DECODING SPEECH IN THE PRESENCE OF OTHER SOURCES

Jon Barker, Martin Cooke

Department of Computer Science,
University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP, UK

j.barker, m.cooke@dcs.shef.ac.uk

Daniel P.W. Ellis

Department of Electrical Engineering,
Columbia University
500 W. 120th Steet, New York NY 10027, USA

dpwe@ee.columbia.edu

June 4, 2002

ABSTRACT

Acoustic interference is arguably the most serious problem facing current speech recognizers. The maturation of statistical pattern recognition techniques has brought us very low word error rates when the training and test material both consist solely of speech. However, in real-world situations, any speech signal of interest will be mixed with background noises coming from the full range of sources encountered in our acoustic environment.

In this paper we present a new technique that extends conventional speech recognition to operate in the situation of loud and variable interfering sounds. Our goal is to develop a technique that decomposes the signal according to the different original sources, and applies pattern matching only to the parts of the signal that truly belong to the target voice. In this way, a single set of clean-speech models can be employed consistently across any range of background interference. Based on insights into human sound organization, we combine low-level signal cues indicating that local 'fragments' of sound energy belong together, with the high-level structural constraints on the allowable acoustic sequences implicit in the trained speech models. This combination is effected through a modified hidden Markov model decoder that 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 the Aurora-2 noise-corrupted digits task, the new approach improves the word error rate in the condition of speech babbel at 5 dB SNR from over 90% for a standard ASR system to around 26%. This is a significant improvement even over the closely-related missing-data approach (which scores around 32% in the same condition) because, unlike missing-data, the speech fragment decoder can search across different segregation hypotheses to find the set of target/background labels most consistent with the speech model constraints.

List of Figures

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 (refs to various robustness workshops). 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 (aurora). Modest departures from such conditions leads 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 (i.e. stored) 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) 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's '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 behavior, 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 2 speech formants are replaced by sinusoids, listeners recognise the resulting sine-wave speech at levels approaching 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, 1997).

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 model 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 of source-driven processing and decoding speed.

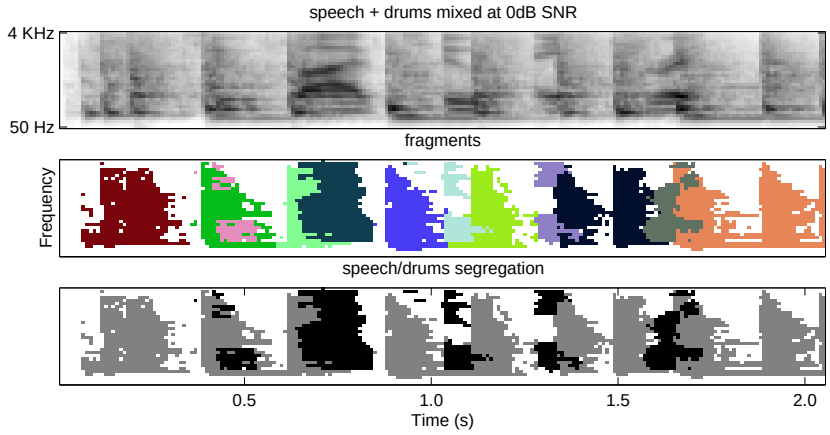

Figure 1: The top panel shows the auditory spectrogram of the utterance "two five two eight three" spoken by a male speaker mixed with drum beats at 0 dB SNR. The lower panel shows the correct segregation of speech energy (black) and drums energy (grey). The centre panel illustrates the set of fragments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

Figure 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 what follows, we refer to these regions as fragments, and use the phrase "segregation model" to indicate the fragmentation process. The aim of the new approach is to search over all admissable fragment combinations to generate the most likely word sequence (or, more generally, model sequence). We show that this can be achieved by decomposing the output probability calculation into 3 parts: a segregation model, a language model, and an acoustic model modified to calculate partial likelihoods. Source-driven processes make up the segregation model, and the resulting fragments occupy arbitrary regions of the time-frequency plane. For instance, fragments need not be compact, not restricted to individual time frames or frequency bands. Fragments may occupy regions of arbitrary size, but will typically be smaller than a syllable and bounded in frequency extent.

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 action of the resulting decoder on both artificial and real noises. Section 4 introduces some extensions to the decoder, and provides results on AURORA 2 task (Pearce and Hirsch, 2000).

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 $\hat{W} = w_1 w_2 ... w_N$ with the maximum *a posteriori* probability given the sequence of acoustic feature vectors observed for the speech, $\mathbf{X} = \mathbf{x}_1 \mathbf{x}_2 ... \mathbf{x}_T$:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|\mathbf{X}) \tag{1}$$

This equation is rearranged using Bayes' rule into:

$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{P(\mathbf{X}|W)P(W)}{P(X)} \tag{2}$$

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, $P(\mathbf{X}|W)$ (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 $P(\mathbf{X}|W)$ is learned by modeling 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, $\hat{W}$, along with the speech/background segregation, $\hat{S}$, which jointly have the maximum posterior probability. [1] 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 $\mathbf{Y}$ to differentiate them from the $\mathbf{X}$ used in our speech-trained acoustic models $P(\mathbf{X}|W)$.

$$\hat{W}, \hat{S} = \underset{W,S}{\operatorname{argmax}} P(W, S|\mathbf{Y}) \tag{3}$$

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

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

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

Since W is independent of S and $\mathbf{Y}$ given $\mathbf{X}$, the first probability simplifies to $P(W|\mathbf{X})$. 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

[1] 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 Equation 3 over the segregation space defining $W' = \operatorname{argmax}_W \sum_S P(W, S|\mathbf{Y})$. However, this integration presents some computational complexity so in practise even if we were not directly interested in the segregation it may be desirable to implement Equation 3 directly and take $\hat{W}$ as an approximation of W'.

