Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 5–25
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Decoding speech in the presence of other sources

J.P. Barker [a,*,1], M.P. Cooke [a,1], D.P.W. Ellis [b]

[a] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
[b] Department of Electrical Engineering, Columbia University, 500 W. 120th Street, New York, NY 10027, USA

Received 7 June 2002; received in revised form 21 November 2003; accepted 20 May 2004

Abstract

The statistical theory of speech recognition introduced several decades ago has brought about low word error rates for clean speech. However, it has been less successful in noisy conditions. Since extraneous acoustic sources are present in virtually all everyday speech communication conditions, the failure of the speech recognition model to take noise into account is perhaps the most serious obstacle to the application of ASR technology.

Approaches to noise-robust speech recognition have traditionally taken one of two forms. One set of techniques attempts to estimate the noise and remove its effects from the target speech. While noise estimation can work in low-to-moderate levels of slowly varying noise, it fails completely in louder or more variable conditions. A second approach utilises noise models and attempts to decode speech taking into account their presence. Again, model-based techniques can work for simple noises, but they are computationally complex under realistic conditions and require models for all sources present in the signal.

In this paper, we propose a statistical theory of speech recognition in the presence of other acoustic sources. Unlike earlier model-based approaches, our framework makes no assumptions about the noise background, although it can exploit such information if it is available. It does not require models for background sources, or an estimate of their number. The new approach extends statistical ASR by introducing a segregation model in addition to the conventional acoustic and language models. While the conventional statistical ASR problem is to find the most likely sequence of speech models which generated a given observation sequence, the new approach additionally determines the most likely set of signal fragments which make up the speech signal. Although the framework is completely general, we provide one interpretation of the segregation model based on missing-data theory. We derive an efficient HMM decoder, which searches both across subword state and across alternative segregations of the signal between target and interference. We call this modified system the *speech fragment decoder*.

The value of the speech fragment decoder approach has been verified through experiments on small-vocabulary tasks in high-noise conditions. For instance, in a noise-corrupted connected digit task, the new approach decreases the word error rate in the condition of factory noise at 5 dB SNR from over 59% for a standard ASR system to less than 22%.

* Corresponding author. Tel.: +44 114 222 1824; fax: +44 114 222 1810.
E-mail address: j.barker@dcs.shef.ac.uk (J.P. Barker).
[1] Partially supported by EPSRC grant GR/47400/01.

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.05.002

Keywords: Robust speech recognition; Signal separation; Missing data recognition; Computational auditory scene analysis; Acoustic mixtures

1. Introduction

In the real world, the speech signal is frequently accompanied by other sound sources on reaching the auditory system, yet listeners are capable of holding conversations in a wide range of listening conditions. Recognition of speech in such 'adverse' conditions has been a major thrust of research in speech technology in the last decade. Nevertheless, the state of the art remains primitive. Recent international evaluations of noise robustness have demonstrated technologically useful levels of performance for small vocabularies in moderate amounts of quasi-stationary noise (Pearce and Hirsch, 2000). Modest departures from such conditions lead to a rapid drop in recognition accuracy.

A key challenge, then, is to develop algorithms to recognise speech in the presence of arbitrary non-stationary sound sources. There are two broad categories of approaches to dealing with interference for which a stationarity assumption is inadequate. *Source-driven* techniques exploit evidence of a common origin for subsets of source components, while *model-driven* approaches utilise prior (or learned) representations of acoustic sources. Source-driven approaches include primitive auditory scene analysis (Brown and Cooke, 1994; Wang and Brown, 1999; see review in Cooke and Ellis, 2001) based on auditory models of pitch and location processing, independent component analysis and blind source separation (Bell and Sejnowski, 1995; Hyvärinen and Oja, 2000) which exploit statistical independence of sources, and mainstream signal processing approaches (Parsons, 1976; Denbigh and Zhao, 1992). The prime examples of model-driven techniques are HMM decomposition (Varga and Moore, 1990) and parallel model combination (PMC) (Gales and Young, 1993), which attempt to find model state sequence combinations which jointly explain the acoustic observations. Ellis' 'prediction-driven' approach (Ellis, 1996) can also be regarded as a technique influenced by prior expectations.

Pure source-driven approaches are typically used to produce a clean signal which is then fed to an unmodified recogniser. In real-world listening conditions, this segregate-then-recognise approach fails (see also the critique in Slaney, 1995), since it places too heavy a demand on the segregation algorithm to produce a signal suitable for recognition. Conventional recognisers are highly sensitive to the kinds of distortion resulting from poor separation. Further, while current algorithms do a reasonable job of separating periodic signals, they are less good both at dealing with the remaining portions and extrapolating across unvoiced regions, especially when the noise background contains periodic sources. The problem of distortion can be solved using missing data (Cooke et al., 1994, 2001) or multiband (Bourlard and Dupont, 1997) techniques, but the issue of sequential integration across aperiodic intervals remains.

Pure model-driven techniques also fail in practice, due to their reliance on the existence of models for all sources present in a mixture, and the computational complexity of decoding multiple sources for anything other than sounds which possess a simple representation.

There is evidence that listeners too use a combination of source and model driven processes (Bregman, 1990). For instance, vowel pairs presented concurrently on the same fundamental can be recognised at levels well above chance, indicating the influence of top-down model-matching behaviour, but even small differences in fundamental—which create a source-level cue—lead to significant improvements in identification indicating that the model-driven search is able efficiently to exploit the added low-level information (Scheffers, 1983). Similarly, when the first three speech formants are replaced by sinusoids, listeners recognise the resulting sine-wave speech at levels approach-

ing natural speech, generally taken as evidence of a purely top-down speech recognition mechanism, since the tokens bear very little resemblance to speech at the signal level (Bailey et al., 1977; Remez et al., 1981). However, when presented with a sine-wave cocktail party consisting of a pair of simultaneous sine-wave sentences, performance falls far below the equivalent natural speech sentence-pair condition, showing that low-level signal cues are required for this more demanding condition (Barker and Cooke, 1999).

In this paper, we present a framework which attempts to integrate source- and model-driven processes in robust speech recognition. We demonstrate how the decoding problem in ASR can be extended to incorporate decisions about which regions belong to the target signal. Unlike pure source-driven approaches, the integrated decoder does not require a single hard-and-fast prior segregation of the entire target signal, and, in contrast to pure model-based techniques, it does not assume the existence of models for all sources present. Since it is an extension of conventional speech decoders, it maintains all of the advantages of the prevailing stochastic framework for ASR by delaying decisions until all relevant evidence has been observed. Furthermore, it allows a tradeoff between the level of detail derived from source-driven processing and decoding speed.

Fig. 1 motivates the new approach. The upper panel shows an auditory spectrogram of the utterance "two five two eight three" spoken by a male speaker mixed with drum beats at a global SNR of 0 dB. The centre panel segments the time-frequency plane into regions, which are dominated (in the sense of possessing a locally-favourable SNR) by one or other source. The correct assignment of regions to the two sources is shown in the lower panel.

In outline, our new formalism defines as an admissible search over all combinations of regions (which we call *fragments*) to generate the most likely word sequence (or, more generally, sequence of source models). This is achieved by decomposing the likelihood calculation into three parts: in addition to the conventional language model term, we introduce a *segregation model*, which defines how fragments are formed, and a *modified acoustic model*, which links the observed acoustics to source models acquired during training.

Section 2 develops the new formalism, and shows how the segregation model and partial acoustic model can be implemented in practice. Section 3 demonstrates the performance of the resulting decoder applied to digit strings with added noise. Section 4 discusses issues that have arisen with the current decoder implementation and future research directions.

2. Theoretical development

The simultaneous segregation/recognition approach can be formulated as an extension of the existing speech recognition theory. When formulated in a statistical manner, the goal of the speech recogniser is traditionally stated as to find the word sequence $\widehat{W} = w_1, w_2, \ldots, w_N$ with the maximum *a posteriori* probability given the sequence of acoustic feature vectors observed for the speech, $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} P(W|\boldsymbol{X}). \tag{1}$$

This equation is rearranged using Bayes' rule into

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \frac{P(\boldsymbol{X}|W)P(W)}{P(X)}, \tag{2}$$

which separates the prior probability of the word sequence alone $P(W)$ (the language model), the distribution of the speech features for a particular utterance, $P(\boldsymbol{X}|W)$ (the acoustic model), and the prior probability of those features $P(X)$ (which is constant over W and thus will not influence the outcome of the argmax). $P(W)$ may be trained from the word sequences in a large text corpus, and $P(\boldsymbol{X}|W)$ is learned by modelling the distribution of actual speech features associated with particular sounds in a speech training corpus.

