ELSEVIER

Speech Communication 34 (2001) 267–285

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Robust automatic speech recognition with missing and unreliable acoustic data

Martin Cooke *, Phil Green, Ljubomir Josifovski, Ascension Vizinho

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK

Received 30 June 1999; received in revised form 26 April 2000; accepted 1 May 2000

Abstract

Human speech perception is robust in the face of a wide variety of distortions, both experimentally applied and naturally occurring. In these conditions, state-of-the-art automatic speech recognition (ASR) technology fails. This paper describes an approach to robust ASR which acknowledges the fact that some spectro-temporal regions will be dominated by noise. For the purposes of recognition, these regions are treated as missing or unreliable. The primary advantage of this viewpoint is that it makes minimal assumptions about any noise background. Instead, reliable regions are identified, and subsequent decoding is based on this evidence. We introduce two approaches for dealing with unreliable evidence. The first – *marginalisation* – computes output probabilities on the basis of the reliable evidence only. The second – *state-based data imputation* – estimates values for the unreliable regions by conditioning on the reliable parts and the recognition hypothesis. A further source of information is the bounds on the energy of any constituent acoustic source in an additive mixture. This additional knowledge can be incorporated into the missing data framework. These approaches are applied to continuous-density hidden Markov model (HMM)-based speech recognisers and evaluated on the TIDigits corpus for several noise conditions. Two criteria which use simple noise estimates are employed as a means of identifying reliable regions. The first treats regions which are negative after spectral subtraction as unreliable. The second uses the estimated noise spectrum to derive local signal-to-noise ratios, which are then thresholded to identify reliable data points. Both marginalisation and state-based data imputation produce a substantial performance advantage over spectral subtraction alone. The use of energy bounds leads to a further increase in performance for both approaches. While marginalisation outperforms data imputation, the latter technique allows the technique to act as a preprocessor for conventional recognisers, or in speech-enhancement applications.

Keywords: Robust ASR; Missing data; Data imputation; HMM; Spectral subtraction

1. Introduction

1.1. The sensory occlusion problem

The identification of coherent objects within the undifferentiated mass of signals reaching sensory receptors is a challenging problem faced by any system or organism charged with making sense of

* Corresponding author. Tel.: +44-114-222-1800; fax: +44-114-222-1810.

E-mail addresses: m.cooke@dcs.shef.ac.uk (M. Cooke), p.green@dcs.shef.ac.uk (P. Green), l.josifovski@dcs.shef.ac.uk (L. Josifovski), a.vizinho@dcs.shef.ac.uk (A. Vizinho).

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00034-0

its environment. One aspect of this challenge is the need to function with missing or unreliable data. In vision, for instance, evidence for individual objects may be incomplete or fragmented due to occlusion. In recent years, a number of studies have examined the 'missing data' problem in computer vision (Ahmed and Tresp, 1993; Ghahramani and Jordan, 1994). The equivalent problem in audition has received far less attention because it is counter to intuition: while objects in a visual scene are predominantly opaque, acoustic signals combine additively. Consequently, techniques for robust automatic speech recognition (ASR) have been developed with the aim of achieving near-perfect allocation of the acoustic mixture into contributions from constituent sources. Examples of such approaches include hidden Markov model (HMM) decomposition (Varga and Moore, 1990), parallel model combination (Gales and Young, 1993) and blind separation (Comon, 1994; Bell and Sejnowski, 1995). However, these techniques presently have a number of limitations, motivating the alternative hypothesis – that incomplete data is a valid characterisation of the normal listening situation – which this paper takes as its starting point.

1.2. Arguments for missing data processing in the auditory system

Several researchers (Lippmann, 1997; Hermansky, 1998; Cooke and Green, in press) argue that the search for robust ASR has much to gain by examining the basis for speech perception in listeners. Lippmann (1997) presents evidence from a wide range of speech recognition tasks which indicates that human error rates are an order of magnitude smaller than those obtained by ASR algorithms for clean speech, and two orders of magnitude smaller for typical noise conditions. Further, for any given signal-to-noise ratio (SNR), listeners generally perform at higher levels of identification for non-stationary maskers (Miller and Licklider, 1950), whereas current robust ASR techniques favour, and often assume, stationary noise back-grounds.

Our proposal is underpinned by evidence that listeners routinely handle the missing data case in everyday sound source processing. Several arguments can be advanced in support of this claim.

1. *Listeners can cope with missing data*. Natural speech signals which have undergone deliberate spectro-temporal excisions typically show remarkably little decrease in intelligibility (Strange et al., 1983; Steeneken, 1992; Warren et al., 1995; Lippman, 1996). For example, normal conversation is viable for speech which has been high- or low-pass filtered with a cutoff frequency of 1800 Hz (Fletcher, 1953; Allen, 1994). Redundancy in the speech signal combats the missing data condition.

2. *The missing data condition occurs naturally*. Whilst the excisions referred to above are deliberate, there are counterparts in everyday listening, e.g. interfering signals, band-restricted transmission and channel noise over telephone lines.

3. *Masked data is effectively missing data*. The neural code for signal detection exhibits what has been called the 'capture effect' (Moore, 1997), in that locally more intense sound components dominate the neural response, whether defined in terms of firing rate or temporal response pattern. Locally weaker sound components do not contribute to the neuronal output: they are masked and therefore can be considered missing for the purposes of further processing. The everyday auditory scene commonly contains several active sound sources, any of which constitute a potential target for the listener's attention. Local spectro-temporal variations in target-to-background intensity will, via the capture effect, lead to incomplete evidence for each constituent source.

4. *The auditory nervous system handles simultaneous signals*. The auditory system must do more than process isolated speech sources. In Bregman's terms (Bregman, 1990), the 'auditory scene analysis' problem involves organising multiple, concurrent signals into associated perceptual streams. There must be a neural mechanism which enables subsequent processes to identify those components which have been grouped into the same stream. Solutions to this binding problem (von der Malsburg and Schneider, 1986; Liu et al., 1994; Brown et al., 1996; Brown and Wang, 1997) typically envisage phase synchronisation of sensory channels containing signal components deemed to

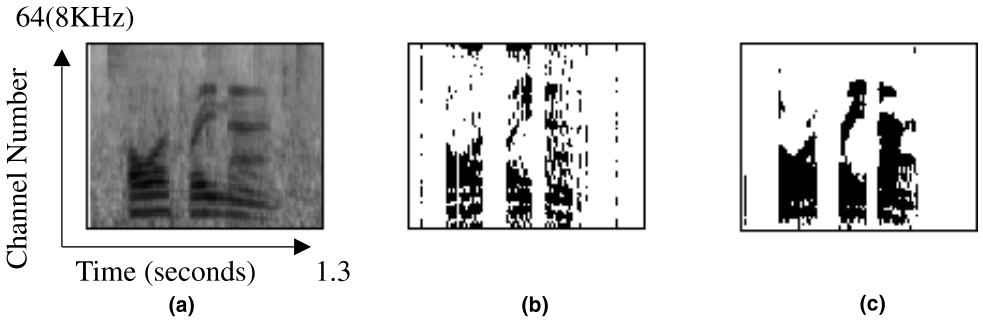

Fig. 1. Identifying reliable evidence by computational auditory scene analysis: (a) auditory spectrogram of the TIDigit sequence 'four three nine' in 10 dB factory noise; (b) reliable regions identified by harmonic cues; and (c) reliable regions identified by spatial location.

originate from the same object, which are desynchronised from channels responding to different objects. Processes sampling the neural array at any instant will receive an uncorrupted, but incomplete, view of one or other constituent source.

1.3. Application to robust ASR

Application of missing data techniques in robust ASR requires a solution to two problems:

(i) identification of reliable spectro-temporal evidence;

(ii) modification of recognition algorithms to handle incomplete data.

In this paper, our emphasis is on (ii), but in this section we discuss (i) and relate our methodology to other approaches to robust ASR.

In the experiments we report, simple noise estimation techniques are used as the basis for evidence selection. Like most approaches to robust ASR (see reviews by Grenie and Junqua, 1992; Gong, 1995; Furui, 1997; Speech Communication (special issue) 25 (1–3) (1998)) we are therefore making use of *noise models* [1] in this paper.