$P(\mathbf{X}|W)$:

$$P(W,S|\mathbf{Y}) = \int \frac{P(\mathbf{X}|W)P(W)}{P(\mathbf{X})} P(\mathbf{X}|S,\mathbf{Y})d\mathbf{X} \cdot P(S|\mathbf{Y}) \quad (6)$$

$$= P(W)\left(\int P(\mathbf{X}|W)\frac{P(\mathbf{X}|S,\mathbf{Y})}{P(\mathbf{X})}d\mathbf{X}\right)P(S|\mathbf{Y}) \quad (7)$$

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

In the case of recognition with hidden Markov models (HMMs), the conventional derivation introduces an unobserved state sequence $Q = q_1, q_2, ..., q_T$ 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 $\mathbf{x_i}$ at time i depend only on the corresponding state q_i, making $P(\mathbf{X}|Q) = \prod_i P(\mathbf{x_i}|q_i)$. The total likelihood of a particular W over all possible state sequences is normally approximated by the score over the single most-likely state sequence (the *Viterbi* path). In our case, this gives:

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

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

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

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

It can be seen that there are three significant differences:

i A new term, $P(S|\mathbf{Y})$ has been introduced. This is the 'segregation model', describing the probability of a particular segragation S given our actual observations $\mathbf{Y}$, 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 organization.

ii The acoustic model score $P(\mathbf{X}|Q)$ is now evaluated over a range of possible values for $\mathbf{X}$, 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 below.

iii 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 admissable 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'. Equation 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 section which follow.

$$\hat{W}, \hat{S} = \underbrace{\underset{W,S}{\operatorname{argmax}} \max_{Q \in Q_W}}_{\text{Search algorithm}} \underbrace{P(S|\mathbf{Y})}_{\text{Segregation model}} \underbrace{P(W)P(Q|W)}_{\text{Language model}} \int \underbrace{P(\mathbf{X}|Q)}_{\text{Acoustic model}} \underbrace{\frac{P(\mathbf{X}|S,\mathbf{Y})}{P(\mathbf{X})}}_{\text{Segregation weighting}} d\mathbf{X}$$

Search algorithm
e.g. modified decoder

Language model
bigrams, dictionary

Segregation weighting
connection to observations

Segregation model
source-level grouping processes

Acoustic model
schema-driven processes

2.1. The Segregation Model

Consider the space of potential speech/background segregations. An acoustic observation vector, $\mathbf{X}$ may be constructed as a sequence of frames $\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$ 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} possible segregations for a one second audio sample.

Fortunately, most of these segregations can be immediately ruled out 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, which we shall hence force call 'coherent fragments', constrain the spectro-temporal pixels contained to share the same speech/background label. So 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). So 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. Figure 2 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).

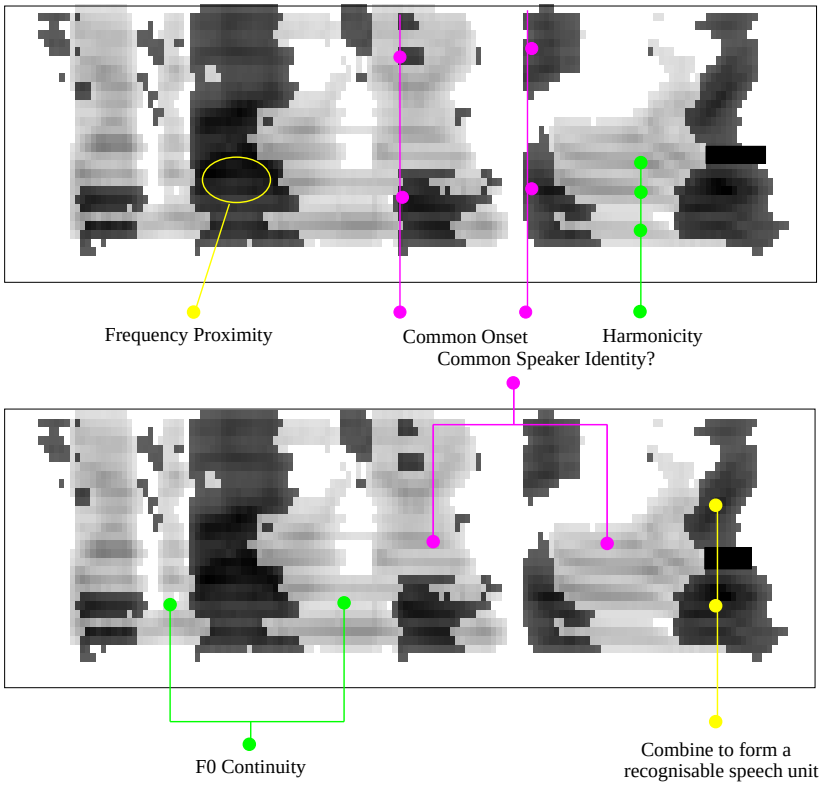

Figure 2: An illustration of short-term (above) and long-term (below) primitive grouping cues which may be exploited to recover partial descriptions of individual sound sources

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 prior probabilities would vary across segregations and would take into account such factors as 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. For further discussion of such 'between-fragment grouping' effects and of the modifications to the search algorithm that they require see Section 5.2.

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, $P(\mathbf{X}|Q)P(\mathbf{X}|S,\mathbf{Y})/P(\mathbf{X})$, is conditioned on both 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 multiplies 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 the naive implementation of the search illustrated in Figure 3. 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 exponential 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 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.

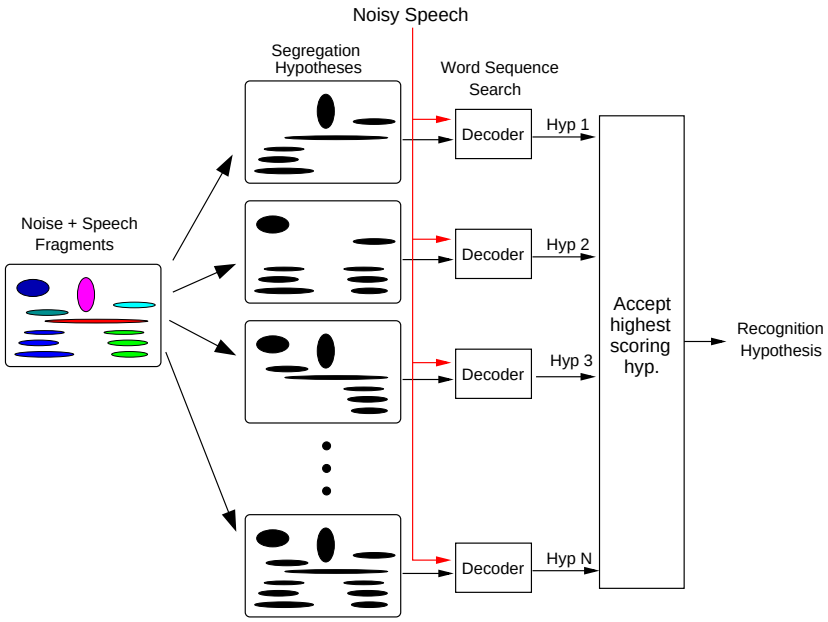

Figure 3: A naive implementation of the segregation/word-sequence search

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 Figure 4, 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 Figure 4. This sharing of computation between pairs of segregation hypotheses, can be

generalised to encompass all segregation hypotheses by arranging them in a tree like structure. As we progress through time, every time a new fragment onsets all the current segregation hypotheses branch to form two complementary cases - in one case the onsetting fragment is considered to be speech and 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 Figure 4). At this point the two competing segregation hypotheses can be compared and the least likely of the pair can be rejected without effecting 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.