Following our considerations above, we may restate this goal as finding the word sequence, $\widehat{W}$,

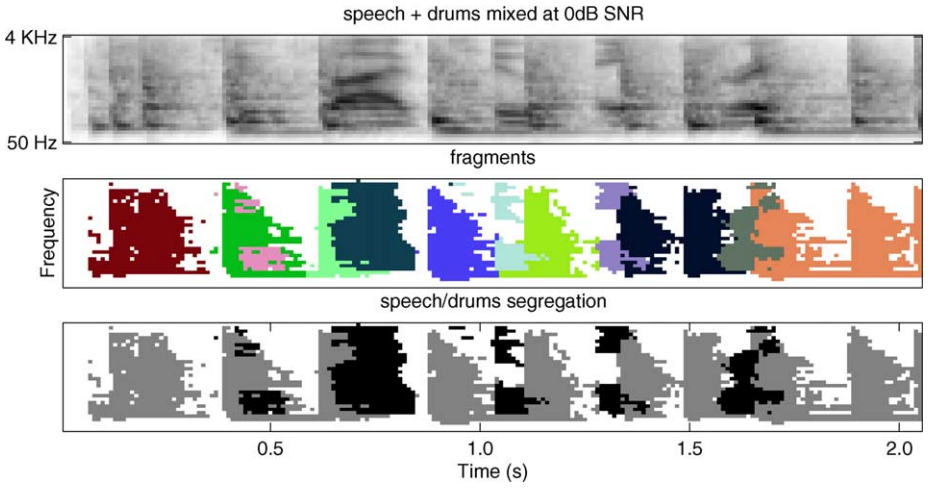

Fig. 1. The top panel shows the auditory spectrogram of the utterance "two five two eight three" spoken by a male speaker mixed with drum beats at 0 dB SNR. The lower panel shows the correct segregation of speech energy (black) and drums energy (grey). The centre panel illustrates the set of fragments generated using knowledge of the speech source and the noise source prior to mixing.

along with the speech/background segregation, $\widehat{S}$, which jointly have the maximum posterior probability.

Further, because the observed features are no longer purely related to speech but in general include the interfering acoustic sources, we will denote them as $\boldsymbol{Y}$ to differentiate them from the $\boldsymbol{X}$ used in our speech-trained acoustic models $P(\boldsymbol{X}|W)$. [2]

$$\widehat{W},\widehat{S} = \underset{W,S}{\operatorname{argmax}}\, P(W,S|\boldsymbol{Y}). \tag{3}$$

To reintroduce the speech features $\boldsymbol{X}$, which are now an unobserved random variable, we integrate the probability over their possible values, and decompose with the chain rule to separate out $P(S|\boldsymbol{Y})$, the probability of the segregation based on the observations:

$$P(W,S|\boldsymbol{Y}) = \int P(W,\boldsymbol{X},S|\boldsymbol{Y})\,\mathrm{d}\boldsymbol{X} \tag{4}$$

$$= \int P(W|\boldsymbol{X},S,\boldsymbol{Y})P(\boldsymbol{X}|S,\boldsymbol{Y})\,\mathrm{d}\boldsymbol{X}\cdot P(S|\boldsymbol{Y}). \tag{5}$$

Since W is independent of S and $\boldsymbol{Y}$ given $\boldsymbol{X}$, the first probability simplifies to $P(W|\boldsymbol{X})$. As in the standard derivation, we can rearrange it via Bayes' rule to obtain a formulation in terms of our trained distribution models $P(W)$ and $P(\boldsymbol{X}|W)$:

$$P(W,S|\boldsymbol{Y}) = \int \frac{P(\boldsymbol{X}|W)P(W)}{P(\boldsymbol{X})} P(\boldsymbol{X}|S,\boldsymbol{Y})\,\mathrm{d}\boldsymbol{X}\cdot P(S|\boldsymbol{Y}) \tag{6}$$

$$= P(W)\left(\int P(\boldsymbol{X}|W)\frac{P(\boldsymbol{X}|S,\boldsymbol{Y})}{P(\boldsymbol{X})}\,\mathrm{d}\boldsymbol{X}\right)P(S|\boldsymbol{Y}). \tag{7}$$

Note that because $\boldsymbol{X}$ is no longer constant, we cannot drop $P(\boldsymbol{X})$ from the integral.

In the case of recognition with hidden Markov models (HMMs), the conventional derivation introduces an unobserved state sequence $Q = q_1, q_2, \ldots, q_T$ along with models for the joint probability of word sequence and state sequence $P(W,Q) = P(Q|W)P(W)$. The Markovian assumptions include making the feature vector $\boldsymbol{x}_i$ at time i depend only on the corresponding state q_i, making $P(\boldsymbol{X}|Q) = \prod_i P(\boldsymbol{x}_i|q_i)$. The total likelihood of a particular W over all possible state sequences is nor-

[2] Note, if we were not interested in the speech/background segregation but only in the most likely word sequence regardless of the actual segregation then it would be more correct to integrate Eq. (3) over the segregation space defining $W' = \operatorname{argmax}_W \sum_S P(W,S|\boldsymbol{Y})$. However, this integration presents some computational complexity so in practice even if we were not directly interested in the segregation it may be desirable to implement Eq. (3) directly and take $\widehat{W}$ as an approximation of W'.

mally approximated by the score over the single most-likely state sequence (the *Viterbi* path). In our case, this gives,

$$\widehat{W},\widehat{S} = \underset{W,S}{\operatorname{argmax}} \max_{Q\in Q_W} P(S|\boldsymbol{Y})P(W)P(Q|W) \times \int P(\boldsymbol{X}|Q)\frac{P(\boldsymbol{X}|S,\boldsymbol{Y})}{P(\boldsymbol{X})}\,\mathrm{d}\boldsymbol{X}, \quad (8)$$

where Q_W represents the set of all allowable state sequences corresponding to word sequence W.

Compare Eq. (8) to the corresponding equation for identifying the word sequence in a conventional speech recogniser:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \max_{Q\in Q_W} P(W)P(Q|W)P(\boldsymbol{X}|Q). \quad (9)$$

It can be seen that there are three significant differences:

(1) A new term, $P(S|\boldsymbol{Y})$ has been introduced. This is the 'segregation model', describing the probability of a particular segregation S given our actual observations $\boldsymbol{Y}$, but independent of the word hypothesis W—precisely the kind of information we expect to obtain from a model of source-driven, low-level acoustic organisation.
(2) The acoustic model score $P(\boldsymbol{X}|Q)$ is now evaluated over a range of possible values for $\boldsymbol{X}$, weighted by their relative likelihood given the observed signal Y and the particular choice of segregation mask S. This is closely related to previous work on missing data theory, and is discussed in more detail in Section 2.3.
(3) The maximisation now occurs over both W and S. Whereas conventional speech recognition searches over the space of words sequences, the extended approach has to simultaneously search over the space of all admissible segregations.

In the terms of Bregman's 'Auditory Scene Analysis' account (Bregman, 1990), the segregation model may be identified as embodying the so-called 'primitive grouping process', and the acoustic model plays the part of the 'schema-driven grouping process'. Eq. (8) serves to integrate these two complementary processes within the probabilistic framework of ASR. The maximisation over W and S can be achieved by extending the search techniques employed by traditional ASR. These three key aspects of the work, namely, the *segregation model*, the *acoustic model* and the *search problem* are addressed in greater detail in the sections which follow (Fig. 2).

2.1. The segregation model

Consider the space of potential speech/background segregations. An acoustic observation vector, $\boldsymbol{X}$ may be constructed as a sequence of frames $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T$ where each frame is composed of observations pertaining to a series of, say F, frequency channels. The observation vector is therefore composed of $T\times F$ spectro-temporal features. A speech/background segregation may be conveniently described by a binary mask in which the label '1' is employed to signify that the feature belongs to the speech source, and a '0' to signify that the feature belongs to the background. As this binary mask has $T\times F$ elements it can be seen that there are 2^{TF} possible speech/background segregations. So, for example, at a typical frame rate of 100 Hz, and with a feature vector employing 32 frequency channels, there would be 2^{3200}

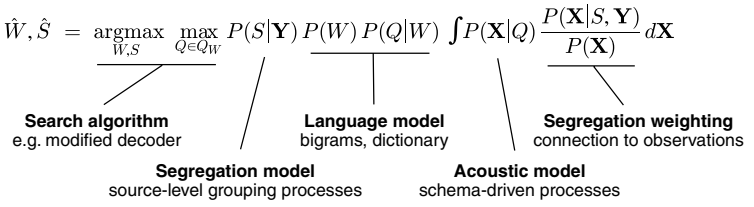

Fig. 2. An overview of the speech fragment decoding equation.

possible segregations for a one second audio sample.

Fortunately, most of these segregations can be ruled out immediately as being highly unlikely and the size of the search space can be drastically reduced. The key to this reduction is to identify spectro-temporal regions for which there is strong evidence that all the spectro-temporal pixels contained are dominated by the same sound source. Such regions constrain the spectro-temporal pixels contained to share the same speech/background label. Hence, for each permissible speech/background segregation, the pixels within any given fragment must either all be labelled as speech (meaning that the fragment is part of the speech source) or must all be labelled as background (meaning that the fragment is part of some other source). Consequently, if the spectro-temporal observation vector can be decomposed into N such fragments, there will be 2^N separate ways of labelling the fragments and hence only 2^N valid segregations. In general each fragment will contain many spectro-temporal pixels, and 2^N will be vastly smaller than the size of the unconstrained segmentation search space, 2^{TF}.