However, *the missing data approach does not presuppose the existence of noise models*. There are alternative solutions to problem (i) based, for instance, on Auditory Scene Analysis (Cooke, 1993; Brown and Cooke, 1994; Ellis, 1996). To illustrate, Fig. 1(a) shows a mixture of speech and factory noise. In Fig. 1(b) the regions in black have been assigned to the speech source solely by making use of harmonicity cues in voiced speech segments. Another cue for grouping evidence from different sources is location: this is illustrated in Fig. 1(c), where inter-aural time differences are used to assign the time-frequency pixels to one of two sources. Here the auditory scene is simulated, so that the speech has an ITD of −0.5 ms relative to the noise. In these examples, properties of the *target source and the auditory scene alone* are being used to identify reliable data.

The 'primitive grouping' (Bregman, 1990) used in these examples exploits properties of the target source alone, i.e. low-level constraints that reach back to the physics of sound and the properties of the auditory and speech production systems. If a speech recogniser is to function in unpredictable listening conditions, with an indefinite number of unidentified sound sources present, which furthermore come and go with time, it seems unrealistic to construct and deploy noise models. However, in situations where noise characteristics are available, noise models can and should be used. This applies in the work reported here, but the theory we present involves no commitment to any particular method of deciding what data is missing. We return to these issues in Section 5.3.

1.4. Previous missing data studies

A number of studies into the missing data problem have been reported (Proc. NIPS, 1996).

[1] We use the term 'noise model' to refer to any structure (such as an estimated distribution or an HMM) from which it is possible to generate typical noise observations.

Naive approaches involve unconditional estimation of missing values. Classification performance using such techniques typically falls very rapidly as the proportion of missing data increases. A more principled method is presented in (Ahmed and Tresp, 1993) for converting an existing classifier to deal with both missing and uncertain (i.e., potentially noise-corrupted) data. A maximum likelihood (ML) framework for both training and recognition with missing data in its most general form is presented in (Ghahramani and Jordan, 1994). The principle advantage of the ML criterion is its ability to generate practical parameter estimation techniques.

Holmes and Sedgwick (1986) proposed a method for handling observations estimated to be noise-dominated. Observations with energies above that of the noise estimate are treated conventionally, while those below the noise estimate are handled using the cumulative probability of all observations below the noise estimate. Varga et al. (1988) compared the performance of the Holmes and Sedgwick approach with noise masking (Klatt, 1976) and noise marking (Bridle et al., 1994), demonstrating its superiority at low SNRs. When reliable data is identified using noise modelling, one of the missing data techniques formulated and evaluated in the current paper – bounded marginalisation (Section 3.3) – is essentially the same as the Holmes and Sedgwick (1986) approach when generalised for CDHMMs. However, formulating the problem as one of missing data (Cooke et al., 1994; Green et al., 1995; Lippmann and Carlson, 1997; Morris et al., 1998) is advantageous for two reasons. First, it highlights the applicability of solution techniques to problems other than dealing with noisy speech. For instance, no noise estimation is involved in dealing with clean, but band-limited speech, yet missing data techniques provide a natural means of handling this condition. Second, the relationship of missing data approaches to results in human speech perception of distorted speech referred to in Section 1.2 is made apparent.

Brendborg and Lindberg (1997) developed a feature-masking technique which has some commonality with the missing data approach in that it marginalises certain features. Missing data imputation prior to decoding was reported by Raj et al. (1998). The missing data approach has some features in common with emerging techniques for robust ASR which attempt classification on the basis of partial information. Notably, the multi-stream approach to ASR (Bourlard and Dupont, 1996; Hermansky et al., 1996) has achieved some success in decoding on the basis of independent processing of different spectral regions, with later recombination.

The work reported here differs from earlier studies in several important respects. Empirical tests of missing data theory, largely in computer vision, have mainly been confined to 'toy' problems with low-dimensional observation vectors and predominantly used random rather than naturally occurring deletion patterns. In contrast, this paper describes practical procedures for dealing with real-world missing data problems in ASR, while keeping to a general theoretical framework which permits these techniques to be transferred directly to other high dimensional classification problem domains. Section 2 describes two distinct approaches to classification with missing data. Section 3 shows how these can be applied to ASR based on continuous-density HMM. Recognition studies using TIDigits in additive noise are described in Section 4.

2. Classification with unreliable data

2.1. The missing data problem

The classification problem in general is to assign an observation vector x to a class C. In the missing data case, some components of x are unreliable or unavailable. In these circumstances, the problem for probabilistic classification is that the likelihood $f(x|C)$ cannot be evaluated in the normal manner.

In the following, it is assumed that some prior process of the form outlined in Section 1 (and exemplified in Section 4.3) has partitioned each data vector x into reliable and unreliable parts, (x_r, x_u). The components of x_r are reliable data, available to the classifier. The components of x_u are distinguished by uncertainty about their true values. Two uncertainty conditions are considered

here. The first is complete ignorance of the unreliable values. This could result from sensor failure, temporary loss of signal, or narrow-bandwidth transmission, for instance. The second condition is knowledge of the interval within which the true data lie. In acoustic signal processing, this common situation simply expresses the case of additive noise masking. Assuming that energy estimates are made over sufficiently large spectro-temporal regions, the observed value defines an upper bound on the true data value, while the lower bound is zero.

It is important to reiterate that these two types of uncertainty require no commitment to noise models. Of course, if such models were available, their effect on the target signal could be expressed probabilistically as, for instance, in (Gales and Young, 1993). However, in many situations of everyday listening, the requirement to have adequate noise models for the whole of the attentional background is unreasonable, and it is of interest to see how well systems with no commitment to noise models can perform.

Here, two approaches to classification with unreliable data are identified: *data imputation* and *marginalisation*. Data imputation is of particular interest when some later process needs actual estimates of the unreliable components, as would be required in speech enhancement or for further data transformation (e.g. to the cepstral domain). Marginalisation is of value in classification tasks which can proceed without reconstruction of unreliable parts.

2.2. Data imputation

In data imputation, the aim is to estimate values for the unreliable components of x, producing a complete observation vector $\hat{x}$, and to then proceed with classification using $f(\hat{x}|C)$.

There are a number of simple but suboptimal approaches to data imputation. Unreliable values could be replaced by the unconditional means for those components in any given class, for example. Previous work with such simplistic techniques has shown them to be inadequate for real-world tasks such as robust ASR (Cooke et al., 1997). A better approach is to use knowledge of the reliable components in conjunction with the covariance structure of each class (i.e., the conditional means, modes, etc.). Specifically, the distribution of unreliable components conditioned on the reliable components can be used for data imputation. Formally, if $f(x|C)$ denotes the distribution of the complete vector in a given class C, then unreliable items are replaced by values drawn from $f(x_u|x_r, C)$. In many cases, values can be chosen using the expectation for this distribution. However, if it is suspected that the distribution is other than unimodal, an estimate of its mode would be more appropriate.

2.3. Marginalisation

An alternative to data imputation is to classify based solely on reliable components, effectively integrating over the unreliable components. For the case of complete ignorance of unreliable values, this is achieved by using the marginal distribution $f(x_r|C)$ in place of $f(x|C)$. The marginal distribution over reliable data has been widely used on its own for classification with incomplete data (Ahmed and Tresp, 1993; Green et al., 1995; Lippmann and Carlson, 1997).

Sections 3.2 and 3.3 develop the data imputation and marginalisation approaches further in the context of HMM-based speech recognition, together with extensions in which unreliable evidence is not dismissed altogether, but serves to bound the possible values which the true observation could take.

3. Application to HMM-based ASR

3.1. Architecture and assumptions

In conventional Continuous Density Hidden Markov Model Speech Recognition, each chosen speech unit is represented by a trained HMM with a number of states. The states correspond to the classes of Section 2. Each state is characterised by a multivariate mixture Gaussian distribution over the components of the acoustic observation vector x, from an observation sequence X. The parameters of these distributions, together with state

transition probabilities within models are estimated during training, commonly using the Baum–Welch algorithm. All this comprises the recogniser's acoustic model. There will also be a language model expressing transition probabilities between the speech units of the acoustic model. A decoder (usually implementing the Viterbi algorithm) finds the state sequence having the highest probability of generating X.

Here, the two approaches outlined in Section 2 are applied to this framework. We assume that HMMs have been trained in the usual way, on clean data. For generality, we avoid training different models for different noise conditions. We further assume that the density in each state C_i can be adequately modelled using mixtures of M Gaussians with diagonal-only covariance structure [2],

$$f(x|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x|k, C_i), \tag{1}$$

where the $P(k|C_i)$ are the mixture coefficients.