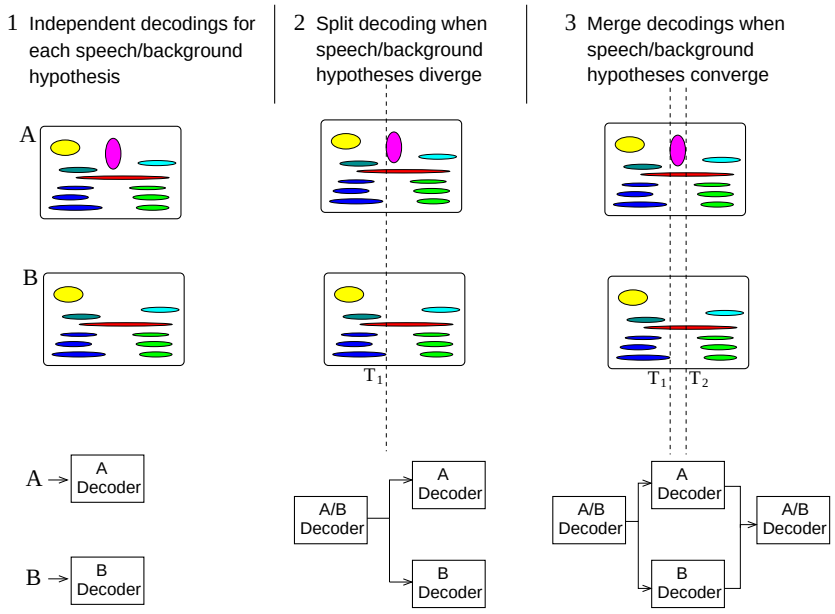

Figure 4: The efficient segregation search exploits the fact that competing segregation hypotheses only differ over a limited number of frames

Figure 5 illustrates the evolution of a set of parallel segregation hypotheses while processing a segment of noisy speech which has been dissected into 3 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 and 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 4 active segregation hypotheses, not the 8 required to consider every possible labelling of each of the 3 fragments.

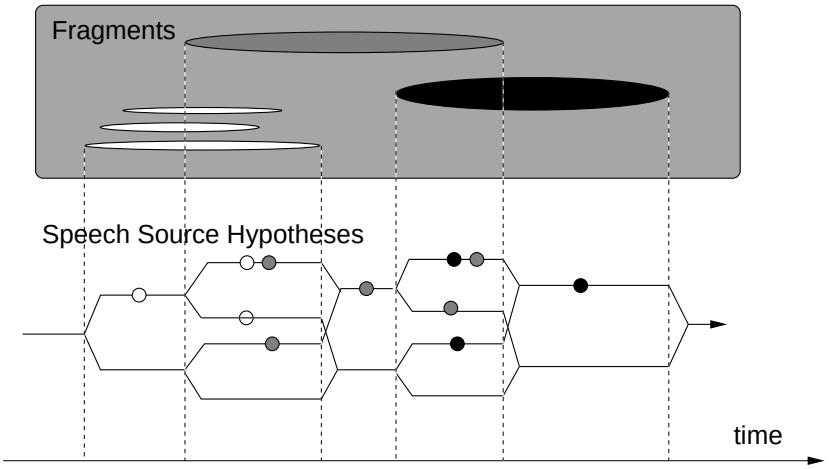

Figure 5: The evolution of a set of segregation hypotheses. Each parallel path represents a separate hypothesis, with the shaded dots indicating which ongoing fragments are being considered as speech part of the speech source.

It is important to understand that the evolution of the segregation hypotheses is dependent on the word sequence hypothesis. For each ongooing 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 effect the admissibility of the search (i.e. it is not a form of hypothesis pruning). The efficient algorithm will return the exact same result a brute-force approach which separately considered 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. It should be noted that 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 (for details see Section 5.2).

2.3. The Acoustic Model

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

$$\int P(\mathbf{X}|Q)\frac{P(\mathbf{X}|S,\mathbf{Y})}{P(\mathbf{X})}d\mathbf{X} \tag{10}$$

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

The integral over the entire space of $\mathbf{X}$ — the full multi-dimensional 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: [2]

$$\int P(\mathbf{X}|Q)\frac{P(\mathbf{X}|S,\mathbf{Y})}{P(\mathbf{X})}d\mathbf{X} \quad = \quad \prod_i \int P(\mathbf{x_i}|q_i)\frac{P(\mathbf{x_i}|S,\mathbf{Y})}{P(\mathbf{x_i})}d\mathbf{x_i} \tag{11}$$

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

$$P(\mathbf{x}|q) = \sum_{k=1}^{M} P(k|q)P(\mathbf{x}|k,q) \tag{12}$$

[2] This step also assumes independence of each timestep for the prior $P(\mathbf{X})$ and for the likelihood of X given the segregation hypothesis and observations, $P(\mathbf{X}|S,\mathbf{Y})$. Both these assumptions are open to serious question, and we return to them in Section 5.