The success of the segregation model depends on being able to identify a reliable set of coherent fragments. The process of dissecting the representation into fragments is similar to the process that occurs in visual scene analysis. The first stage of interpreting a visual scene is to locate regions within the scene that are components of larger objects. For this purpose all manner of primitive processes may be employed: edge detection, continuity, uniformity of colour, uniformity of texture etc. Analogous processes may be used in the analysis of the auditory 'scene', for example, spectro-temporal elements may be grouped if they form continuous tracks (i.e. akin to visual edge detection), tracks may be grouped if they lie in harmonic relation, energy regions may grouped across frequency if they onset or offset at the same time. Fig. 3 illustrates some of the mechanisms that may be used to bind spectro-temporal regions to recover partial descriptions of the individual sound sources. A detailed account of these so-called 'primitive grouping processes' is given in (Bregman, 1990).

In the experiments that follow, each of the 2^N valid segregations is allocated an equal prior probability. This stands as a reasonable first approximation. However, a more detailed segregation model could be constructed in which the segregation priors vary across segregations. Such a model would take into account factors like the relationship between the individual fragments of which they are composed. For example, if there are two fragments which cover spectro-temporal regions in which the acoustic data is periodic and has the same fundamental frequency, then these two fragments are likely to be parts of the same sound source, and hence segregations in which they are labelled as either both speech or both background should be favoured. Section 4.3 discusses further such 'between-fragment grouping' effects and of the modifications to the search algorithm that they require.

2.2. The search problem

The task of the extended decoder is to find the most probable word sequence and segregation given the search space of all possible word sequences and all possible segregations. Given that the acoustic match score:

$$P(\boldsymbol{X}|Q)P(\boldsymbol{X}|S,\boldsymbol{Y})/P(\boldsymbol{X}), \tag{10}$$

is conditioned both on the segregation S and the subword state Q, the (S,Q) search space cannot in general be decomposed into independent searches over S and Q. Since the size of the S space expands the overall search space it is imperative that the search in the plane of the segregation space is conducted as efficiently as possible.

To illustrate this point, imagine a naive implementation of the search illustrated in Fig. 4. In this approach, each segregation hypothesis is considered independently, and therefore requires a separate word sequence search. If the segregation model has identified N coherent fragments, then there will be 2^N segregation hypotheses to consider. Hence, the total computation required for the decoding will scale exponentially with the number of fragments. The total number of fragments is likely to be a linear function of the duration of the acoustic mixture being processed, therefore the

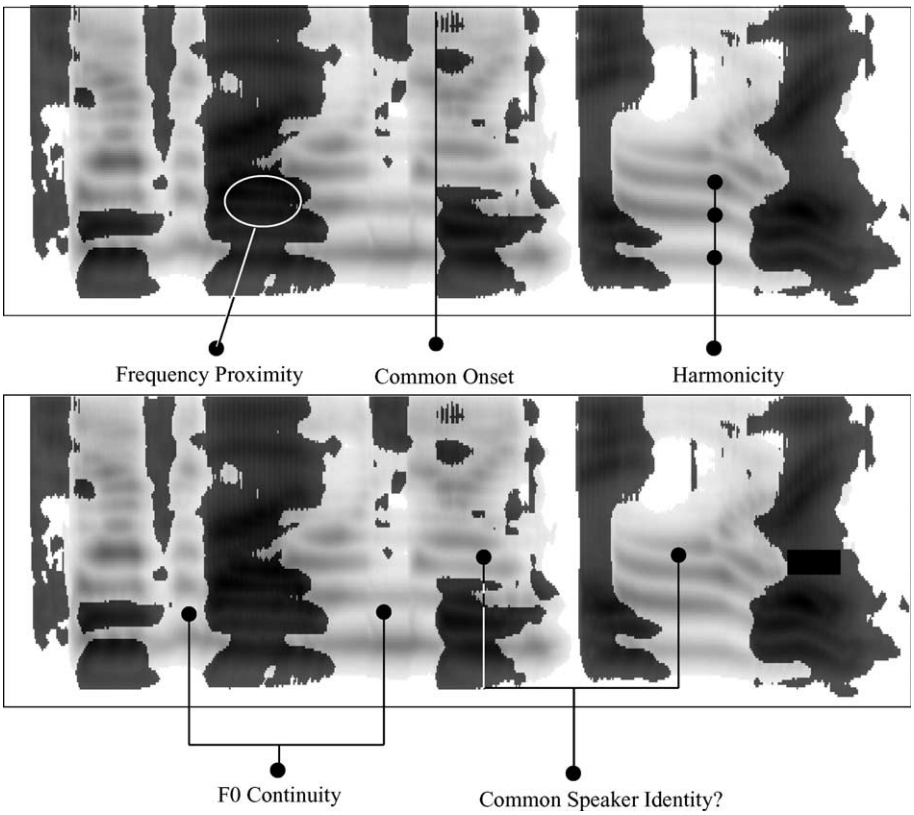

Fig. 3. An illustration of short-term (above) and long-term (below) primitive grouping cues which may be exploited to recover partial descriptions of individual sound sources. The figure shows a time-frequency representation of two simultaneous speech utterances. Regions where the energy of one source dominate are shown in dark grey, while those of the other source are in light grey.

computation required will be an exponential function of this duration. For sufficiently large vocabularies, the cost of decoding the word sequence typically makes up the greater part of the total computational cost of ASR. It is clear that the naive implementation of the word sequence/segregation search is unacceptable unless the total number of fragments is very small.

The key to constructing an efficient implementation of the search is to take advantage of similarities that exist between pairs of segregation hypotheses. Consider the full set of possible segregations. There is a unique segregation for every possible assignment of speech/background labelling to the set of fragments. For any given pair of hypotheses, some fragments will have the same label. In particular, some hypotheses will differ only in the labelling of a single fragment. For such pairs, the speech/background segregation will be identical up to the time frame where the differing fragment onsets, and identical again from the frame where the fragment offsets. The brute-force search performs two independent word sequence searches for two such similar segregation hypotheses (see Fig. 5, column 1). The computational cost of these two independent searches may be reduced by allowing them to share processing up to the time frame where the segregation hypotheses differ—i.e. the onset of the fragment that is labelled differently in each hypothesis, marked as time T1 in column 2 of Fig. 5. This sharing of computation between pairs of segregation hypotheses can be generalised to encompass all segregation hypotheses by arranging them in a graph structure. As we

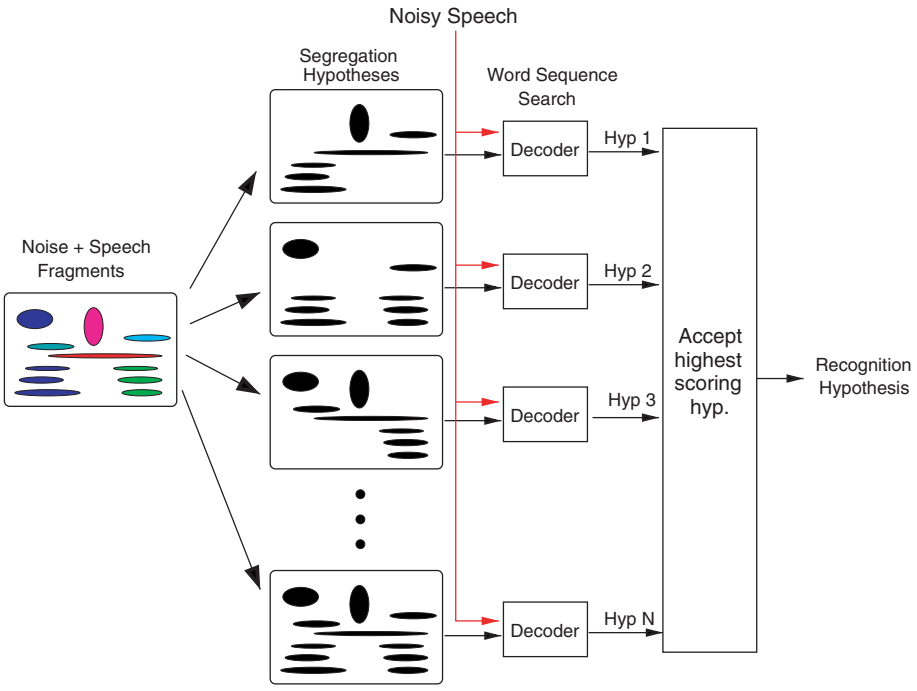

Fig. 4. The figure illustrates a naive implementation of the segregation/word-sequence search. From a set of N noise and speech fragments, 2^N speech/noise segregation hypotheses can be generated. It is then possible to search for the best word sequence given each of these 2^N segregation hypotheses. The overall best hypothesis can then be found by comparing the scores of these independent searches.

progress through time, new fragment onsets cause all current segregation hypotheses to branch, forming two complementary sets of paths. In one set, the onsetting fragment is considered to be speech while in the other it is considered to be background. However, although this arrangement saves some computation, the number of segregation hypotheses under consideration at any particular frame still grows exponentially with time. This exponential growth may be prevented by noting that segregation hypotheses will become identical again after the offset of the last fragment by which they differ (marked as time T2 in column 3 of Fig. 5). At this point, the two competing segregation hypotheses can be compared and the least likely of the pair can be rejected without affecting the admissibility of the search. Again, this step can be generalised to encompass all segregation hypotheses and effectively brings together the branches of the diverging segregation hypothesis tree.