This is a reasonable assumption if sufficient numbers of mixture components are used since any covariance can theoretically be approximated with such a mixture. In previous studies, the use of full covariance structures in missing data ASR has been demonstrated to be computationally prohibitive (Morris et al., 1998).

3.2. Data imputation

Following the approach outlined in Section 2.2, the conditional density $f(x_u|x_r, C_i)$ is required for each state C_i. By Bayes Rule

$$f(x_u|x_r, C_i) = \frac{f(x_u, x_r|C_i)}{f(x_r|C_i)}. \tag{2}$$

But $f(x_u, x_r|C_j)$ is just $f(x|C_i)$. Substituting (1) above, we obtain

$$f(x_u|x_r, C_i) = \frac{\sum_{k=1}^{M} P(k|C_i) f(x_u, x_r|k, C_i)}{f(x_r|C_i)}. \tag{3}$$

[2] In this paper, $P(x)$ denotes the probability of x and $f(x)$ denotes the probability density at x.

Since each mixture component indexed by k is modelled using diagonal-only covariance, the independence assumption can be applied *at the level of each mixture* (i.e., while $f(x_u, x_r|C_i) \neq f(x_r|C_i) f(x_u|C_i)$, $f(x_u, x_r|k, C_i) = f(x_r|k, C_i) f(x_u|k, C_i)$, for each k), allowing (3) to be rewritten as

$$f(x_u|x_r, C_i) = \frac{\sum_{k=1}^{M} P(k|C_i) f(x_r|k, C_i) f(x_u|k, C_i)}{f(x_r|C_i)}.$$

An application of Bayes' rule results in the following expression for the conditional density:

$$f(x_u|x_r, C_i) = \sum_{k=1}^{M} P(k|x_r, C_i) f(x_u|k, C_i). \tag{4}$$

Any value can be imputed in place of x_u. However, some values are more probable than others. Here, we impute the most probable value of the conditional density $f(x_u|x_r, C_i)$. For a unimodal distribution, this is the expected value,

$$E_{x_u|x_r, C_i}\{x_u\} = \int f(x_u|x_r, C_i) x_u \, dx_u. \tag{5}$$

Substituting (4) and moving the integral inwards produces

$$E_{x_u|x_r, C_i}\{x_u\} = \sum_{k=1}^{M} P(k|x_r, C_i) \int f(x_u|k, C_i) x_u \, dx_u.$$

The integral expression is the expectation of $f(x_u|k, C_i)$, which is just the mean of this mixture component for the unreliable portions of the data vector, as estimated during model training, $\mu_{u|k,C_i}$. Hence, data imputation estimates unreliable components of the observation vector using

$$\hat{x}_{u,i} = \sum_{k=1}^{M} P(k|x_r, C_i) \mu_{u|k,C_i}, \tag{6}$$

where the $P(k|x_r, C_i)$ can be considered as responsibilities for mixture component k given the reliable data in each state,

$$P(k|x_r, C_i) = \frac{P(k|C_i) f(x_r|k, C_i)}{\sum_{k=1}^{M} P(k|C_i) f(x_r|k, C_i)}. \tag{7}$$

Note that each state C_i gives rise to different imputed values $\hat{x}_{u,i}$. In the decoding algorithm, these state-dependent vectors replace the normal

state-independent feature vector. After decoding, the imputed vectors for the states on the winning path can be used to provide a single restoration, for instance in speech enhancement (see Section 5.1).

The term $f(x_r|k, C_i)$ also appears in the marginalisation approach for handling missing data. Its evaluation is detailed in Section 3.4.

3.3. Marginalisation

Here, the aim is to compute the HMM state output probabilities using a reduced distribution based solely on reliable components. We require the marginal determined by integrating over all missing components,

$$f(x_r|C_i) = \int f(x_u, x_r|C_i)\,\mathrm{d}x_u. \tag{8}$$

Substituting Eq. (1) and exploiting the independence, within each mixture component, of reliable and unreliable subvectors, we obtain

$$f(x_r|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x_r|k, C_i) \int f(x_u|k, C_i)\,\mathrm{d}x_u. \tag{9}$$

3.4. Evaluation of the marginal

Eqs. (6), (7) and (9) identify the computations required for data imputation and marginalisation, respectively, for the case of densities representable by mixtures of diagonal-covariance Gaussians. Both use the marginal $f(x_r|k, C_i)$ which is easily evaluated in the case of diagonal-covariance multivariate Gaussians using the following procedure.

Let $N(x; \mu_{k,i}, \sigma^2_{k,i})$ denote the Gaussian for mixture component k of model i. Using the partition of the observation vector $x = (x_r, x_u)$ into reliable and unreliable components, the mean and variance vectors for mixture component k of model i are similarly partitioned,

$$\mu = (\mu_{r,k,i}, \mu_{u,k,i}),$$

$$\sigma^2 = (\sigma^2_{r,k,i}, \sigma^2_{u,k,i}).$$

The marginal is then obtained as (Morrison, 1990)

$$f(x_r|k, C_i) = N(x_r; \mu_{r,k,i}, \sigma^2_{r,k,i}). \tag{10}$$

3.5. Bounded marginalisation

The integral in Eq. (9) reduces to unity if the unreliable components are missing altogether. However, if knowledge of the unreliable subvector exists in the form of bounds $[x_{low}, x_{high}]$, and in the case of diagonal Gaussian mixture components, the integral can be evaluated as a vector difference of multivariate error functions,

$$\int f(x_u|k, C_i)\,\mathrm{d}x_u = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x_{high,u} - \mu_{u,k,i}}{\sqrt{2}\sigma_{u,k,i}}\right) - \mathrm{erf}\left(\frac{x_{low,u} - \mu_{u,k,i)}}{\sqrt{2}\sigma_{u,k,i}}\right)\right]. \tag{11}$$

Any further prior knowledge of noise characteristics could be employed at this point.

3.6. Utilising bounds with data imputation

Knowledge of the unreliable subvector in the form of bounds $[x_{low}, x_{high}]$ can also be used to constrain the imputed values $\hat{x}_u$ (in the following, the subscript i indicating the state to which the imputed value corresponds is dropped for clarity). Eq. (11) can be used in the calculation of the responsibilities (7) to produce values which discriminate better between the Gaussians in the mixture, especially in the frames with little or no reliable data.

It is also necessary to constrain the choice of imputed values for unreliable data so that they lie within the bounds of the spectro-temporal energy surface. If the individual Gaussians in the state pdfs are well-separated, their means can be taken as the modes of the mixture distribution.

For each individual Gaussian, we compute the most likely value within the bounds,

$$\hat{x}_{u,k} = \begin{cases} \mu_{u,k}, & x_{low} \leqslant \mu_{u,k} \leqslant x_{high}, \\ x_{high}, & x_{high} < \mu_{u,k}, \\ x_{low}, & x_{low} > \mu_{u,k}, \end{cases}$$

and choose the most likely $\hat{x}_{u,k}$ in each state as our imputed value.

3.7. Computational complexity and the use of energy bounds

Data imputation (6) doubles the computational complexity compared to conventional output probability evaluation. Marginalisation without bounds requires less computation than a conventional output probability calculation since probability evaluation of (univariate) Gaussians is required only for x_r rather than the whole of x. With bounds (11), complexity is greater than a conventional approach, and its tractability depends on the availability of a fast error function implementation. If the observation vector represents spectral energies, as in the experiments reported below, $x_{low} = 0$ and x_{high} is the observed energy. In these conditions, the second error function in Eq. (11) can be precomputed.

4. Recognition experiments

4.1. Task details

To evaluate the approaches presented in Section 3, we report recognition studies using the TIDigits connected digit corpus (Leonard, 1984). Acoustic vectors were obtained via a 64-channel auditory filter bank (Cooke, 1993) with centre frequencies spaced linearly in ERB-rate from 50 to 8000 Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a first-order filter with 8 ms time constant, and sampled at a frame-rate of 10 ms. The training section of the corpus was used to train 12 word-level HMMs (1–9, 'oh', 'zero' and a silence model), each with 8 emitting states. Observations in each state were modelled with a 10 component mixture. Models were trained on clean training data. Testing was performed on a 240-utterance subset of the TIDigits test set. HTK Version 1.5 (Young and Woodland, 1993) was used for training, and a local MATLAB decoder adapted for missing data was used for all recognition tests.