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

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

Assuming a similar decomposition of the prior $P(X)$, we can take the integral of Equation 11 inside the summation to give:

$$\int P(\mathbf{x}|q) \frac{P(\mathbf{x}|S,\mathbf{Y})}{P(\mathbf{x})} d\mathbf{x} = \sum_{k=1}^{M} P(k|q) \prod_j \int P(x_j|k,q) \frac{P(x_j|S,\mathbf{Y})}{P(x_j)} dx_j \tag{14}$$

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

We can consider the factor

$$\frac{P(x_j|S,\mathbf{Y})}{P(x_j)} \tag{15}$$

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 marginalization* (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 [3] from the observed signal $\mathbf{Y}$ and hence the segregation weighting will be a Dirac delta at the calculated value, x^*:

$$P(x_j|S,\mathbf{Y}) \quad = \quad \delta(x_j - x^*) \tag{16}$$

$$\int P(x_j|k,q) \frac{P(x_j|S,\mathbf{Y})}{P(x_j)} dx_j \quad = \quad P(x^*|k,q)/P(x^*) \tag{17}$$

[3] The observed signal $\mathbf{Y}$ will in general be a richer representation than simply the subband energies that would have formed $\mathbf{x}$ 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 $\mathbf{x}$ will be completely defined given $\mathbf{Y}$ in the case of a segregation hypothesis that rates the whole spectrum as unmasked for that time slice.

The more interesting case comes when 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 $P(x|S,\mathbf{Y})$ as proportional to the prior $P(x)$ for $x \leq x^*$, and zero for $x > x^*$. Thus,

$$P(x_j|S,\mathbf{Y}) = \begin{cases} F \cdot P(x_j) & x_j \leq x^* \\ 0 & x_j > x^* \end{cases} \tag{18}$$

$$\int P(x_j|k,q)\frac{P(x_j|S,\mathbf{Y})}{P(x_j)}dx_j \quad = \quad \int_{-\infty}^{x^*} P(x_j|k,q) \cdot F dx_j \tag{19}$$

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

$$F = \frac{1}{\int_{-\infty}^{x^*} P(x_j)dx_j} \tag{20}$$

In Equation 19, 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 equation 19 and the evaluation of the point-likelihood in Equation 17 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 Equation 17 for the unmasked dimensions and Equation 19 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 Equation 14 which becomes:

$$\int P(\mathbf{x}|q)\frac{P(\mathbf{x}|S,\mathbf{Y})}{P(\mathbf{x})}d\mathbf{x} = \sum_{k=1}^{M} P(k|q) \prod_{j\in S_O} \frac{P(x_j^*|k,q)}{x_{max}} \prod_{j\in S_M} \int P(x_j|k,q)\cdot\frac{x_{max}}{x_j^*}dx_j \quad (21)$$

where S_O is the set of directly observed (not masked) dimensions of $\mathbf{x}$, 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 Equation 11 to give the data likelihood for an entire sequence.

Consider the effect of x_{max} in Equation 21. Holding everything else constant, the likelihood is proportional to $(x_{max})^{\#M-\#O}$ where $\#M$ is the number of channels treated as masked, and $\#O$ is the count of channels believed to be directly observable. Since the decoder is searching through segregation space S the hypotheses that it compares do not necessarily share the same mask. Some hypotheses have more missing data than others, and hence these hypotheses will be effected to a greater degree by any bias to the likelihood term introduced by the missing feature likelihood computation. As a result it may be observed that using equation 21 leads to a bias toward hypotheses in which too many fragments have been labelled as background, or alternatively toward hypotheses in which too many fragments have been labelled as speech. As an approximate solution to this problem, the results of the integrations across the masked dimensions (only) are scaled by a tuning parameter α shifting the relative likelihood of 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.

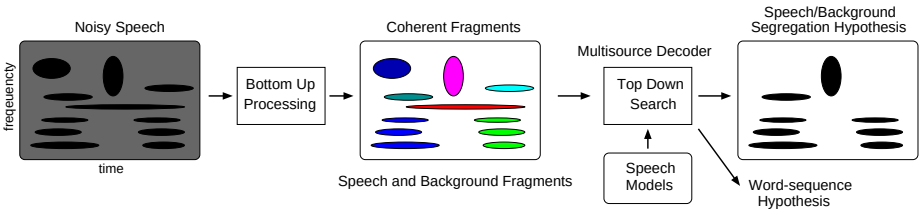

Figure 6: An overview of the speech fragment decoding system. Bottom-up processes are employed to locate 'coherent fragments' (regions of representation that are due entirely to one source) and then a top-down search with access to speech models is used to search for the most likely combination of fragment labelling and speech model sequence.

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, as well as the corresponding word sequence.

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

Generating the 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'.

Constructing the fragments

The fragments were generated by the following steps:

i For each noisy utterance the first 10 frames of the auditory spectrogram are averaged to estimate a stationary noise spectrum.[4]

ii The noise spectrum estimate is used to estimate the local SNR for each frame and frequency channel of the noisy utterance.

[4]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

iii 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.

iv The initial approximation of the speech segment is dissected by first dividing it into four frequency bands.

v Each contiguous regions within each of the four subbands is defined to be a separate fragment.

vi The set of fragments and the noisy speech representation are passed to the speech fragment decoder.

The fragmentation process is summarised in Figure 7.

Training the 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.

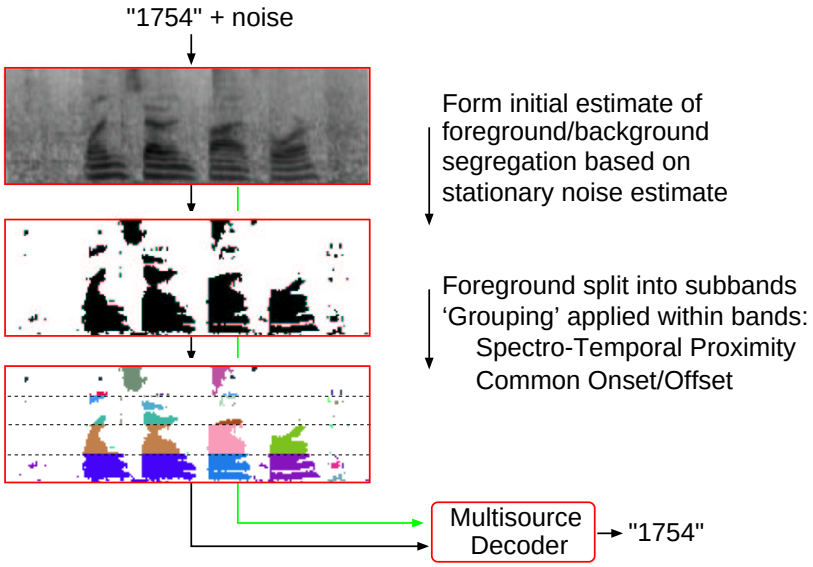

Figure 7: The speech fragment decoder with SNR-based fragments.