Fig. 6 illustrates the evolution of a set of parallel segregation hypotheses while processing a segment of noisy speech which has been dissected into three fragments (shown schematically by the shaded regions in the figure). When the first fragment (white) commences, two segregation hypotheses are formed. In one hypothesis, the white fragment is labelled as speech, while in the other it is assigned to the background. When the grey fragment starts, all ongoing hypotheses are again split with each pair covering both possible labellings for the grey fragment. When the white fragment ends, pairs of hypotheses are merged if their labelling only differs with regard to the white fragment. This pattern of splitting and merging continues until the end of the utterance. Note that at any instant there are at most four active segregation hypotheses, not

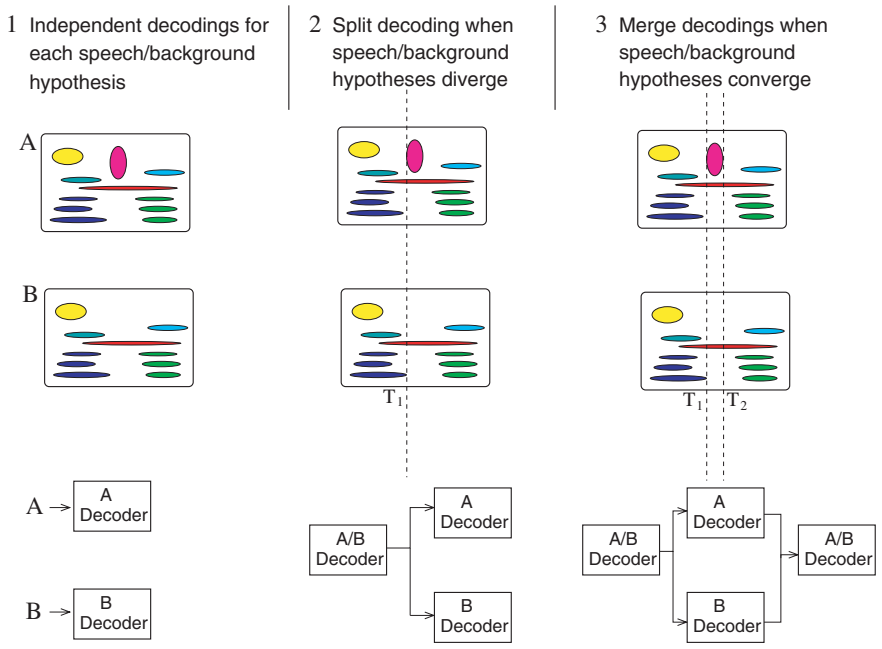

Fig. 5. The efficient segregation search exploits the fact that competing segregation hypotheses only differ over a limited number of frames.

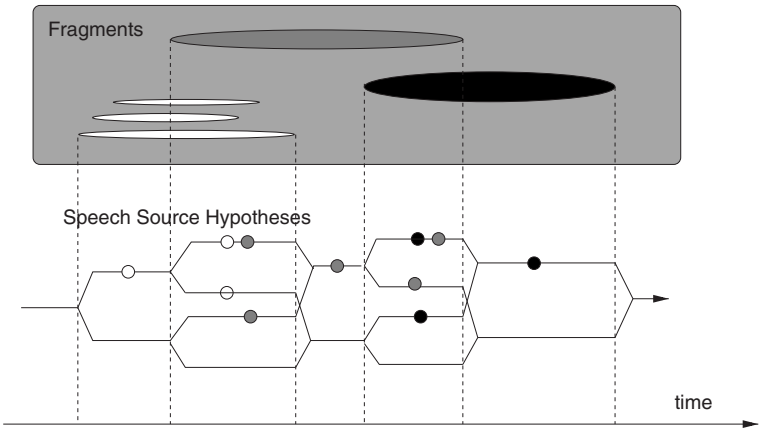

Fig. 6. The evolution of a set of segregation hypotheses. Each parallel path represents a separate hypothesis, with the shaded dots indicating which ongoing fragments are being considered as speech part of the speech source.

the eight required to consider every possible labelling of each of the three fragments.

It is important to understand that the evolution of segregation hypotheses is dependent on the word sequence hypothesis. For each ongoing word sequence being considered by the decoder, a particular corresponding optimal segregation is simultaneously developed.

If the word sequence is modelled using HMMs, then the segregation/word-sequence decoder can be implemented by extending the token-passing Viterbi algorithm employed in conventional ASR:

- Tokens keep a record of the fragment assignments they have made, i.e. each token stores its labelling of each fragment encountered as either *speech* or *background*.
- *Splitting*: When a new fragment starts all existing tokens are duplicated. In one copy the new fragment is labelled as speech and in the other it is labelled as background.
- *Merging*: When a fragment ends, then for each state we compare tokens that differ only in the label of the fragment that is ending. The less likely token or tokens are deleted.
- At each time frame, tokens propagate through the HMM as usual. However, each state can hold as many tokens as there are different labellings of the currently active fragments. When tokens enter a state only those with the same labelling of current active fragments are directly compared. The token with the highest likelihood score survives and the others are deleted.

It should be stressed that the deletion of tokens in the 'merging' step described above does not affect the admissibility of the search (i.e. it is not a form of hypothesis pruning). The efficient algorithm will return an identical result to that of a brute-force approach, which separately considers every word-sequence/segregation hypothesis. This is true as long as the Markov assumption remains valid. In the context of the above algorithm this means that the future of a partial hypothesis must be independent of its past. This places some constraints on the form of the segregation model. For example, the Markov assumption may break down if the segregation model contains between-fragment grouping effects in which the future scoring of a partial hypothesis may depend on which groups it has previously interpreted as part of the speech source. In this case the admissibility of the search can be preserved by imposing extra constraints on the hypothesis merging condition. This point is discussed further in Section 4.3.

2.3. The acoustic model

In Eq. (8), the acoustic model data likelihood $P(\boldsymbol{X}|Q)$ of a conventional speech recogniser is replaced by an integral over the partially-observed speech features $\boldsymbol{X}$, weighted by a term conditioned on the observed signal features $\boldsymbol{Y}$ and the segregation hypothesis S:

$$\int P(\boldsymbol{X}|Q)\frac{P(\boldsymbol{X}|S,\boldsymbol{Y})}{P(\boldsymbol{X})}\mathrm{d}\boldsymbol{X}, \tag{11}$$

where $P(\boldsymbol{X}|Q)$ is the feature distribution model of a conventional recogniser trained on clean speech, and $P(\boldsymbol{X}|S,\boldsymbol{Y})/P(\boldsymbol{X})$ is a likelihood weighting factor introducing the influence of the particular (noisy) observations $\boldsymbol{Y}$ and the assumed segregation S.

The integral over the entire space of $\boldsymbol{X}$—the full multidimensional feature space at every time step—is clearly impractical. Fortunately, it can be broken down into factors. Firstly, the Markov assumption of independent emissions given the state sequence allows us to express the likelihood of the sequence as the product of the likelihoods at each time step i: [3]

$$\int P(\boldsymbol{X}|Q)\frac{P(\boldsymbol{X}|S,\boldsymbol{Y})}{P(\boldsymbol{X})}\mathrm{d}\boldsymbol{X} = \prod_i \int P(\boldsymbol{x}_i|q_i)\frac{P(\boldsymbol{x}_i|S,\boldsymbol{Y})}{P(\boldsymbol{x}_i)}\mathrm{d}\boldsymbol{x}_i. \tag{12}$$

Secondly, in a continuous-density (CDHMM) system, $P(\boldsymbol{x}|q)$ is modelled as a mixture of M multivariate Gaussians, usually each with a diagonal covariance matrix:

$$P(\boldsymbol{x}|q) = \sum_{k=1}^{M} P(k|q)P(\boldsymbol{x}|k,q), \tag{13}$$

where $P(k|q)$ are the mixing coefficients. Since the individual dimensions of a diagonal-covariance Gaussian are independent, we can further factorise the likelihood over the feature vector elements x_j:

$$P(\boldsymbol{x}|q) = \sum_{k=1}^{M} P(k|q)\prod_j P(x_j|k,q). \tag{14}$$

[3] This also assumes independence of each time step for the prior $P(\boldsymbol{X})$ and for the likelihood of X given the segregation hypothesis and observations, $P(\boldsymbol{X}|S,\boldsymbol{Y})$. Both these assumptions are open to serious question, and we return to them in Section 4.

Assuming a similar decomposition of the prior $P(X)$, we can take the integral of Eq. (12) inside the summation to give:

$$\int P(\boldsymbol{x}|q)\frac{P(\boldsymbol{x}|S,\boldsymbol{Y})}{P(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \sum_{k=1}^{M} P(k|q)\prod_j \int P(x_j|k,q)\frac{P(x_j|S,\boldsymbol{Y})}{P(x_j)}\mathrm{d}x_j, \tag{15}$$

where $P(x_j|k,q)$ is now a simple unidimensional Gaussian.