Three noise signals (car, factory, Lynx helicopter) from the NOISEX corpus (Varga et al., 1992) were added with random offsets at a range of SNRs from −5 dB to 20 dB in 5 dB steps. Auditory 'spectrograms' or *rate maps* of these signals are shown in Fig. 2 for several SNRs. These noise types present differing degrees of challenge for robust ASR. Most of the noise power density of this car noise sample is concentrated in the sub-200 Hz region, and hence is a comparatively benign masker of speech formant regions. It is also highly stationary. The helicopter noise has significant energy (in the form of relatively narrow peaks) in the mid-frequency region, and is less stationary than the car noise. The factory noise sample is the least stationary of the three, and is characterised both by energy peaks in the formant region and by impulsive energetic regions (hammer blows?). These characteristics are reflected in the recognition results obtained.

4.2. Baseline systems

Tables 1–3 give baseline accuracies for car, helicopter and factory noise, respectively. Results for recognition of the noisy rate maps alone, and in combination with a simple spectral subtraction

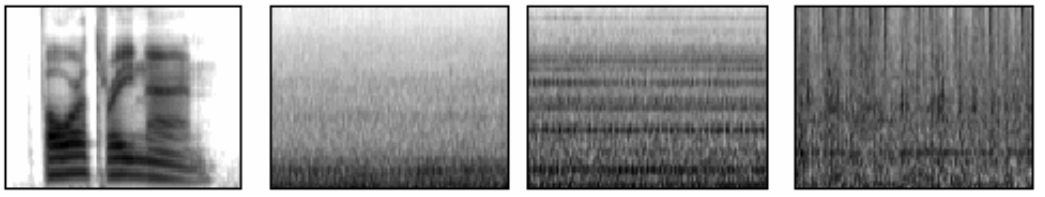

Fig. 2. Auditory spectrograms of, from left to right, the TIDigit sequence 'four three nine', car noise, helicopter noise and factory noise. The noise segments (from NOISEX) are illustrative portions only. Random offsets into a longer sample of each are used for the experiments.

Table 1
Baseline accuracies for car noise

	Clean	20	15	10	5	0	−5
MFCC13_D_A	99	99	99	98	98	97	92
MFCC13_D_A + CMN	99	99	99	99	98	98	97
RATE64	97	97	96	93	86	70	50
RATE64 + SS	97	96	96	95	92	87	70

Table 2
Baseline accuracies for helicopter noise

	Clean	20	15	10	5	0	−5
MFCC13_D_A	99	99	98	91	69	18	01
MFCC13_D_A + CMN	99	99	96	88	62	21	01
RATE64	97	73	56	34	20	11	8
RATE64 + SS	97	93	84	63	29	2	0

Table 3
Baseline accuracies for factory noise

	Clean	20	15	10	5	0	−5
MFCC13_D_A	99	97	90	56	14	0	0
MFCC13_D_A + CMN	99	97	92	75	41	10	0
RATE64	97	58	47	33	14	10	7
RATE64 + SS	97	86	65	37	16	7	3

scheme are shown. A simple non-adaptive implementation of spectral subtraction was employed in which a fixed noise estimate is computed as the mean of the initial 10 frames of each noisy digit sequence. More advanced schemes such as adaptive spectral subtraction (Mokbel, 1992), energy histograms (Hirsch and Ehrlicher, 1995) and minimum statistics (Martin, 1993) could be employed to derive better estimates of the noise spectrum.

For comparison with a conventional robust ASR technique, results for a system trained on a 13-element MFCC parameterisation with deltas and accelerations are also shown, both with and without cepstral mean normalisation (CMN).

For each of the four combinations, raw recognition accuracy deteriorates with increasing noise levels. The MFCC system operates at a higher level of accuracy than that based on rate maps: orthogonalised representations are more accurately modelled by diagonal Gaussian mixtures. Spectral subtraction results in the expected improvement over baseline performance, although for helicopter and factory noise this advantage is small for SNRs below 10 and 15 dB, respectively. The departure from absolute stationarity makes these noise types less suitable for the simple spectral subtraction approach used here. Similarly, CMN improves the MFCC system baseline. In both cases (MFCC with CMN and rate map with SS), the gain is relatively modest.

4.3. Limits on performance

To reveal the potential of the missing data technique, we examine the claim that recognition can be based on the reliable evidence subset in the artificial condition where the spectro-temporal location of reliable evidence is known a priori. Any choice of spectro-temporal regions can be visualised as a binary mask. There are a number of ways to compute such a mask if the clean speech and noise signals are available prior to mixing. The scheme adopted here forms the mask from those regions whose energy in the mixture is within 3 dB of the energy in the clean

speech (corresponding to a local SNR of 7.7 dB – see Section 4.5). Examples of such 'a priori masks' and corresponding noisy spectrograms are shown in Figs. 3–5 for the three noise types and a range of SNRs.

It is clear that the a priori masks allow through most of the salient phonetic information, even at high noise levels. Note that as the SNR decreases, the 3 dB criterion will progressively omit lower energy regions of the speech signal as they become noise-dominated. Taking the helicopter noise masker in Fig. 4 as an example, evidence for the initial weak fricative /f/ and the final nasal /n/ of '439' is lost as the SNR decreases from 20 to 0 dB.

Performance of the missing data techniques with a priori masks is given in Tables 4–6 as described below, and highlights the potential of our approach: baseline performance and a priori performance for helicopter and factory noise are plotted against global SNR in Fig. 6. It should be noted that these results do not represent an upper limit for missing data techniques in general: an ideal mask might, for instance, employ different local SNR thresholds dependent on the amount of missing data.

4.4. Missing data results using a negative energy criterion

Section 1.4 outlined several methods for the identification of reliable data. One cause of suspect data are artefacts introduced by signal processing. Here, spectral subtraction can leave negative energy values. Missing data techniques can be used to counter these anomalies, as suggested by Drygajlo and El-Maliki (1998), who demonstrated a significant improvement over the use of spectral subtraction alone in a speaker identification task. If the observed magnitude in any frame is denoted by $|s+n|$ and the estimated noise magnitude spectrum by $\hat{n}$, then the *negative energy criterion* removes spectral regions from the mask if

$$|s+n| - |\hat{n}| < 0. \tag{12}$$

Column 3 of Figs. 3–5 shows example masks which result from this criterion.

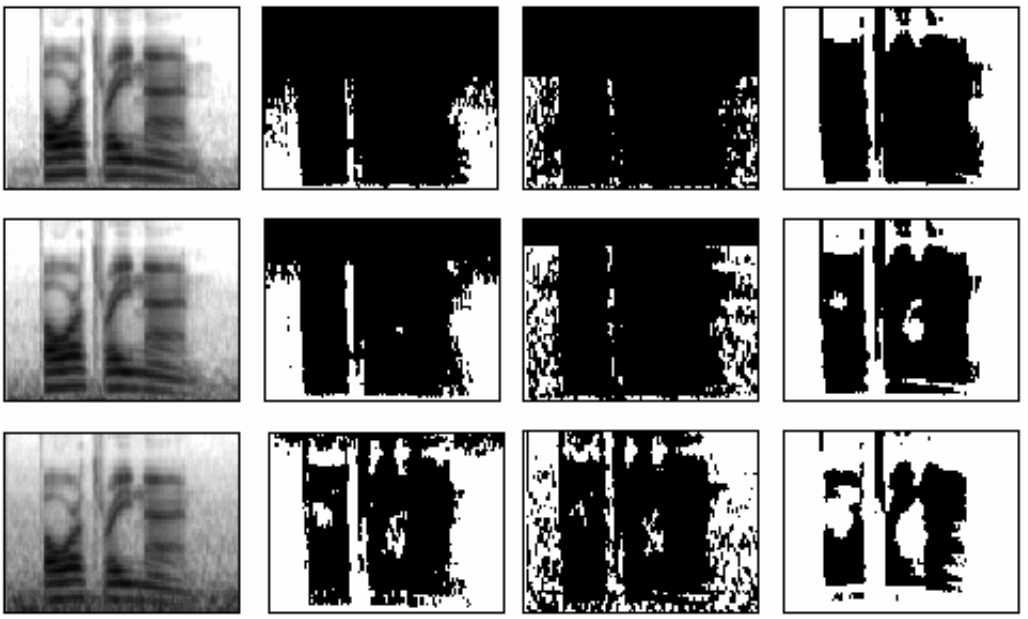

Fig. 3. Columns 1–4 depict auditory spectrograms, a priori reliable data masks, masks produced by the negative energy criterion (12) and masks produced by the joint criterion (Eqs. (12) and (13)), respectively. Rows 1–3 differ in the level of car noise added (20 dB, top; 10 dB, middle; 0 dB, bottom).