The scaling constant, α, that is 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 frag-

ment decoder should allow incorrectly assigned regions to be rejected from the speech source, thereby producing a better recognition hypothesis. This effect is illustrated in figure 12, where broad-band noise bursts have been artificially added to the noisy data representation. These unexpected components appear as bands in the present data mask and hence disrupt the standard missing data 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 "61159").

Figure 8 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.

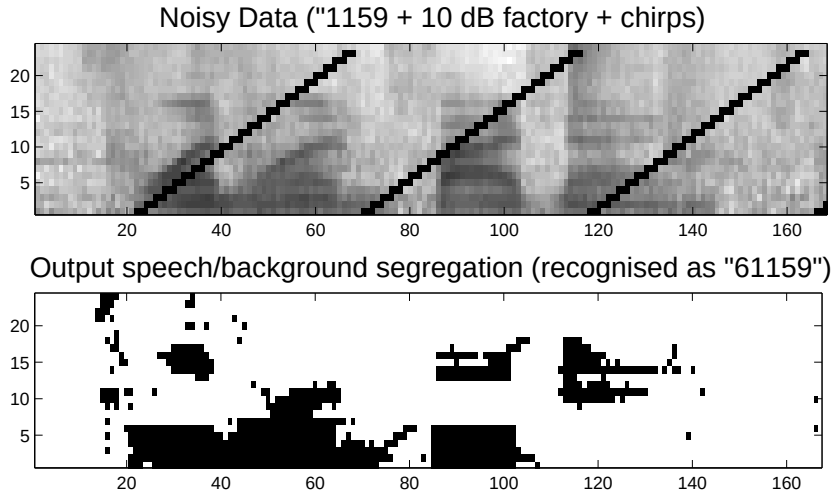

Figure 8: Another example of the speech fragment decoding for data corrupted with artificial chirps.

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). NOISEX factory noise provides a challenge for robust ASR systems as although it has a stationary background, it also has high intensity non-stationary components such as hammer blows etc. which are highly unpredictable.

Figure 9 compares the performance of the speech fragment decoding technique, over that of a recogniser 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.

3.4. Discussion

The results in figure 9 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 clumping high energy regions in the SNR mask. If the noise and speech sources occupy

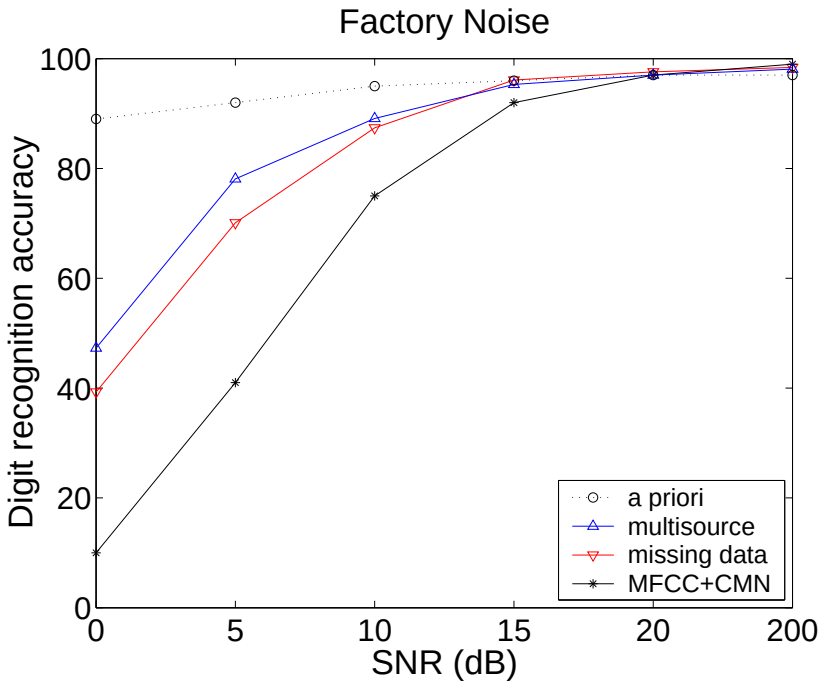

Figure 9: Recognition results for a baseline MFCC system, a missing data system, and the speech fragment decoder system. The "a priori" line represents results that are potentially achievable if the speech can be perfectly segregated from the noise.

adjoining spectro temporal regions this technique will not be able to separate them. This is evident is figures 12 and 8 where, as a result of both noise and speech being mixed in the same fragment, a lot of clean speech energy has been removed from the masks and some of the noise energy has survived.

The artificial examples highlight that the point that the success of the system is strongly dependent on the quality of the segregation model. By producing incoherent fragment the segregation model limits the performance of the recogniser as it has effectively made hard decisions that can not 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 increase 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. SPEECH FRAGMENT DECODING WITH SOFT DECISIONS

In the previous experiment the speech fragment decoder assumed a discrete speech/background classification for each time-frequency 'pixel'. That is to say that each time frequency element belongs exclusively to one fragment, and each fragment is hypothesised to be either part of the speech source or part of another competing sound source. However recent research has shown that, when using missing-data techniques to handle masked speech, better ASR results are obtained by softening the speech/background decisions and assigning pixels with a *probability* of being speech rather than a binary speech/background label (Barker *et al.*, 2000).

This section employs an extension to the theory described thus far designed to carry the advantages of soft speech/background decisions through to the speech fragment decoder architecture. The new system is similar to that of previous experiment in that it employs estimated local SNR to produce the initial set of fragments and the decoder hypothesises hard speech/background assignments at the level of fragments. However, following the soft missing data approach reported in (Barker *et al.*, 2000), the acoustic model is adapted to employ estimated $P(SNR > 0)$ probabilities at the 'pixel' level *within* each fragment when calculating the likelihood of matches to the clean speech models given the hypothesised speech/background segregation.

4.1. Procedure

Generating the feature vectors

The experiments employed speech data from the Aurora 2.0 speaker independent connected digit recognition task (Pearce and Hirsch, 2000). The feature vectors were generated in a similar manner to those described in Section 3.1 except this time 32 channels were employed rather than 24, and the centre frequencies were evenly spaced (on an ERB scale) between 50 Hz and 3750 Hz rather than between 50 Hz and 8 kHz.[5]

Constructing the fragments

The fragmentation of the auditory spectrogram follows a similar procedure to that employed in the previous experiment. A local SNR estimate is employed to make an initial speech/background segregation. The speech segment, which will generally contain some erroneous background regions due to error in the SNR estimation, is then split into a number of fragments.