We can consider the factor

$$\frac{P(x_j|S,\boldsymbol{Y})}{P(x_j)} \tag{16}$$

as the 'segregation weighting'—the factor by which the prior probability of a particular value for the speech feature is modified in light of the segregation mask and the observed signal. Since we are working with models of subband spectral energy, we can use a technique closely related to the missing-data idea of *bounded integration* (Cooke et al., 2001): For subbands that are judged to be dominated by speech energy (i.e., under the segregation hypothesis S, not one of the 'masked' channels), the corresponding feature values x_k can be calculated directly [4] from the observed signal $\boldsymbol{Y}$ and hence the segregation weighting will be a Dirac delta at the calculated value, x^*:

$$P(x_j|S,\boldsymbol{Y}) = \delta(x_j - x^*), \tag{17}$$

$$\int P(x_j|k,q)\frac{P(x_j|S,\boldsymbol{Y})}{P(x_j)}\mathrm{d}x_j = P(x^*|k,q)/P(x^*). \tag{18}$$

[4] The observed signal $\boldsymbol{Y}$ will in general be a richer representation than simply the subband energies that would have formed $\boldsymbol{x}$ in the noise-free case, since it may include information such as spectral fine-structure used to calculate pitch cues used in low-level segregation models, etc. However, the information in $\boldsymbol{x}$ will be completely defined given $\boldsymbol{Y}$ in the case of a segregation hypothesis that rates the whole spectrum as unmasked for that time slice.

The other case is that the subband corresponding to x is regarded as masked under the segregation hypothesis. We can still calculate the spectral energy x^* for that band, but now we assume that this level describes the masking signal, and the speech feature is at some unknown value smaller than this. In this case, we can model $P(x|S,\boldsymbol{Y})$ as proportional to the prior $P(x)$ for $x \leqslant x^*$, and zero for $x > x^*$. Thus,

$$P(x_j|S,\boldsymbol{Y}) = \begin{cases} F\cdot P(x_j) & x_j \leqslant x^*, \\ 0 & x_j > x^*, \end{cases} \tag{19}$$

$$\int P(x_j|k,q)\frac{P(x_j|S,\boldsymbol{Y})}{P(x_j)}\mathrm{d}x_j = \int_{-\infty}^{x^*} P(x_j|k,q)\cdot F\,\mathrm{d}x_j, \tag{20}$$

where F is a normalisation constant to keep the truncated distribution a true pdf i.e.

$$F = \frac{1}{\int_{-\infty}^{x^*} P(x_j)\,\mathrm{d}x_j} \tag{21}$$

In Eq. (20), the likelihood gets smaller as more of the probability mass associated with a particular state lies in the range precluded by the masking level upper bound; it models the 'counterevidence' (Cunningham and Cooke, 1999) against a particular state. For example, given a low x^* the quieter states will score better then more energetic ones. Since the elemental distributions $P(x_j|k,q)$ are simple Gaussians, each integral is evaluated using the standard error function.

Both the scaling factor F in Eq. (20) and the evaluation of the point-likelihood in Eq. (18) require a value for the speech feature prior $P(x_j)$. In the results reported below we have made the very simple assumption of a uniform prior on our cube-root compressed energy values between zero and some fixed maximum $x_{\max}$, constant across all feature elements and intended to be larger than any actual observed value. This makes the prior likelihood $P(x_j)$ equal a constant $1/x_{\max}$ and $F = x_{\max}/x^* \propto 1/x^*$.

Using Eq. (18) for the unmasked dimensions and Eq. (20) for the masked dimensions we can evaluate the acoustic data likelihood (or 'acoustic match score') for a single state at a particular time slice with Eq. (15) which becomes:

$$\int P(\boldsymbol{x}|q)\frac{P(\boldsymbol{x}|S,\boldsymbol{Y})}{P(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \sum_{k=1}^{M} P(k|q)\prod_{j\in S_O} P(x_j^*|k,q)\cdot x_{\max}\prod_{j\in S_M} \times \int P(x_j|k,q)\cdot\frac{x_{\max}}{x_j^*}\mathrm{d}x_j, \quad (22)$$

where S_O is the set of directly observed (not masked) dimensions of $\boldsymbol{x}$, S_M are the remaining, masked, dimensions, and x_j^* is the observed spectral energy level for a particular band j. This per-time likelihood can then be combined across all timeslices using Eq. (12) to give the data likelihood for an entire sequence.

In practise it has been observed that Eq. (22) exhibits a bias towards favouring hypotheses in which too many fragments have been labelled as background, or alternatively towards hypotheses in which too many fragments have been labelled as speech. The reasons for this bias are presently unclear, but one possibility is that is it introduced by the uniform prior employed for $P(x_j)$. As an approximate solution to this problem, the results of the integrations across the masked dimensions are scaled by a tuning parameter α shifting the relative likelihood of the missing and present dimensions. Giving α a high value tunes the decoder toward favouring hypotheses in which more fragments are labelled as background, while a low value favours hypotheses in which more fragments are labelled as speech. Experience has shown that the appropriate value of α depends largely on the nature of the fragments (i.e. the segregation model) and little on the noise type or noise level. Hence, it is easy to tune the system empirically using a small development data set.

Finally, it is instructive to compare the speech fragment decoding approach being proposed here with the missing data approach proposed in earlier work (Cooke et al., 1994, 2001). Basic missing data recognition consists of two separate steps performed in sequence: first, a 'present-data' mask is calculated, based, for instance, on estimates of the background noise level. Second, missing data recognition is performed by searching for the most likely speech model sequence consistent with this evidence. By contrast, the speech fragment decoding approach integrates these two steps, so that the search includes building the present-data mask to find the subset of features most likely to correspond to a single voice, while simultaneously building the corresponding most likely word sequence (Fig. 7).

2.4. Illustrative example

Fig. 8(A) shows the spectrogram of the utterance "seven five", to which a stationary background noise and a series of broadband high-energy noise bursts have been added (panel B). The initial frames of the signal can be employed to estimate and identify the stationary noise component, leaving the unmasked speech energy and the non-stationary noise bursts as candidate 'present data', as shown in panel C. This however must be broken up into a set of fragments to permit searching by the speech fragment decoder.

In order to confirm that the top-down process in the decoder is able to identify the valid speech

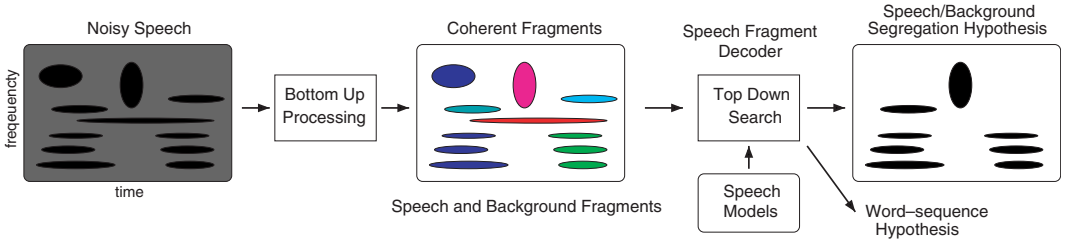

Fig. 7. An overview of the speech fragment decoding system. Bottom-up processes are employed to locate 'coherent fragments' (regions of representation that are due entirely to one source) and then a top-down search with access to speech models is used to search for the most likely combination of fragment labelling and speech model sequence.