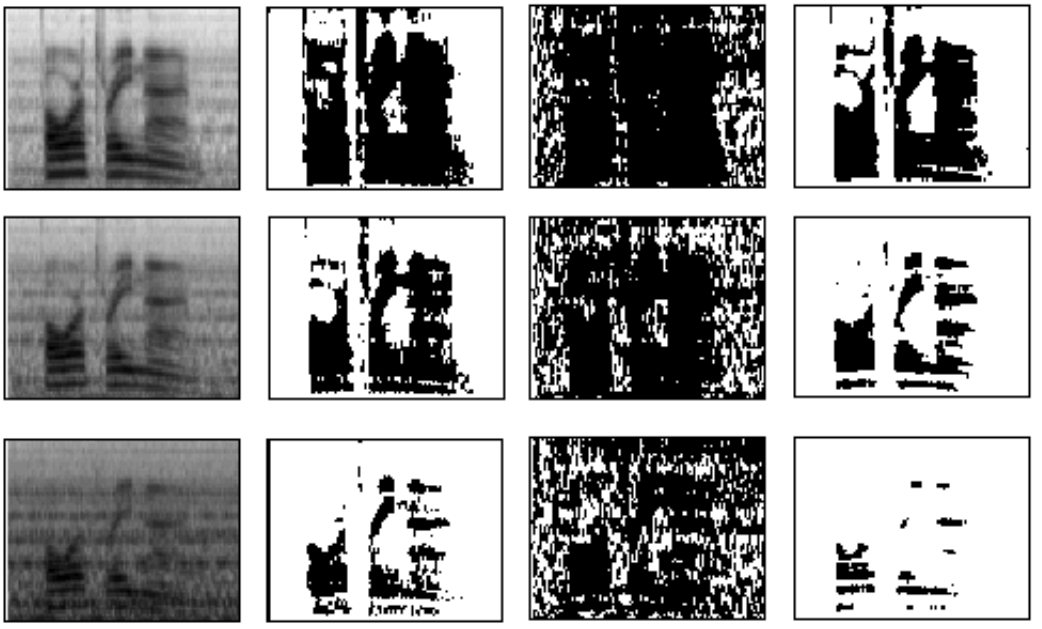

Fig. 4. As for Fig. 3 using helicopter noise rather than car noise.

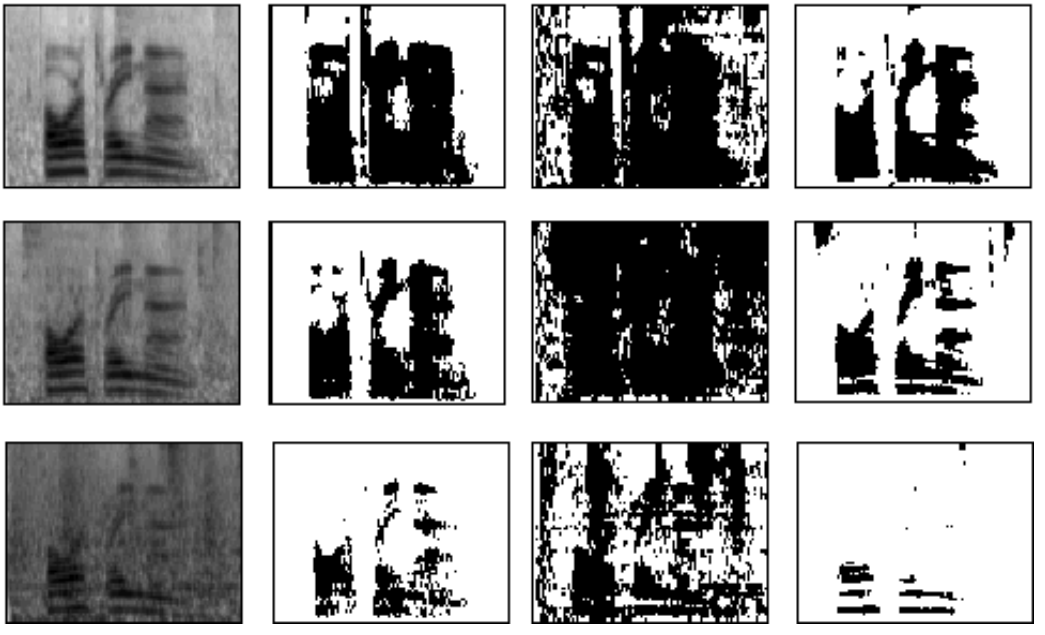

Fig. 5. As for Fig. 4 but with factory noise.

Tables 7–9 provide results for the missing data techniques described in Section 3 after applying this criterion. In each case, missing data processing results in an improvement over spectral subtraction alone. Versions of the two missing data algorithms of Section 3 which make use of bounds

Table 4
Missing data recognition accuracies for car noise, local SNR criterion[a]

	Clean	20	15	10	5	0	−5
Marginalisation	82 (97)	79 (97)	75 (97)	76 (96)	76 (91)	75 (82)	71 (77)
Bounded marginalisation	96 (97)	97 (97)	97 (97)	97 (97)	96 (97)	95 (97)	95 (96)
Imputation	96 (97)	96 (97)	94 (97)	92 (96)	88 (90)	79 (87)	70 (79)
Bounded imputation	97 (97)	97 (97)	96 (97)	96 (97)	96 (97)	95 (96)	91 (94)

[a] Data in parenthesis represent a priori mask accuracies.

Table 5
Missing data recognition accuracies for helicopter noise, local SNR criterion[a]

	Clean	20	15	10	5	0	−5
Marginalisation	82 (97)	76 (75)	76 (74)	73 (77)	64 (77)	46 (74)	29 (68)
Bounded marginalisation	96 (97)	96 (97)	95 (97)	93 (96)	88 (93)	69 (81)	41 (62)
Imputation	96 (97)	81 (92)	75 (88)	67 (82)	51 (74)	32 (66)	19 (52)
Bounded imputation	97 (97)	94 (96)	91 (95)	83 (92)	69 (86)	52 (75)	30 (61)

[a] Data in parenthesis represent a priori mask accuracies.

Table 6
Missing data recognition accuracies for factory noise for RATE64, local SNR criterion[a]

	Clean	20	15	10	5	0	−5
Marginalisation	82 (97)	63 (74)	61 (74)	49 (75)	40 (76)	25 (74)	15 (71)
Bounded marginalisation	96 (97)	94 (97)	91 (97)	81 (95)	59 (87)	34 (69)	13 (55)
Imputation	96 (97)	78 (91)	68 (87)	53 (81)	36 (74)	18 (63)	10 (50)
Bounded imputation	97 (97)	90 (96)	84 (93)	67 (89)	47 (80)	25 (65)	11 (43)

[a] Data within parenthesis represent a priori mask accuracies.

information outperform unbounded versions. From these results, it is not clear whether marginalisation or imputation is superior.

4.5. Missing data results using a local SNR criterion

While treating negative energy artefacts as missing produces some improvements, a further gain might be expected by identifying regions with low local SNR. Such regions are likely to be dominated by noise and it may be profitable to treat them as missing too. We refer to this as the SNR criterion, defined as treating as missing those regions whose estimated local SNR falls below some threshold δ,

$$10 \log \frac{\hat{s}^2}{\hat{n}^2} < \delta,$$

i.e.

$$\hat{s}^2 < \beta \hat{n}^2, \quad \text{where } \beta = 10^{\delta/10}. \tag{13}$$

The negative energy criterion (12) implies the SNR criterion, so this artefact is taken into account by (13). Following experimental verification, we set the value of the estimated local SNR threshold in the work reported below to 7.7 dB. This value is equivalent to a 'posterior SNR' (El-Maliki and Drygajlo, 1999) of 10.7 dB.

Columns 4 of Figs. 3–5 show SNR criterion masks. Note their similarity to the a priori masks in column 2 and their dissimilarity with the negative energy masks of column 3.

Recognition results obtained by treating as missing those points which satisfy the SNR criterion are shown in Tables 4–6. The accuracies obtained demonstrate a further large improvement over the negative energy criterion for all noise types and methods. Here, a clear superiority of marginalisation over imputation is revealed.