There are three details in which the fragmentation procedure differs from that employed previously.

First, the local SNR estimate is based on an adaptive noise estimation technique rather than a stationary noise estimate. The adaptive noise estimate is initialised using the stationary noise estimate computed by averaging the first ten frames of the representation (as employed previously). Then starting at the $11th$ frame the noise estimate is adapted using spectro-temporal points at which the estimated local SNR is below some minimum threshold i.e. glimpses of the noise that are observed in low energy speech regions. Both the noise mean and variance are computed and from this $P(SNR > 0)$ is computed at each point (i.e. the probability that the energy is predominantly speech energy). Then the initial speech/background segregation is made by taking the speech segment to be those point where $P(SNR > 0) > 0.5$.

[5] Note, the Aurora data has a 8 kHz sampling rate, whereas the TIDigit data employed in the previous experiment is sampled at 20 kHz.

Second, the initial speech segment is cut across frequency at certain points along the temporal axis. These cuts are made at points where the signals changes from being predominantly harmonic to predominantly inharmonic or vice versa. This extra step helps to separate harmonic from inharmonic sound sources that happen to have energy in overlapping spectro-temporal regions and would thus otherwise merge to form an incoherent fragment. The degree of harmonicity at each frame is measured using a technique based on the autocorrelogram (see (Barker *et al.*, 2001) for details).

Third, fragments are identified without the prior splitting into frequency subbands that was employed in the previous experiment. The auditory spectrogram employed in this experiment has a higher resolution on the frequency axis (32 channel spanning 50 Hz to 4 kHz, as opposed to 24 channels spanning 50 Hz to 8 kHz) and therefore better resolves speech formants and harmonics. The resulting peakier nature of the spectral cross-sections means that the initial estimate of the speech segment tends to be composed of a fairly large number of separated regions. This natural separation of the energetic regions reduces the need to split the frequency axis into arbitrary subbands.

The acoustic model

Each of the eleven words in the Aurora 2.0 vocabulary (digits one to nine, plus the two pronunciations of 0, namely "oh" and "zero") are modelled with a 16-state HMM. An additional 3-state model is used to model the silence before and after each utterance, and the pauses that may occur between digits. Each state of the digit models has only two transitions; a self transition and a transition to the following state. The emission distribution for each state is modelled using a mixture of 7 (XXX-CHECK-XXX) Gaussian distributions each with a diagonal covariance matrix. These models are trained on the Aurora 2.0 clean speech training set using HTK and the Aurora 2.0 training scripts.[6]

During testing the acoustic match probability is modified to take advantage

[6]The Aurora 2.0 training scripts are adapted to employ the auditory spectrogram representation and to extend the number of Gaussian mixtures from three to seven.

of the fact that each spectro-temporal pixel has a 'probability of speech' value rather than a discrete speech/background label. Previously, each element of the feature vector was either labelled as speech (i.e. a reliable observation) or background (i.e. an unreliable speech observation). Given these discrete labels the acoustic match score is given by equation 21 from section 2.3, repeated below:

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

where S_O is the set of channels tagged as directly observable (not masked) in the corresponding segregation hypothesis, S_M is the set of masked (missing) channels, k indexes across the diagonal-Gaussian mixture elements used to model state q, and α is our empirical 'balancing' factor between observed and masked dimensions (see Section 2.3). If the jth element of the observation vector has a probability p_j of being a reliable speech observation — i.e. $j \in S_O$ — and a probability $1 - p_j$ of being speech that has been masked by the background — $j \in S_M$ — then the acoustic match score becomes:

$$\sum_{k=1}^{M} P(k|q) \prod_{j} p_j \frac{P(x_j^*|k,q)}{x_{max}} + (1-p_j) \int P(x_j|k,q)\frac{\alpha x_{max}}{x_j^*}dx_j \quad (23)$$

So to recap, the speech fragment decoder explores hypotheses in which the fragments are themselves assigned discrete speech/background labels, but these fragment level discrete labellings are translated into a set of continuous probabilities at the time-frequency pixel level. In the current system when a fragment label is hypothesised to be speech then the pixels it contains are assigned the speech probabilities, $p_j = P(SNR > 0)$ directly from the adaptive noise estimate, and when a fragment is label is hypothesised to be background then the pixels are assigned speech probabilities of $1-P(SNR > 0)$. These continuous p_j values are then employed in the evaluation of the acoustic model $f(x|q)$ through the use of Equation 23.

For the most part the $P(SNR > 0)$ of data in the fragments are close to 1 meaning that the p_j values will be valued as either 1 or 0. For discrete values of p_j equation 23 becomes 22 and hence the acoustic model becomes equivalent to the discrete speech/background version employed in the previous sections. However, in the areas around the edge of each fragment the background noise may make a significant contribution to the observed energy. In these areas $P(SNR > 0)$ may be closer to 0.5, this means that the values of $P(SNR > 0)$ and $1 - P(SNR > 0)$ become more similar and hence the data at the edge of the fragment effectively makes a reduced contribution to the speech/background fragment labelling decision. So although the fragments are themselves still discrete entities, the change in the acoustic model means they have 'softer' edges and hence the recognition system is more able to tolerate errors in the fragment boundaries that arise due to the approximate nature of the fragmentation process.

4.2. Artificial Examples

Figure 13 (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. Adaptive noise estimation identifies the stationary component, leaving the unmasked speech energy and the nonstationary 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 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 Figure 13).

Given these fragments, the decoder is able to correctly recognise the utter-

ance 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.)