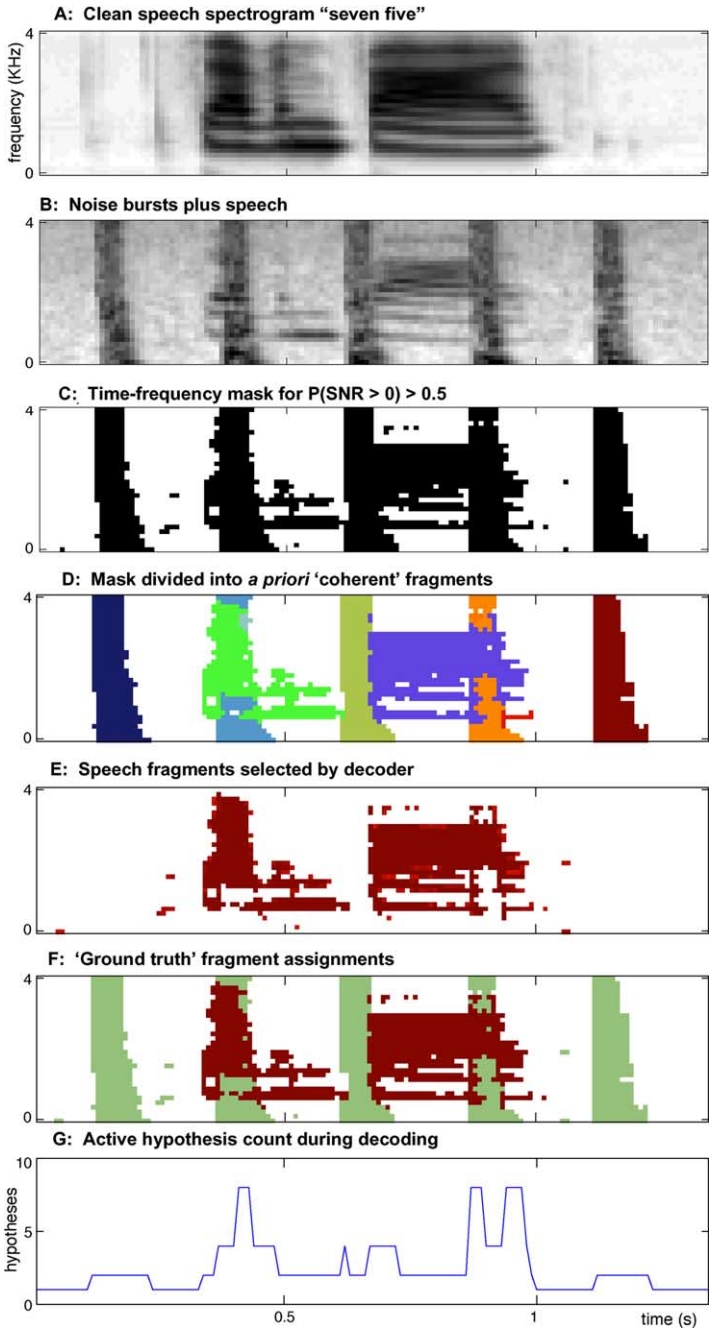

Fig. 8. An example of the speech fragment decoder's operation on a single noisy utterance: Panel A shows a spectrogram of the utterance ''seven five''. Panel B shows the same signal but after adding a two state noise source. Panel C shows the components of the mixture that are not accounted for by the adaptive background noise model. Panel D displays a test set of perfectly coherent fragments generated using a priori knowledge of the clean signal. Panel E shows the groups that the speech fragment decoder identifies as being speech groups. The correct assignment is shown in panel F. Panel G plots the number of grouping hypotheses that are being considered at each time frame.

fragments, its performance was tested using a small set of 'ideal' coherent fragments. These can be generated by applying a priori knowledge of the clean speech, i.e. comparing the clean and noisy spectrograms to mark out the exact regions where either the speech or the noise bursts dominate. The ideal fragments are simply the contiguous regions which are formed by this segregation process (see Panel D of Fig. 8).

Given these fragments, the decoder is able to correctly recognise the utterance as "seven five", using the fragments in panel E as evidence of the speech. The correct speech/noise fragment labelling is shown in panel F. Comparing E and F, it can be seen that the decoder has accepted all the speech fragments, while correctly rejecting all the larger fragments of noise. (Some small noise regions have been included in the speech, implying their level was consistent with the speech models.)

3. Experiments employing SNR-based fragments

The first set of experiments employ a connected digit recognition task and compare the performance of the speech fragment decoding technique with that of previously reported missing data techniques in which the speech/background segregation is effectively decided before proceeding with recognition (Cooke et al., 2001). The segregation model employed has been kept extremely simple. The coherent fragments are approximated directly from the acoustic mixture by using a simple noise estimation technique. The techniques presented here serve as a useful baseline against which the performance of more sophisticated segregation models can be compared.

3.1. Procedure

3.1.1. Feature vectors

The experiments in this section employ TIDigit utterances (Leonard, 1984) mixed with NOISEX factory noise (Varga et al., 1992) at various SNRs. NOISEX factory noise has a stationary background component but also highly unpredictable components such as hammer blows etc. which make it particularly disruptive for recognisers.

To produce the acoustic feature vectors the noisy mixtures were first processed with a 24 channel auditory filterbank (Cooke, 1991) 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 an 8 ms time constant, and sampled at a frame-rate of 10 ms. Finally, cube-root compression was applied to the energy values.

This forms a spectro-temporal sound energy representation that is suitable for segregation. This representation will henceforth be referred to as an 'auditory spectrogram'.

3.1.2. Fragments

The fragments were generated by the following steps:

(1) For each noisy utterance the first 10 frames of the auditory spectrogram are averaged to estimate a stationary noise spectrum. [5]
(2) The noise spectrum estimate is used to estimate the local SNR for each frame and frequency channel of the noisy utterance.
(3) The spectro-temporal region where the local SNR is above 0 dB is identified. This provides a rough approximation of the speech/background segregation.

If the additive noise source were stationary then the first three steps would provide the correct speech/background segregation and the speech fragment decoder technique would not be needed. However, if the competing noise source is non-stationary then some of the regions that are identified as speech will in fact be due to the noise. Hence, we now proceed with the following steps, which allow the speech fragment decoder technique to improve on the recognition result that would have been achieved if we had used the initial approximation to the speech/background segregation.

[5] This technique assumes that there is a delay before the speech source starts and hence the first frames provide a reliable measure of the noise background.

(4) The initial approximation of the speech segment is dissected by first dividing it into four frequency bands.
(5) Each contiguous region within each of the four subbands is defined to be a separate fragment.
(6) The set of fragments and the noisy speech representation are passed to the speech fragment decoder.

The fragmentation process is summarised in Fig. 9.

3.1.3. Acoustic models

An 8-state HMM was trained for each of the eleven words in the TIDigit corpus vocabulary (digits 'one' to 'nine', plus the two pronunciations of 0, namely 'oh' and 'zero'). The HMM states have two transitions each; a self transition and a transition to the following state. The emission distribution of each state was modelled by a mixture of 10 Gaussian distributions each with a diagonal covariance matrix. An additional 3-state HMM was used to model the silence occurring before and after each utterance, and the pauses that may occur between digits.

The scaling constant, α, required to balance missing and present data (see Section 2.3), was empirically tuned by maximising recognition performance on a small set of noisy utterances with an SNR of 10 dB. The value $\alpha = 0.3$ was found to give best performance. This value was then used for all noise levels during testing.

3.2. Artificial examples

As explained above, if the background noise is non-stationary the local SNR estimates (which have been based on the assumption that the noise is stationary), may be grossly inaccurate. A local peak in noise energy can lead to a spectro-temporal region that is mistakenly labelled as having high local SNR. This error then generates a region in the initial estimate of the speech/background segregation that is incorrectly identified as belonging to the speech source. If this segregation is used directly in conjunction with standard missing data techniques then the error will lead to poor recognition performance.

Fragmenting the initial speech segregation and applying the speech fragment decoder should allow incorrectly assigned regions to be rejected from the speech source, thereby producing a better recognition hypothesis. This effect is illustrated in Fig. 10, where the spectro-temporal signal representation has been altered to simulate broad-band noise bursts. These unexpected components appear as bands in the present data mask and hence disrupt the standard missing data

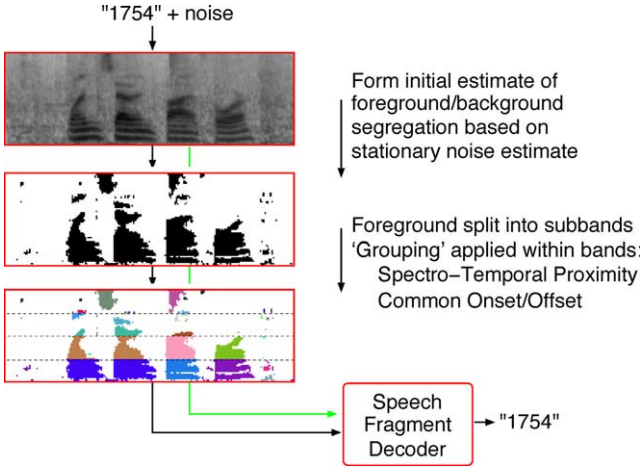

Fig. 9. A summary of the front-end processing used to generate the fragments employed in the experiments described in Section 3.