The tables also show results obtained if masks were obtained by prior knowledge of the local SNR. These figures illustrate the potential of techniques which distinguish between reliable and unreliable evidence.

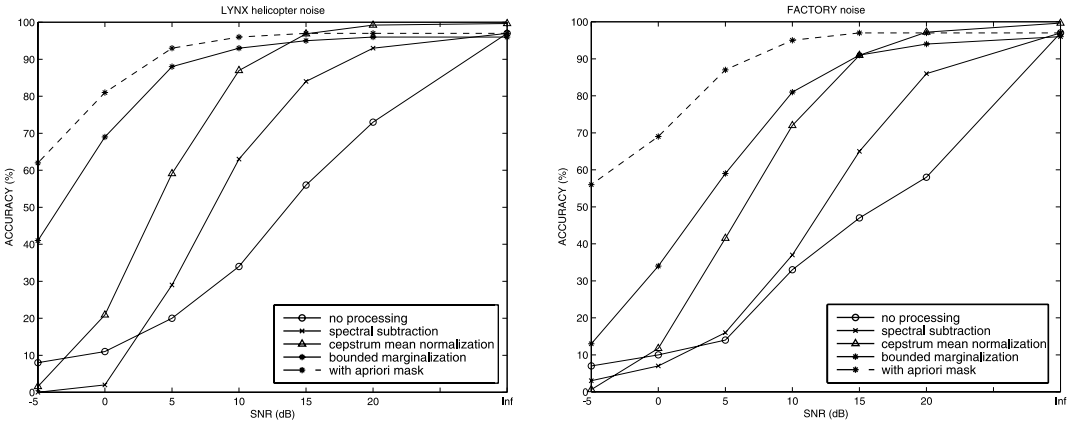

Fig. 6. Summary of results for helicopter noise (left) and factory noise (right). Recognition accuracy is plotted against SNR for no processing, spectral subtraction, MFCCs with cepstral mean normalisation and missing data recognition (with bounded marginalisation). Fo missing data, results with masks derived from noise estimates, and from the true noise value ('a priori masks' are shown).

4.6. Summary of results

Fig. 6 summarises our results for helicopter and factory noise, showing accuracies against global SNR for the best-performing missing data technique (bounded marginalisation) compared to no processing, spectral subtraction and cepstral mean normalisation. There is a clear win below 10 dB for missing data recognition. The gap widens as the amount of noise increases. Performance is better for helicopter noise than factory noise: this is to be expected for any system making use of simple noise estimates, because of the unpredictable impulsive factory noise components. We also show the missing data performance using bounded marginalisation with a priori masks. In the case of helicopter noise the performance gap between these masks and masks based on SNR estimates is encouragingly narrow. It is wider for factory noise, possibly for the reasons given above.

5. Discussion

5.1. Marginalisation versus imputation

While both methods produce a similar pattern of improvement over the baseline spectral subtraction technique, marginalisation is superior, since it involves no commitment to choosing a single estimate to represent an uncertain value; rather, it takes into account the distribution of the missing points. For imputation, introduction of the SNR criterion (13) is not beneficial unless the bounds constraint is also applied. This finding suggests that, compared to marginalisation, data imputation is more heavily affected by data sparsity than data reliability.

If restored observation vectors are not required in subsequent processing, marginalisation is the method of choice. However, the restored vector produced by data imputation offers considerable practical advantages for further processing. For example, it opens the way to applications such as full-band restoration and speech enhancement, as illustrated in Fig. 7. Furthermore, restored vectors can be used as input to any ASR system (Statistical or Hybrid) and can undergo useful transformations such as (approximate) orthogonalisation via the DCT, differencing and normalisation.

The imputation method used here is only one of many possible schemes and no claims to its optimality can be made. If the conditional distribution $f(x_u|x_r, C_i)$ is multimodal, then the values of choice for imputation are the modes. However, they are not as easy to find as the mean. Further, if there are several modes, a suitable criterion is needed to pick and impute one of them only. Using the highest (most probable) mode may be the natural choice in the absence of other knowledge.

Table 7
Missing data recognition accuracies for car noise, negative components criterion

	Clean	20	15	10	5	0	−5
Marginalisation	97	97	97	96	96	93	89
Bounded marginalisation	97	97	97	96	96	94	90
Imputation	97	97	97	96	96	94	91
Bounded imputation	97	97	97	96	95	94	91

Table 8
Missing data recognition accuracies for helicopter noise, negative components criterion

	Clean	20	15	10	5	0	−5
Marginalisation	97	94	90	76	48	19	9
Bounded marginalisation	95	95	91	78	51	22	9
Imputation	97	95	92	80	57	25	11
Bounded imputation	97	95	92	83	59	28	10

Table 9
Missing data recognition accuracies for factory noise for RATE64, negative components criterion

	Clean	20	15	10	5	0	−5
Marginalisation	97	89	76	52	27	16	11
Bounded marginalisation	97	91	80	57	32	16	10
Imputation	97	91	79	56	29	14	4
Bounded imputation	97	92	81	60	32	15	10

Constraints such as bounds on the imputed values or continuity (smoothness) of the features (imputed and present) may help in choosing the correct mode for imputation. For the case of mixtures of diagonal Gaussians, if the Gaussians are far enough apart, the distribution will be multimodal. A fuller discussion of mode-finding is contained in (Carreira-Perpinan, 1999).

5.2. Use of energy bounds

Both the estimated masks and those obtained through a priori knowledge of the local SNR result in systems with surprisingly poor performance at high SNRs (>10 dB). Analysis of the results shows that the drop in accuracy from clean to 20 dB SNR is due to a large number of insertion errors in the decoder output. This can be explained by the observation (see Figs. 3–5) that, in the a priori masks, periods without speech energy are noise dominated and therefore not present in the mask even for high SNRs. For any frame in which very little reliable evidence is available, the missing data output probability calculation will yield similar likelihoods for all model states. In the limit where no data is available at all, these will collapse to unity. If the period of little or no reliable data is short, the decoder may overcome this temporary lack of discriminatory evidence. However, an extended period will produce equalised path likelihoods and hence the decoder result will, in the absence of priors (e.g. from a language model), contain arbitrary insertions. An analysis of the correctness versus accuracy scores confirms that the insertions occur when there is no speech energy and are evenly distributed across models. Without additional information, this is the best that the decoder can achieve.

Fortunately, more information is available in the form of bounds on the spectro-temporal energy surface. Intuitively, this constraint measures the degree to which the hypothesised acoustics could fit beneath the observed total energy. For high SNRs, we would expect the relatively low noise level to provide an effective masker for the silence model, but not for others. The use of

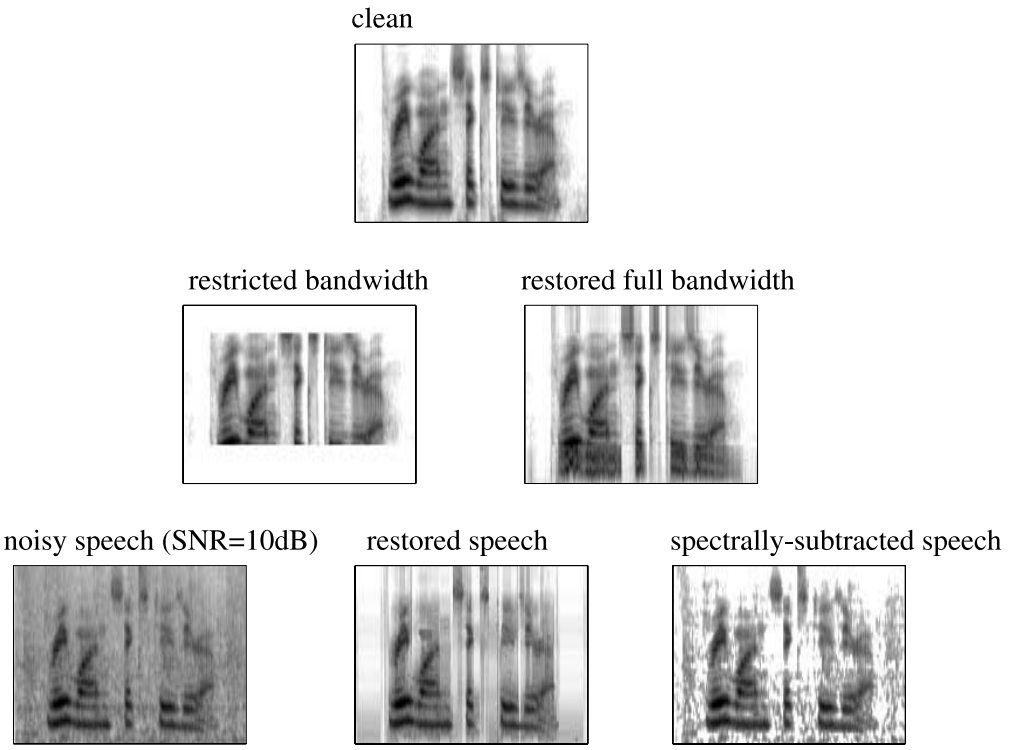

Fig. 7. Illustration of data imputation and applications. The middle row shows restoration of limited bandwidth speech. The bottom row shows restoration of noisy speech. Comparison with the rightmost panel illustrates the benefits of the imputation approach over spectral subtraction alone.