4.3. Experiments Employing the Aurora 2 Connected Digit Task

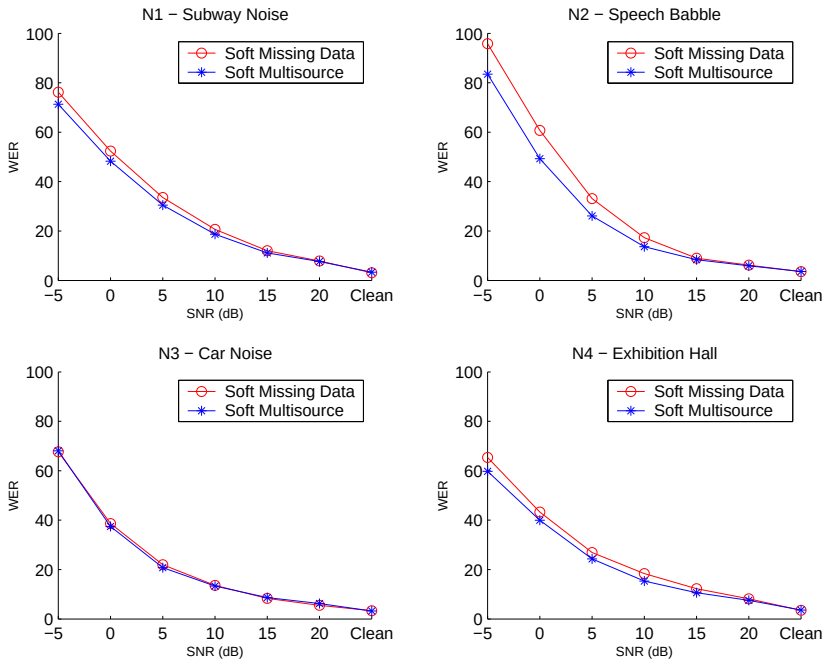

Figure 10: Results for the Aurora Test Set A (see text)

The soft speech fragment decoder system was tested using the Aurora 2.0 speaker independent connected digit recognition task (Pearce and Hirsch, 2000).

Experiments compared the full speech fragment decoder system as described above with a fixed-mask soft-decision missing data system based on the same $P(SNR > 0)$ probabilities calculated from the adaptive noise estimates.

Results for the four noise conditions is the Aurora test set A are shown in

Figure 10.[7] For three of the four noise conditions the speech fragment decoder processing achieves a better performance than the standard missing data system. For the highly non-stationary speech babble noise the performance improvements at low SNRs are fairly large. The only noise for which no improvement is seen is the car noise (N3). Examination of the noises shows that the car noise is the most stationary of the four and is well modelled by the adaptive noise estimate. It is therefore not surprising that for this noise type the speech fragment decoding technique, which is designed to deal with non-stationary noise events, can do little to improve over the already strong performance of the standard missing data technique.

5. DISCUSSION

Having explained the motivation and form of the speech fragment decoder, in this section we discuss some of the issues that have arisen with our current implementation, and the directions we plan to pursue in the future.

5.1. 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 will result in significant improvement gains for the technique. 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

[7] The results presented here are for systems that are not employing temporal difference features and hence the baseline is somewhat lower than similar results published in previous papers e.g. (Barker *et al.*, 2000)

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

5.2. Between-fragment grouping

Psycho-acoustic 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 frequency. 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 intra-fragment grouping effects have a probabilistic nature and their influence can be overcome by learned patterns, such as musical melody or speech.

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 employed in Section 4.2. 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 intra-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 intra-fragment grouping effects to a fixed number of frames, say T frames[8], 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 intra-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. So there is a trade-off between the temporal extent of the intra-fragment grouping influences and the size of the segregation search space (and hence computational cost of the decoding procedure). The exact nature of this trade-off will depend

[8]That is to say that intra-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

on the form of the fragments themselves. As an example see Figure 11, which is generated using ideal coherent fragment data generated from a 0 dB SNR mixture of speech and noise. The graph plots the average number of active segregation hypotheses per frame that would result from a range of values for the temporal extent of intra-fragment grouping effects.

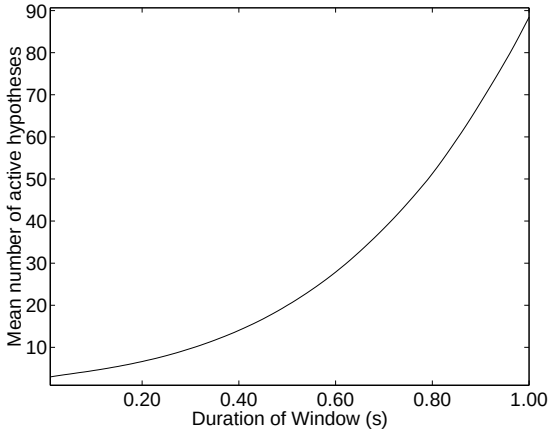

Figure 11: Average simultaneous segregation hypotheses versus temporal extent of intra-fragment grouping effects. The graph has been calculated from fragment data generated from speech plus noise at 0 dB SNR.

5.3. Approximating $P(\mathbf{X})$

In equation 11, 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:

$$\frac{P(\mathbf{X}|S,\mathbf{Y})}{P(\mathbf{X})} \quad = \quad \prod_i \frac{P(\mathbf{x_i}|S,\mathbf{Y})}{P(\mathbf{x_i})} \tag{24}$$

This independence assumption is certainly incorrect, however if we are calculating only the ratio of two unrealistically low probability estimates, it may be that the ratio itself comes out somewhat closer to the the ideal 'true' value. On the other hand, when the segregation hypthesis considers certain dimensions to be directly observed, the math requires scaling by $P(\mathbf{x_i})$ alone. This

is another potential source of imbalance between directly-observed and masked dimensions, which we have patched for the time being with the tuning factor α.

Apart from the independence assumption, our model for the prior distribution of speech feature vector elements within a single time frame, $P(x_j)$, as uniform between zero and some global constant x_{max} is clearly very weak. 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 $\mathbf{X}$ rather than to the unpredictable noisy observations $\mathbf{Y}$, we already have the training data we need.

5.4. 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 nonspeech 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 nonspeech — 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.

5.5. Computational complexity

In the Aurora experiments, the number of fragments per utterance often exceeded 100. However, as illustrated in Figure 13 (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 hypothesis 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.

5.6. 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 again 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.

6. CONCLUSION

We have presented a technique for 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 organization 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.