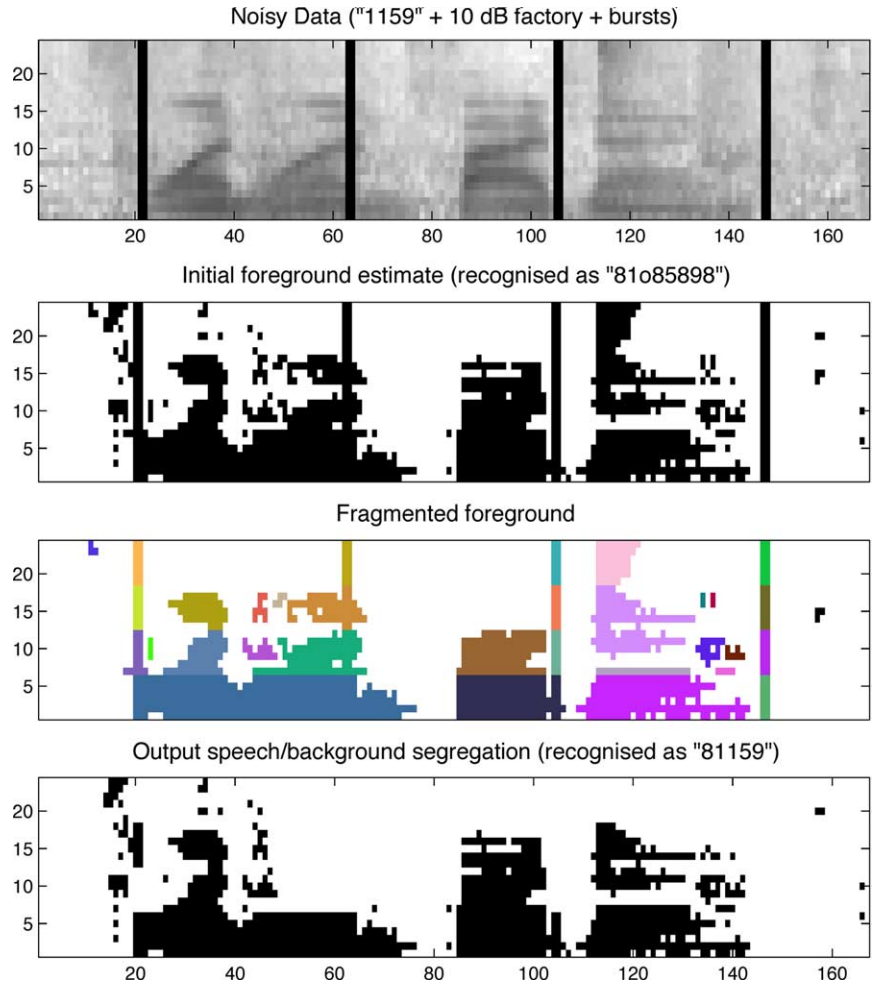

Fig. 10. An example of the speech fragment decoder system performance when applied to data corrupted by artificial transients (see text).

recognition technique ('1159' is recognised as '81o85898'). The third image in the figure shows how the mask is now dissected before being passed into the speech fragment decoder. The final panel shows a *backtrace* of the fragments that the speech fragment decoder marks as present in the winning hypothesis. We see that the noise pulse fragments have been dropped (i.e. relabelled as 'background'). Recognition performance is now much improved ('1159' is recognised as '81159').

Fig. 11 shows a further example with a different pattern of artificial noise—a series of chirps—imposed upon the same utterance. Again, noise contaminated fragments are mostly placed into the background by the decoder.

3.3. Results with real noise

The examples discussed in the previous section were artificial and the background intrusions in the data mask were very distinct. The experiments in this section test the technique with speech mixed with factory noise taken from the NOISEX corpus (Varga et al., 1992).

Fig. 12 compares the performance of the speech fragment decoding technique with that of a recog-

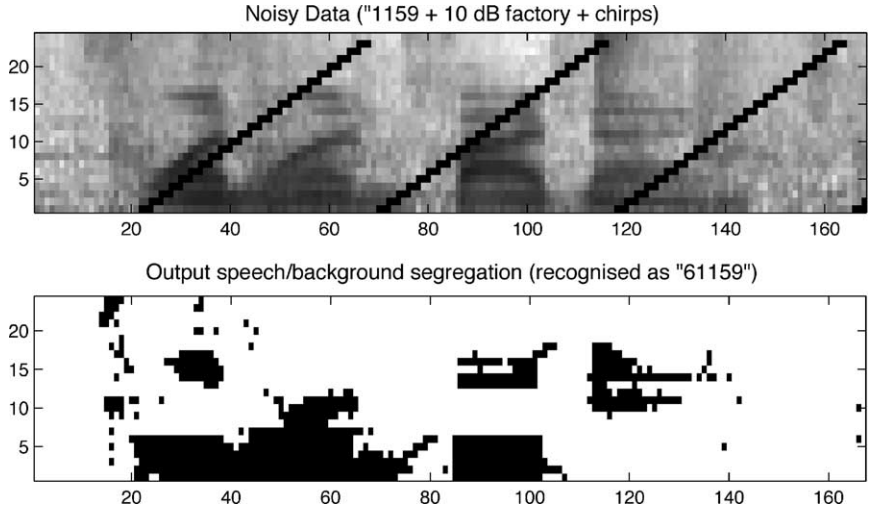

Fig. 11. Another example of the speech fragment decoding for data corrupted with artificial chirps.

niser using the stationary SNR-based speech/background segregation in conjunction with missing data techniques.

It can be seen that speech fragment decoding provides a significant improvement at the lower SNRs, e.g. at 5 dB recognition accuracy is improved from 70.1% to 78.1%—a word-error rate reduction from 29.9% to 21.9%, or 26.7% relative.

Also shown on the graph are results using a traditional MFCC system with 13 cepstral coefficients, deltas and accelerations, and cepstral mean normalisation (labelled MFCC + CMN). This demonstrates that the speech fragment decoding technique is providing an improvement over a missing data system that is already robust by the standards of traditional techniques.

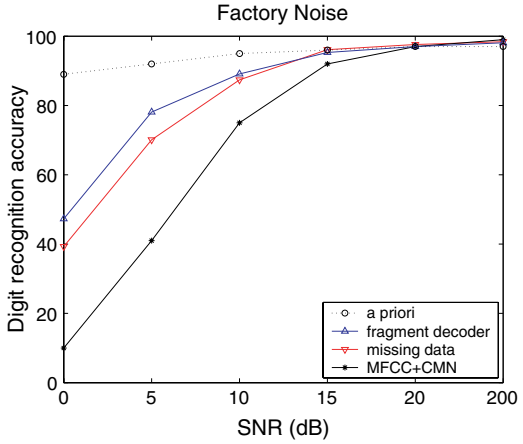

Fig. 12. Recognition results for a baseline MFCC system, a missing data system, and the speech fragment decoder system. The 'a priori' line represents results that are potentially achievable if the speech can be perfectly segregated from the noise.

3.4. Discussion

The results in Fig. 12 labelled 'a priori' show the performance achieved using missing data techniques if prior knowledge of the noise is used to create a perfect local SNR mask. Even using the speech fragment decoding technique results fall far short of this upper limit as the noise level rises above 10 dB SNR.

One possible cause of this this significant performance gap is that the fragments supplied to the speech fragment decoder are not sufficiently coherent. In this work we have used a simple set of fragments generated by aggregating high energy regions in the SNR mask. If the noise and speech sources occupy adjoining spectro-temporal regions this technique will not be able to separate them. This is evident is Figs. 10 and 11 where, as a result of both noise and speech being mixed in the same fragment, much clean speech energy has been

removed from the masks and some of the noise energy has survived.

The artificial examples highlight that the success of the system is strongly dependent on the quality of the segregation model. By producing incoherent fragments, the segregation model limits the performance of the recogniser as it has effectively made hard decisions that cannot be undone at a later stage. Of course, the coherence of the fragments can be easily increased by splitting them into smaller and smaller pieces. At the extreme each fragment may contain a single spectro-temporal pixel which by definition must be coherent. However, over zealous fragmentation also has undesirable consequences. First, it greatly increases the size of the segregation search space and hence increases the computational cost of the decoding process. Second, it weakens the constraints imposed by the segregation model. If there are a very large number of small fragments, the decoder is more able to construct spurious speech descriptions by piecing together spectro-temporal pieces from the collection of sound sources present.

4. Discussion

In this paper we have laid the foundation for a statistical approach to computational auditory scene analysis. In the sections that follow, we discuss some of the issues that have arisen with our current implementation and suggest some possible future research directions.

4.1. Improvements to fragment generation

The fragments in the current system rely on a very simple and crude model—mainly that energy below an estimate 'noise floor' is to be ignored, and the remainder can be divided up according to some simple heuristics. It is likely that more powerful fragmentation techniques will result in significant performance gains. In general, one can imagine a two-phase process in which cues for auditory grouping (as listed, for example, in Bregman, 1990, and Table 1 of Cooke and Ellis, 2001) are applied to aggregate auditory filter outputs across time and frequency, followed by the application of segregation principles which serve to split the newly-formed regions. In contrast with earlier approaches to grouping and segregation, such a strategy can afford to be conservative in its application of grouping principles, since some of the work of aggregation can be left to the decoder. In fact, since any groups formed at this stage cannot later be split, it is essential that any hard-and-fast decisions are based on reliable cues for grouping. In practice, this can be achieved both by adopting more stringent criteria for incorporation of time-frequency regions into groups and by weakening criteria for the splitting of groups.

For instance, within the regions currently marked as 'voiced', subband periodicity measures could indicate whether frequency channels appear to be excited by a single voice, or whether multiple pitches suggest the division of the spectrum into multiple voices (as in Brown and Cooke, 1994). Sudden increases in energy within a single fragment should also precipitate a division, on the basis that this is strong evidence of a new sound source appearing.

The application of stricter grouping criteria may appear to result in a loss of valuable information about which regions are likely to belong together. However, we show in the following section that such information can be employed during the decoding stage.

4.2. Statistical versus ruled-based segregation models

The speech fragment decoding theory is expressed in terms of a statistical segregation model. However, the primitive grouping principles described in the previous section have tended to be modelled by essentially rule-based systems and have previously lacked a clear statistical footing. Psychoacousticians have set out to look for grouping rules using reductionist approaches—essentially by studying the percepts generated by highly simplified acoustic stimuli. Rule-based models that rely on a small sets of parameters can be hand tuned to fit such empirical psychoacoustic data.