bounds leads to a very significant reduction in error rate at most SNRs. For car noise at −5 dB SNR, the error rate is reduced from 95% to 30% by spectral subtraction alone, to 11% using the negative energy criterion, and to 5% with both criteria and bounds. In helicopter noise, use of the SNR criterion and bounds produces even larger gains: at +5 dB SNR, whereas the negative energy criterion reduces error rate from 80% to 52%, the additional constraints produce 12%. The less stationary factory noise shows a similar pattern of improvement.

5.3. Criteria for choosing unreliable regions

This study confirms the finding of Drygajlo and El-Maliki (1998) that negative energy artefacts resulting from noise overestimation can be handled using missing data. However, we have also shown that a far larger gain in accuracy can be obtained by thresholding points based on local SNR estimates. Masks of present/absent components in Figs. 3–5 illustrate the negative energy criterion (column 3) and the SNR criterion (column 4). It is clear that these criteria differ substantially in their assignment of reliability: negative energy alone produces masks which are only loosely related to the a priori masks (column 2), while SNR yields a much closer approximation to the ideal.

Comparison of columns 2 and 4 shows that the estimated noise masks differ less from frame to frame than the a priori masks. This is a consequence of using a constant noise estimate. In reality, the noise level varies somewhat from frame to frame. The limitations of a stationary noise estimate can be seen in the case of factory noise, where energy from intermittent impulsive hammer blows escapes spectral subtraction. This effect is

visible, for instance, at the beginning and the end of the column 4 mask in Fig. 5 at 10 dB SNR.

Such unpredictable non-stationarity could also escape adaptive noise estimation schemes and points to the need for techniques with an auditory motivation. A preprocessor based on computational auditory scene analysis makes few assumptions about the number or type of sources present. These techniques (Cooke, 1993; Brown and Cooke, 1994; Ellis, 1996) are informed by experimental studies (Bregman, 1990; Darwin and Carlyon, 1995) which have explored the perceptual organisation of sound. For example, components that are harmonically related, or have synchronous onsets, or arrive from the same location, appear to be grouped into a single perceptual description. Harmonic- and location-based grouping processing was illustrated in Fig. 1. The need for such processing is revealed, for example, in our factory noise results where simple noise estimation cannot capture the unpredictable impulsive components. A system using principles based on human sound source separation is needed so that such components can be treated as part of a different structure. Ellis (1996) uses an initial decomposition of arbitrary sound 'scenes' into impulsive, noise-like and periodic objects which are suitable for subsequent grouping into coherent source descriptions. We are currently pursuing such approaches in conjunction with the missing data techniques.

5.4. Further improvements

Our focus to date has been on the performance of missing data techniques within a relatively simple ASR system. A variety of improvements can be envisaged. First, the auditorily motivated rate map parameterisation, employed for compatibility with our interest in auditory scene analysis preprocessing, is known to result in suboptimal performance on ASR tasks. This is partly due to variability caused by the resolution of individual harmonics in the low-frequency region. Such resolution is beneficial for source separation. Additionally, spectral representations typically perform less well than those which have undergone cepstral transformation, as illustrated by our baseline results. Unfortunately, the cepstral transform smears localised spectral uncertainty into global cepstral uncertainty, so such approaches are not viable here. However, it has recently been demonstrated that certain localised spectral filtering operations can produce parameterisations which are competitive with MFCCs (Nadeu et al., 1997). These operations result in far less uncertainty smearing and may be applicable to the missing data approach.

Further optimisations may be gained by considering an adaptive local SNR threshold rather than the single global threshold used in the studies reported here. For high noise levels, data sparsity may be a limiting factor on performance, and in such circumstances it may be preferable to lower the reliable data threshold. At low noise levels, a higher threshold could be employed since sufficient components of the observation vector are available.

5.5. Relation to other approaches

It is interesting to contrast the missing data approach with other robust recognition techniques:

HMM decomposition (Varga and Moore, 1990) is a search of model state combinations for that combination sequence which has maximum likelihood. It is a general technique applicable to any signals for which models are available. In contrast, the missing data approach attempts to exploit properties of speech (redundancy, harmonicity,...) and the auditory scene (source location, physics of sound,...), in order to inform the recognition decoding.

In 'full-combination multistream recognition' (Morris et al., 1999), the 'union model' (Ming and Smith, 1999) and 'acoustic back-off' (de Veth et al., 1999) the assumption is that nothing is known a priori about which portions of the speech evidence are clean and which are corrupted. The full-combination and union approaches attack this problem by considering all possible noise positions, in order to find the best match. Acoustic back-off equates unreliable data with distribution outliers. In contrast, missing data techniques employ prior knowledge of the location of the reliable regions.

Two possibilities for combining these approaches are likelihood weighting in multiple streams and the introduction of a 'soft' reliable/unreliable decision.

6. General discussion and conclusions

Missing data methods can produce striking performance benefits for connected digit recognition in noise when used in combination with a simple noise estimation technique. We have recently reported results with more advanced noise estimation approaches (Vizinho et al., 1999): some modest further gains are obtained. The framework we have developed for classification with missing data for robust ASR has a number of potential advantages. It makes no assumptions about the noise (or, more generally, the collection of unattended sources) and is viable where the noise cannot be modelled in advance. Thus, there is no requirement to retrain models for each noise condition. Within a CDHMM system, it requires only minor modifications to the output probability calculation.

Finally, outside the context of speech technology, we are investigating the role that missing data processing may play in a general account of speech perception in listeners (Cooke and Green, in press). Since many experimental manipulations such as bandpass filtering involve a potential reduction in the amount of information available to listeners, it is of interest to determine how well the missing data approach is able to predict any reduction in intelligibility. To date, missing data processing has been used to model the perception of sine-wave speech (Barker and Cooke, 1997, 1999), low and high pass filtered digits (Lippmann and Carlson, 1997), narrow band speech (Cunningham and Cooke, 1999) and vowels (de Cheveigne and Kawahara, 1999).

Acknowledgements

This work was supported by the EPSRC Communications Signal Processing and Coding Initiative (Research Grant GR/K18962). It is now supported within the SPHEAR Transfer and Mobility of Researchers network (ERB-FMRXCT971050), the ESPRIT LTR Project fRESPITE (28149). L. Josifovski's doctoral work is supported by Motorola. Miguel Carreira-Perpignan, Andrew Morris, Guy Brown and Jon Barker made valuable suggestions contributing to this work. We thank the four referees for incisive comments.

References

Ahmed, S., Tresp, V., 1993. Some solutions to the missing feature problem in vision. In: Hanson, S.J., Cowan, J.D., Giles, C.L. (Eds.) Advances in Neural Information Processing Systems, Vol. 5. Morgan Kaufmann, San Mateo, CA, 393–400.

Allen, J.B., 1994. How do humans process and recognize speech? IEEE Trans. Speech Audio Process. 2 (4), 567–577.

Barker, J., Cooke, M.P., 1997. Modelling the recognition of spectrally reduced speech. In: Proc. Eurospeech '97, pp. 2127–2130.

Barker, J., Cooke, M.P., 1999. Is the sine-wave speech cocktail party worth attending? Speech Communication 27 (3–4), 159–174.

Bell, A.J., Sejnowski, T.J., 1995. An information–maximisation approach to blind separation and blind deconvolution. Neural Comput. 7 (6), 1004–1034.

Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. In: Proc. ICSLP '96.

Bridle, J.S., Ponting, K.M., Brown, M.D., Borrett, A.W., 1994. A noise compensating spectrum distance measure applied to automatic speech recognition. In: Proc. Inst. Acoust., pp. 307–314.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA.