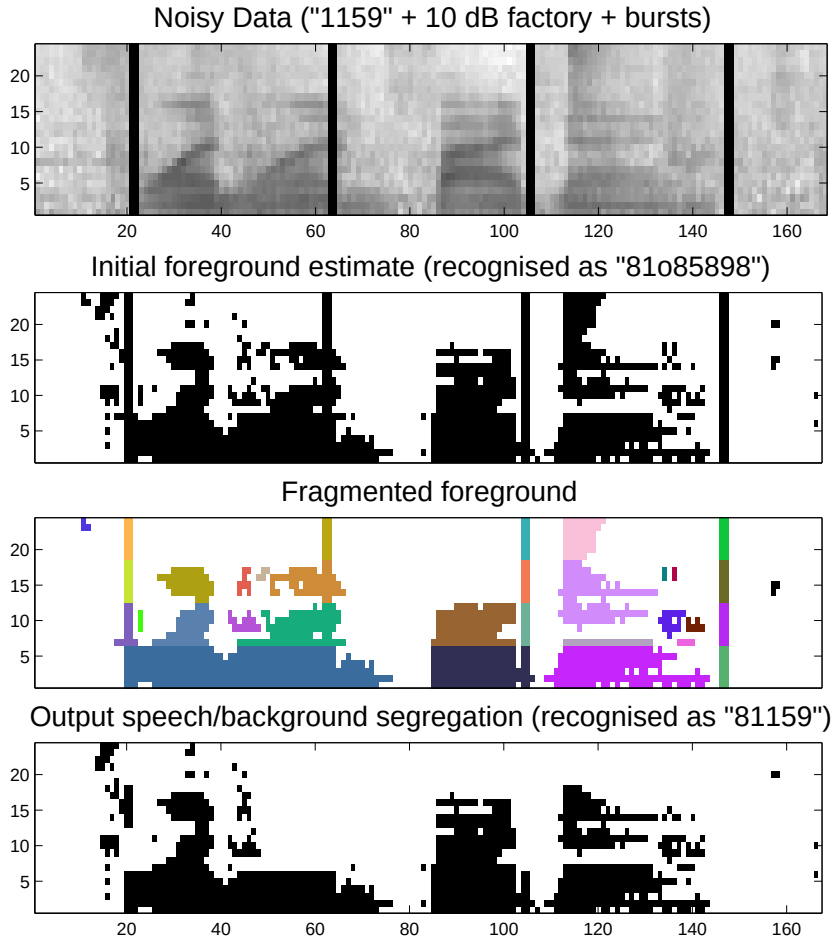

Figure 12: An example of the speech fragment decoder system performance when applied to data corrupted by artificial transients (see text).

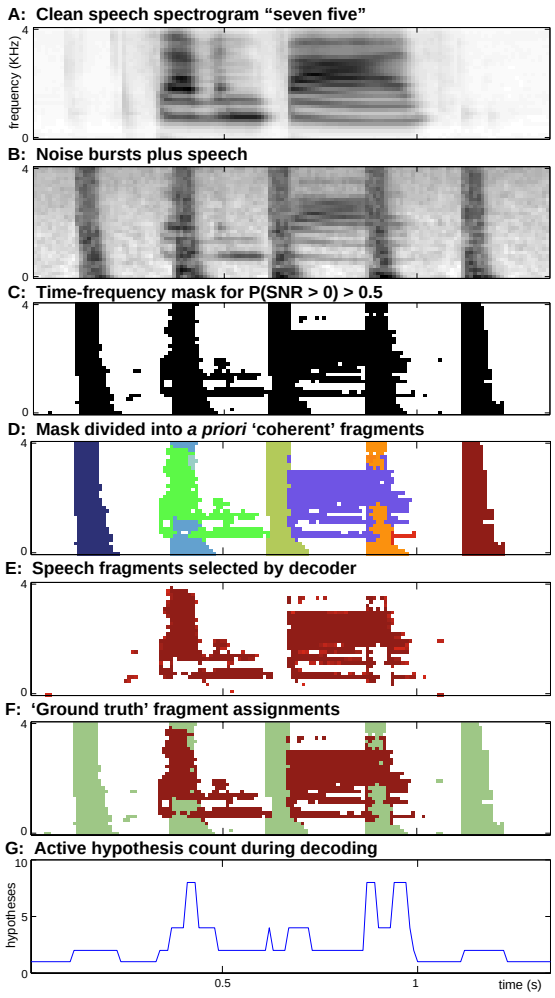

Figure 13: An example of the speech fragment decoder's operation on a single noisy utterance: Panel A shows a spectrogram of the utterance "seven five". Panel B shows the same signal but after adding a two state noise source. Panel C shows the components of the mixture that are not accounted for by the 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.

References

P.J. Bailey, M.F. Dorman, and A.Q. Summerfield. Identification of sine-wave analogs of CV syllables in speech and non-speech modes. *Journal of the Acoustical Society of America*, 61, 1977.

J.P. Barker and M.P. Cooke. Is the sine-wave cocktail party worth attending? In *Proceedings of the 2nd Workshop on Computational Auditory, Scene Analysis*, Nagoya, Japan, 1997. Int. Joint Conf. Artificial Intelligence.

J.P. Barker, L. Josifovski, M.P. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proc. ICSLP '00*, Beijing, China, October 2000.

J.P. Barker, P. Green, and M.P. Cooke. Linking auditory scene analysis and robust ASR by missing data techniques. In *Proc. WISP '01*, Stratford-upon-Avon, UK, 2001.

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

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

A. S. Bregman. *Auditory scene analysis.* MIT Press, 1990.

G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336, 1994.

M.P. Cooke and D.P.W. Ellis. The auditory organisation of speech and other sound sources in listeners and computational models. *Speech Communication*, 2001. Accepted for publication.

M. Cooke, P. Green, and M. Crawford. Handling missing data in speech recognition. In *Proc. ICSLP '94*, Yokohama, Japan, September 1994.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech

recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, 2001.

M. P. Cooke. *Modelling auditory processing and organisation.* PhD thesis, Department of Computer Science, University of Sheffield, 1991.

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

P.N. Denbigh and J. Zhao. Pitch extraction and separation of overlapping speech. *Speech Communication*, 11:119–125, 1992.

D. P. W. Ellis. *Prediction–driven computational auditory scene analysis.* PhD thesis, Department of Electrtical Engineering and Computer Science, M.I.T., 1996.

M. J. F. Gales and S. J. Young. HMM recognition in noise using parallel model combination. In *Eurospeech'93*, volume 2, pages 837–840, 1993.

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

T.W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 60(4):911–918, 1976.

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

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

M.T.M. Scheffers. *Sifting vowels: auditory pitch analysis and sound segregation.* PhD thesis, University of Groningen, The Netherlands, 1983.

M. Slaney. A critique of pure audition. In *Proceedings of the 1st Workshop on*

Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence, Montreal, 1995.

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

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

D.L. Wang and G.J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10(3):684–697, 1999.