 1997.

B. Raj, R. Singh, and R. M. Stern. Inference of missing spectrographic features for robust speech recognition. In *Proc. ICSLP*, pages 1491–1494, 1998.

K. Rao and P. Yip. *Discrete Cosine Transform, Algorithms, Advantages, Applications*. Academic Press, 1990.

R.E. Remez, P.E. Rubin, S.M. Berns, J.S. Pardo, and J.M. Lang. On the pereceptual organization of speech. *Psychological review*, 101(1):129–156, 1994.

P. Renevey. *Speech recognition in noisy conditions using missing feature approach*. PhD thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, EPFL, 2000.

P. Renevey and A. Drygajlo. Missing feature hteory and parallel model combination for robust speech recognition. In *Robust Methods for Speech Recognition in Adverse Conditions*, pages 215–218, Tampere, Finland, may 1999.

P. Renevey and A. Drygajlo. Introduction of a reliability measure in missing data approach for robust speech recognition. In *Proc. EUSPICO'2000*, Tampere, Finland, sep 2000a.

P. Renevey and A. Drygajlo. Statistical estimation of unreliable features for robust speech recognition. In *Proc. ICASSP*, volume 3, pages 1731–1734, 2000b.

M.D. Richard and R.P. Lippmann. Neural network classifiers estimate Bayesean aposteriori probabilities. *Neural Computation*, 3:361–483, 1991.

C. Ris. Using artifical neural network to predict the mask for missing data. Personal communication, mar 2000. RESPITE: FPM bi-monthly report: Task 2.2.

C. Ris and S. Dupont. Assessing local noise level estimation methods: Application to noise robust asr. *Speech communication*, 34(1–2):141–158, 2001.

J. Roberts. Modification to piecewise LPC. Technical Report Working paper WP–21752, MITRE, may 1978.

A.J. Robinson, G.D. Cook, D.P.W. Ellis, E. Fosler-Lussier, S.J. Renals, and D.A.G. Williams. Connectionist speech recognition of broadcast news. *Speech communication*, 2000. submitted.

R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE transactions of speech and audio processing*, 2(2):245–257, apr 1994.

D.F. Rosenthal and H.G. Okuno, editors. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, New Jersey, 1998.

S. Roweis. One microphone source separation. In *Neural Information Processing Systems 13 (NIPS'00)*, 2000.

D.B. Rubin. *Multiple imputation in for nonresponse in surveys*. John Wiley, New York, 1987.

S. Sagayama and A. Kiyoami. Issues relating the future of asr for telecommunications applcications. In *Proc. ETRW*, pages 75–81, 1997.

R. Sarikaya and J. N. Gowdy. Subband based classification of speech under stress. In *Proc. ICASSP*, pages 569–572, 1998.

V. Schless and F. Class. SNR–Dependent flooring and noise overestimation for joint application of spectral subtraction and model combination. In *Proc. ICSLP*, pages 1495–1498, 1998.

M.L. Seltzer, B. Raj, and R.M. Stern. Classifier–based mask estimation for missing feature methods of robust speech recognition. In *Proc. ICSLP*, volume 3, pages 538–541, 2000.

A. Shankar and C.-H. Lee. Robust speech recognition based on stochastic matching. In *Proc. ICASSP*, pages 121–124, 1995.

L. Singh and S. Srdiharan. Speech enhancement using critical band spectral subtraction. In *ICSPL'98*, pages 2827–2830, 1998.

O. Siohan, Y. Gong, and J.-P. Haton. Noise adaptation using linear regression for continuous noisy speech recognition. In *Proc. Eurospeech*, pages 465–468, sep 1995.

V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spctreal subtraction and Wiener filtering. In *Proc. ICASSP*, volume 3, pages 1875–1878, 2000.

H.J.M. Steeneken. *On measuring and predictiong speech intelligibility*. PhD thesis, University of Amsterdam, 1992.

R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima. Signal processing for robust speech recognition. In C.-H. Lee and F. Soong, editors, *Speech recognition*, pages 351–378. Kluwer Academic Publishers, Boston, 1996.

R. M. Stern, B. Raj, and P. J. Moreno. Compensation for environmental degradation in automatic speech recognition. In *Robust speech recognition using unknown communication channels*, pages 33–42. ESCA-NATO Tutorial and Research Workshop, apr 1997.

B. Strope and A. Alwan. Robust word recognition using threaded spectral peaks. In *Proc. ICASSP*, pages 625–628, 1998.

T. Takiguchi, S. Nakamura, and K. Shikano. Speech recognition for a distant moving speaker based on hmm composition and separation. In *Proc. ICASSP*, volume 3, pages 1403–1406, 2000.

J. Tian, R. Hariharan, and K. Laurila. Noise–robust two stream auditory feature extraction method for speech recognition. In *Proc. ICSLP*, pages 991–994, 1998.

S. Tibrewala and H. Hermansky. Multi–band and adaptation approaches to robust speech recognition. In *Proc. Eurospeech*, pages 2619–2622, 1998.

K. Torkkola. Blind separation of delayed sources based on information maximization. In *Proc. ICASSP*, volume 6, pages 3509–3511, 1996.

V. Tresp, S. Ahmad, and R. Neuneier. Training neural networks with deficient data. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advanes in Neural Information Processing Systems 6*, pages 128–135. Morgan Kaufmann, San Mateo, CA, 1994.