Brendborg, M.K., Lindberg, B., 1997. Noise robust recognition using feature selective modelling. In: Proc. Eurospeech '97, pp. 295–298.

Brown, G.J., Cooke, M.P., 1994. Computational auditory scene analysis. Comput. Speech Language 8, 297–336.

Brown, G.J., Wang, D.L., 1997. Modelling the perceptual segregation of double vowels with a network of neural oscillators. Neural Networks 10, 1547–1558.

Brown, G.J., Cooke, M.P., Mousset, E., 1996. Are neural oscillations the substrate of auditory grouping? In: Proceedings of the Workshop on the Auditory Basis of Speech Perception, Keele University.

Carreira-Perpinan, M.A., 1999. Mode-finding for mixtures of Gaussian distributions. Technical report CS-99-03, Department of Computer Science, University of Sheffield, UK.

Cheveigne, A., Kawahara, H., 1999. Missing data model of vowel identification. J. Acoust. Soc. Am. 105 (6), 3497–3508.

Comon, P., 1994. Independent component analysis. A new concept? Signal Process. 36, 287–314.
Cooke, M.P., 1993. Modelling Auditory Processing and Organisation. Cambridge University Press, Cambridge, MA.
Cooke, M.P., Green, P.D., in press. Auditory organisation and speech perception: pointers for robust ASR. In: Greenberg, S. (Ed.), Listening to Speech, to appear.
Cooke, M.P., Green, P.G., Crawford, M.D., 1994. Handling missing data in speech recognition. In: Proc. ICSLP '94, pp. 1555–1558.
Cooke, M.P., Morris, A., Green, P.D., 1997. Missing data techniques for robust speech recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 863–866.
Cunningham, S., Cooke, M., 1999. The role of evidence and counter-evidence in speech perception. In: Int. Congress Phonetic Sci., to appear.
Darwin, C.J., Carlyon, R.P., 1995. Auditory Grouping. In: Moore, B.C.J. (Ed.), The Handbook of Perception and Cognition, Vol. 6. Hearing. Academic Press, New York, pp. 387–424.
de Veth, J., de Wet, F., Cranen, B., Boves, L., 1999. Missing feature theory in ASR: make sure you have the right type of features. In: Proceedings of the Workshop on Robust Methods for Speech recognition in Adverse Conditions, Tampere, Finland, May '99, pp. 231–234.
Drygajlo, A., El-Maliki, M., 1998. Speaker verification in noisy environment with combined spectral subtraction and missing data theory. In: Proc. ICASSP-98, Vol. 1, pp. 121–124.
Ellis, D.P.W., 1996. Prediction-driven computational auditory scene analysis. Ph.D. Thesis. MIT Press, Cambridge, MA.
El-Maliki, M., Drygaglo, A., 1999. Missing feature detection and compensation for GMM-based speaker verification in noise. In: Proceedings of the COST 250 Workshop on Speaker Recognition in Telephony, Rome, 10–12 November.
Fletcher, H., 1953. Speech and Hearing in Communication. Van Nostrand, New York.
Furui, S., 1997. Recent advances in robust speech recognition. In: Proceedings of the ESCA–NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, France, pp. 11–20.
Gales, M.J.F., Young, S.J., 1993. HMM recognition in noise using parallel model combination. In: Proc. Eurospeech '93, pp. 837–840.
Ghahramani, Z., Jordan, M.I., 1994. Supervised learning from incomplete data via an EM approach. In: Cowan, J.D., Tesauro, G., Alspector, J. (Eds.), Advances in Neural Information Processing Systems, Vol. 6. Morgan Kaufmann, San Mateo, CA, pp. 120–129.
Gong, Y., 1995. Speech recognition in noise environments: A survey. Speech Communication 16, 261–291.
Green, P.D., Cooke, M.P., Crawford, M.D., 1995. Auditory scene analysis and HMM recognition of speech in noise. In: Proc. ICASSP '95, pp. 401–404.
Grenie, M., Junqua, J.-C. (Eds.), 1992. Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, Cannes (ISSN 1018–4554).
Hermansky, H., 1998. Should recognisers have ears? Speech Communication 25 (1–3), 3–28.
Hermansky, H., Tibrewala, S., Pavel, M., 1996. Towards ASR on partially corrupted speech. In: Proc. ICSLP '96.
Hirsch, H.G., Ehrlicher, C., Noise estimation techniques for robust speech recognition. In: Proc. ICASSP '95, pp. 153–156.
Holmes, J.N., Sedgwick, N.C., 1986. Noise compensation for speech recognition using probabilistic models. In: Proc. ICASSP '86, pp. 741–744.
Klatt, D.H., 1976. A digital filter bank for spectral matching. In: Proc. ICASSP '76, pp. 573–576.
Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proc. ICASSP '84, pp. 111–114.
Liu, F., Yamaguchi, Y., Shimizu, H., 1994. Flexible vowel recognition by the generation of dynamic coherence in oscillator neural networks: speaker-independent vowel recognition. Biol. Cybern. 54, 29–40.
Lippman, R.P., 1996. Accurate consonant perception without mid-frequency speech energy. IEEE Trans Speech Audio Process. 4 (1), 66–69.
Lippmann, R.P., 1997. Speech recognition by machines and humans. Speech Communication 22 (1), 1–16.
Lippmann, R.P., Carlson, B.A., 1997. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise. In: Proc. Eurospeech '97, pp. 37–40.
Martin, R., 1993. An efficient algorithm to estimate the instantaneous SNR of speech signals. In: Proc. Eurospeech '93, pp. 1093–1096.
Miller, G.A., Licklider, J.C.R., 1950. The intelligibility of interrupted speech. J. Acoust. Soc. Am. 22, 167–173.
Ming, J., Smith, F.J., 1999. Union: a new approach for combining sub-band observations for noisy speech recognition. In: Proceedings of the Workshop on Robust Methods for Speech recognition in Adverse Conditions, Tampere, Finland, May '99, pp. 175–178.
Mokbel, C., 1992. Reconnaissance de la parole le bruit: bruitage/debruitage. Ecole Nationale Superiore de Telecommunications.
Moore, B.C.J., 1997. An Introduction to the Psychology of Hearing, 4th ed. Academic Press, New York.
Morris, A.C., Cooke, M.P., Green, P.D., 1998. Some solutions to the missing feature problem in data classification, with application to noise-robust ASR. In: ICASSP-98, Seattle.
Morris, A.C., Hagen, A., Bourlard, H., 1999. The full combination sub-bands approach to noise robust hmm/ann-based asr. In: Proc. Eur. Conf. Speech Commun. Technol., pp. 599–602.
Morrison, D.F., 1990. Multivariate Statistical Methods. 3rd Ed. McGraw-Hill.
Nadeu, C., Pachés-Leal, P., Juang, B., 1997. Filtering the time sequences of spectral parameters for speech recognition. Speech Communication 22, 315–322.
Proceedings of the NIPS-95 workshop on Missing data: Methods and Models, Denver. MIT Press, Cambridge, MA, 1996.

Raj, B., Singh, R., Stern, M., 1998. Inference of missing spectrographic features for robust speech recognition. In: Proc. ICSLP '98, pp. 1491–1494.

Steeneken, H.J.M., 1992. On measuring and predicting speech intelligibility. Ph.D. thesis, University of Amsterdam, unpublished.

Strange, W., Jenkins, J.J., Johnson, T.L., 1983. Dynamic specification of coarticulated vowels. J. Acoust. Soc. Am. 74 (3), 695–705.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. ICASSP '90, pp. 845–848.

Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical Report, Speech Research Unit, Defence Research Agency, Malvern, UK.

Varga, A.P., Moore, R.M., Bridle, J.S., Ponting, K., Russell, M.J., 1988. Noise compensation algorithms for use with hidden Markov model-based speech recognition. In: Proc. IEEE ICASSP, pp. 481–484.

Vizinho, A., Green, Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study. In: Proc. Eurospeech, 99, pp. 2407–2410.

von der Malsburg, C., Schneider, W., 1986. A neural cocktail-party processor. Biol. Cybern. 54, 29–40.

Warren, R.M., Riener, K.R., Bashford, J.A., Brubaker, B.S., 1995. Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. Perception Psychophys. 57 (2), 175–182.

Young, S.J., Woodland, P.C., 1993. HTK Version 1.5: User, Reference and Programmer Manual Cambridge University Engineering Department, Speech Group.