An alternative approach, is to build a statistical segregation model from labelled noisy data. Labels can be attached to the noisy acoustic data if the contributions of the individual sources are known prior to mixing. A corpus of synthetic sound scenes could be used to achieve this aim.

4.3. Between-fragment grouping

Psychoacoustic experiments provide evidence that weak grouping effects may exist between the tightly bound local spectro-temporal fragments. For example, a sequence of tonal elements are more likely to be perceived as emanating from the same sound source if they have similar frequencies (Van Noorden, 1975). These grouping effects may allow a fragment to have an influence on the evolving source interpretation that spans over a considerable temporal window. However, such between-fragment grouping effects have a probabilistic nature and their influence can be overcome by learned patterns, such as musical melody (Hartmann and Johnson, 1991) or speech (Culling and Darwin, 1993).

Between-fragment grouping effects may be best modelled as soft biases rather than hard and fast rules. One approach would be to estimate prior probabilities of the segregation hypotheses according to various distance measures between the fragments composing the sources that the segregation describes. A suitable distance measure may be based on the similarity of a vector of fragment properties such as mean frequency, spectral shape, spatial location, mean energy. The posterior probability of pairs of fragments belonging to the same source given their properties could then be learnt using training data employing a priori fragments similar to those used in Section 2.4. Such probabilities could be added into the segregation model by appropriately adjusting the scores for each evolving segregation hypotheses as each new fragment is considered by the decoding process.

When including long term between-fragment grouping probabilities into the segregation model some care has to be taken with the speech fragment decoding algorithm to ensure that the Markov property is preserved and that the segregation/word-sequence search remains admissible. In the version of the algorithm described in Section 2.2, decisions about the best labelling of a fragment are made at the instant at which the fragment offsets. However, allowing for between-fragment effects, it is not possible to know at this time point how the labelling of the present fragment will influence the labelling of fragments occurring in the future. This problem can be overcome by first limiting the temporal extent of the between-fragment grouping effects to a fixed number of frames, say T,[6] and second, delaying the decision over how to label a given fragment until the decoder has passed the offset of the fragment by T frames.

Note that the delay in fragment labelling decisions necessitated by between-fragment grouping effects will mean that there are on average more active hypotheses at any instant. The growth in the number of hypotheses will in general be an exponential function of the length of the delay which, in turn, has to be the same duration as the extent of the temporal influence between fragments. Consequently, there is a trade-off between the temporal extent of the between-fragment grouping influences and the size of the segregation search space (and hence computational cost of the decoding procedure).

4.4. Approximating $P(\boldsymbol{X})$

In Eq. (12), we factored the ratio of the likelihood of the speech features conditioned on segregation and mask to their prior values by essentially assuming their values were independent at each time step i, i.e. we took:

$$\frac{P(\boldsymbol{X}|S,\boldsymbol{Y})}{P(\boldsymbol{X})} = \prod_i \frac{P(\boldsymbol{x}_i|S,\boldsymbol{Y})}{P(\boldsymbol{x}_i)}. \tag{23}$$

This independence assumption is certainly incorrect, but difficult to avoid in practical systems. We note, however, that depending on how $P(\boldsymbol{x}_i|S,\boldsymbol{Y})/P(\boldsymbol{x}_i)$ is calculated, the ratio may be reasonable even when the numerator and denominator include systematic error factors, as long as those factors are similar.

[6] That is to say that between-fragment grouping probabilities are included for interactions between the fragment that is ending and each fragment that overlaps a window that extends back T frames before the fragment ended.

A second weak point is our model for the prior distribution of individual speech feature elements at a single time frame, $P(x_j)$, as uniform between zero and some global constant $x_{\max}$. It would be relatively simple to improve this, e.g. by using individual single-Gaussian models of the prior distribution of features in each dimension. Since this applies only to the clean speech features $\boldsymbol{X}$ rather than to the unpredictable noisy observations $\boldsymbol{Y}$, we already have the training data we need.

4.5. Three-way labelling of time-frequency cells

Although the primary purpose of the current system is to decide which time-frequency pixels can be used as evidence for the target voice, we note that there is actually a three-way classification occurring, firstly between stationary background and foreground (by the initial noise estimation stage), then of the foreground energy into speech and non-speech fragments (by the decoding process). This special status of the stationary background is not strictly necessary—those regions could be included in the search, and would presumably always be labelled as non-speech—but it may reveal something more profound about sound perception in general. Just as it is convenient and efficient to identify and discard the 'background roar' as the first processing stage in this system, perhaps biological auditory systems perform an analogous process of systematically ignoring energy below a slowly varying threshold.

4.6. Computational complexity

In the Aurora experiments, the number of fragments per utterance often exceeded 100. However, as illustrated in Fig. 8(G), the maximum number of simultaneous fragments was never greater than 10 and the average number of hypotheses per frame computed over the full test set was below 4. Although the decoder is evaluating on average roughly four times as many hypotheses as a standard missing data decoder, much of the probability calculation may be shared between hypotheses and hence the computational load is increased by a much smaller factor.

4.7. Decoding multiple sources

A natural future extension would be to search for fits across multiple simultaneous models, possibly permitting the recognition of both voices in simultaneous speech. This resembles the ideas of HMM decomposition (Varga and Moore, 1990; Gales and Young, 1993). However, because each 'coherent fragment' is assumed to correspond to only a single source, the likelihood evaluation is greatly simplified. The arguments about the relationship between large, coherent fragments and search efficiency remain unchanged.

5. Conclusion

We have presented a statistical foundation to computational auditory scene analysis, and developed from this framework an approach to recognising speech in the presence of other sound sources that combines (i) a bottom up processing stage to produce a set of source fragments, with (ii) a top-down search which, given models of clean speech, uses missing data recognition techniques to find the most likely combination of source speech/background labelling and speech model sequence. Preliminary ASR experiments show that the system can produce recognition performance improvements even with a simplistic implementation of the bottom-up processing. We believe that through the application of more sophisticated CASA-style sound source organisation techniques, we will be able to improve the quality of the fragments fed to the top-down search and further improve the performance of the system.

References

Bailey, P., Dorman, M., Summerfield, A., 1977. Identification of sine-wave analogs of CV syllables in speech and non-speech modes. J. Acoust. Soc. Amer., 61.

Barker, J., Cooke, M., 1999. Is the sine-wave speech cocktail party worth attending? Speech Commun. 27, 159–174.

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

Bourlard, H., Dupont, S., 1997. Subband-based speech recognition. In: Proc. ICASSP'97, Munich, Germany, April 1997. pp. 1251–1254.

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

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

Cooke, M.P., 1991. Modelling auditory processing and organisation. Ph.D. Thesis, Department of Computer Science, University of Sheffield.

Cooke, M., Ellis, D., 2001. The auditory organisation of speech and other sound sources in listeners and computational models. Speech Commun. 35, 141–177.

Cooke, M., Green, P., Crawford, M., 1994. Handling missing data in speech recognition. In: Proc. ICSLP'94, Yokohama, Japan, September 1994. pp. 1555–1558.

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

Culling, J., Darwin, C., 1993. Perceptual separation of simultaneous vowels: within and across-formant grouping by F0. J. Acoust. Soc. Amer. 93 (6), 3454–3467.

Cunningham, S., Cooke, M., 1999. The role of evidence and counter-evidence in speech perception. In: ICPhS'99. pp. 215–218.

Denbigh, P., Zhao, J., 1992. Pitch extraction and separation of overlapping speech. Speech Commun. 11, 119–125.

Ellis, D.P.W., 1996. Prediction-driven computational auditory scene analysis. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT.

Gales, M.J.F., Young, S.J., 1993. HMM recognition in noise using parallel model combination. In: Eurospeech'93, Vol. 2. pp. 837–840.

Hartmann, W., Johnson, D., 1991. Stream segregation and peripheral channeling music perception. Music Percept. 9 (2), 155–184.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: Algorithms and applications, neural networks. Neural Networks 13 (4–5), 411–430.

Leonard, R., 1984. A database for speaker-independent digit recognition. In: Proc. ICASSP'84. pp. 111–114.

Parsons, T., 1976. Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Amer. 60 (4), 911–918.

Pearce, D., Hirsch, H.-G., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ICSLP'00, Vol. 4, Beijing, China, October 2000, pp. 29–32.

Remez, R., Rubin, P., Pisoni, D., Carrell, T., 1981. Speech perception without traditional speech cues. Science 212, 947–950.

Scheffers, M., 1983. Sifting vowels: auditory pitch analysis and sound segregation. Ph.D. Thesis, University of Groningen, The Netherlands.

Slaney, M., 1995. A critique of pure audition. In: Proc. 1st Workshop on Computational Auditory Scene Analysis, Internat. Joint Conf. Artificial Intelligence, Montreal. pp. 13–18.

Van Noorden, L., 1975. Temporal coherence in the perception of tone sequences. Ph.D. Thesis, Eindhoven University of Technology.

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

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

Wang, D., Brown, G., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks 10 (3), 684–697.