V. Tresp, R. Neuneier, and S. Ahmad. Efficient methods for dealing with missing data in supervised learning. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advanes in Neural Information Processing Systems 7*, pages 689–696. MIT Press, Cambridge, MA, 1995.

M. Trompf, R. Richter, H. Eckhardt, and H. Hackbarth. Combination of distortion–robust feature extraction and neural noise reduction for ASR. In *Proc. Eurospeech*, volume 2, pages 1039–1042, 1993.

A.J. van der Kouwe, D.L. Wang, and G.J. Brown. A comparison of auditory and blind separation techniques for speech segregation. Technical Report OSU-CISRC-6/99-TR15, Departmenf of Computer and Information Science, The Ohio State University, Columbus, Ohio 43210-1277, 1999.

A. Varga, R. Moore, J. Bridle, K. Ponting, and M. Russel. Noise compensation algorithms for use with Hidden Markov model based speech recognition. In *Proc. ICASSP*, pages 481–484, 1988.

A. Varga and K. Ponting. Control experiments on noise compensation in Hidden Markov Model based continuous word recognition. In *Proc. Eurospeech*, volume 1, pages 167–170, 1989.

A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. ICASSP*, volume 2, pages 845–848, 1990.

A. P. Varga and R. K. Moore. Simultaneous recognition of concurent speech signals using Hidden Markov model decomposition. In *Proc. Eurospeech*, pages 1175–1178, 1991.

A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, Speech Research Unit, Defence Research Agency, Malvern, UK, 1992.

S. V. Vaseghi and B. P. Milner. Noise–adaptive Hidden Markov models based on Wiener filters. In *Proc. Eurospeech*, volume 2, pages 1023–1026, 1993.

S. V. Vaseghi and B. P. Milner. Noise compensation methods for Hidden Marko Model speech recognition in adverse environments. *IEEE transactions on speech and audio processing*, 5(1): 11–21, jan 1997.

O. Vikki and K. Laurila. Noise robust HMM–based speech recognition using segmental cepstral feature vector normalization. In *Robust speech recognition using unknown communication channels*, pages 107–110. ESCA-NATO Tutorial and Research Workshop, apr 1997.

N. Virag. Speech enhancement based on masking properties of the auditory system. In *Proc. ICASSP*, pages 796–799, 1995.

D.L. Wang and G.J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on neural networks*, 10(3):684–697, may 1999.

R. M. Warren, K. R. Riener, J. A. Bashford, and B. S. Brubaker. Spectral redundancy: Intelligibility of sentences heard through narrow spectral tilts. *Perception and Psychophysics*, 57(2): 175–182, 1995.

R.M. Warren. Perceptual restoration of missing speech sounds. *Science*, 167:392–393, 1970.

R.M. Warren, J.A. Bashford Jr., E.W. Healy, and B.S. Brubaker. Auditory induction: Reciprocal changes in the alternating sounds. *Pereception and Psychophysics*, 55(3):313–322, 1994.

M. Weintraub. *A theory and computational model of auditory monoaural sound separation.* PhD thesis, Department of Electrical Engineering, Stanford University, 1985.

K. F. Wong, S. H. Leung, and H. C. Ng. Noisy speech recognition using singular value decomposition and two –sided linear prediction. In *Proc. Eurospeech*, pages 1027–1030, 1993.

S.-K. Wong and B. Shi. A non–linear model transformation for ml stochastic matching in additive noise. In *Second workshop on multimedia signal processing*, pages 143–148, dec 1998.

H.-C. Wu, J. Principe, and D. Xu. Exploring the tempo–frequency micro–structure of speech for blind source separation. In *Proc. ICASSP*, volume 2, pages 1145–1148, 1998a.

S.-L. Wu, B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Performance improvements through combining phone and syllable–scale information in automatic speech recognition. In *Proc. ICSLP*, pages 459–462, 1998b.

F. Xie and D. V. Campernolle. Speech enhancement by nonlinear spectral estimation–a unifying approach. In *Proc. Eurospeech*, volume 1, pages 617–620, 1993.

R. Yang and P. Haavisto. Noise compensation for speech recognition in car noise environments. In *Proc. ICASSP*, pages 433–436, 1995.

R. Yang, M. Mjaniemi, and P. Haavisto. Dynamic parameter compensation for speech recognition in noise. In *Proc. Eurospeech*, pages 469–472, sep 1995.

B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy. Speech enhancement using linear prediction residual. *Speech Communication*, 28:25–42, 1999.

N.B. Yoma, F.R. McInnes, and M.A. Jack. Improving performance of spectral subtraction in speech recognition using a model for additive noise. *IEEE Transactions on speech and audio processing*, 6(6):579–582, nov 1998.

S. J. Young and P. C. Woodland. *HTK Version 1.5: User, reference and programmer manual.* Cambridge University Engineering Department, Speech Group, 1993.

S.J. Young, N.H. Russel, and J.H.S. Thornton. Token passing: A simple conceptual model for connected speech recognition systems. Technical Report TR38, Cambridge University Engineering Department, jul 1989.

K.-H. Yuo and H.-C. Wang. Robust features for noisy speech recognition based on temporal trajectory filtering of short–time autocorrelation sequences. *Speech Communication*, 28:13–24, 1999.

M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition. Technical Report No. CS99–1, University of New Mexico, Albuquerque, NM 87131, USA, jul 1999.