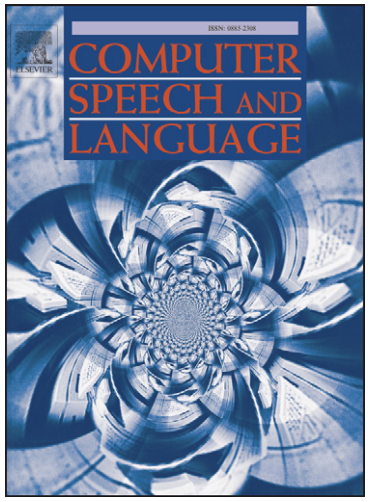
Speech fragment decoding techniques for simultaneous speaker identification and speech recognition

Jon Barker, Ning Ma, André Coy, Martin Cooke



Speech fragment decoding techniques for simultaneous speaker identification and speech recognition

Jon Barker, Ning Ma, André Coy, Martin Cooke

PII: S0885-2308(08)00031-4
DOI: 10.1016/j.csl.2008.05.003
Reference: YCSLA 381

To appear in: *Computer Speech and Language*

Received Date: 18 September 2007
Revised Date: 8 May 2008
Accepted Date: 12 May 2008

Please cite this article as: Barker, J., Ma, N., Coy, A., Cooke, M., Speech fragment decoding techniques for simultaneous speaker identification and speech recognition, *Computer Speech and Language* (2008), doi: 10.1016/j.csl.2008.05.003

Speech fragment decoding techniques for simultaneous speaker identification and speech recognition.

Jon Barker, * Ning Ma, André Coy and Martin Cooke

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

Abstract

This paper addresses the problem of recognising speech in the presence of a competing speaker. We review a speech fragment decoding technique that treats segregation and recognition as coupled problems. Data-driven techniques are used to segment a spectro-temporal representation into a set of fragments, such that each fragment is dominated by one or other of the speech sources. A speech fragment decoder is used which employs missing data techniques and clean speech models to simultaneously search for the set of fragments and the word sequence that best matches the target speaker model. The paper investigates the performance of the system on a recognition task employing artificially mixed target and masker speech utterances. The fragment decoder produces significantly lower error rates than a conventional recogniser, and mimics the pattern of human performance that is produced by the interplay between energetic and informational masking. However, at around 0 dB the performance is generally quite poor. An analysis of the errors shows that a large number of target/masker confusions are being made. The paper presents a novel fragment-based speaker identification approach that allows the target speaker to be reliably identified across a wide range of SNRs. This component is combined with the recognition system to produce significant improvements. When the target and masker utterance have the same gender, the recognition system has a performance at 0 dB equal to that of humans; in other conditions the error rate is roughly twice the human error rate.

Key words: speech recognition, speech separation, speaker identification, simultaneous speech, auditory scene analysis, noise robustness.

* Corresponding author. Tel.: +44 114 222 1824, fax: +44 114 222 1810
Email address: {j.barker, n.ma, a.coy, m.cooke}@dcs.shef.ac.uk (Ning Ma, André Coy and Martin Cooke).

1 Introduction

Despite many years of research focused on robustness, automatic speech recognition (ASR) remains characterised by its fragility. This fragility is perhaps most apparent in the response of ASR to background noise. In the main, ASR systems either employ close-talking microphones, for which a high signal-to-noise ratio can be assured (i.e. the expected operating condition of all commercially available dictation systems), or they are designed to operate with small vocabularies in narrowly specified and highly predictable noise conditions (c.f. Hirsch and Pearce, 2000). For ASR to approach the robustness of human speech recognition (HSR) new approaches are needed that are able to operate without having to make strong assumptions about the acoustic environment. This paper investigates the performance of one technique, Speech Fragment Decoding (SFD), which is designed to operate with minimal assumptions about the nature of the background noise. Here, it is applied to the particularly difficult task of recognising speech in the presence of a second speaker.

Until recently, the challenge of simultaneous speech recognition has not been widely appreciated by the ASR community. The difficulties posed by the unpredictability of the competing speaker, and the similarity between the signal and the 'noise', are sufficient to foil most robust ASR systems. However, there have been many studies of *human* simultaneous speech recognition performance. Early research focused on the identification of pairs of simultaneously presented artificial vowels. These studies have demonstrated a clear effect of pitch difference: vowels are difficult to separately identify if they are presented with the same pitch, but identification scores improve as the pitch separation between the vowels increases (Assmann and Summerfield, 1990). Double-vowel experiments have inspired numerous models of vowel separation (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992; de Cheveigné, 1993) most of which work by first grouping narrowband frequency channels according to the pitch that dominates them, and then applying a template matching algorithm to compare partial spectra with the expected spectra of the possible vowels. To extend these techniques beyond stationary double-vowels it is necessary to deal with the fact that, in real speech, pitch is not stationary (even during a vowel) and that simple template matching algorithms, although adequate for discriminating a small number of artificial stationary vowels, are not suitable for continuous speech recognition even when vocabulary sizes are small. The SFD technique offers solutions to both of these problems (Barker et al., 2005).

The SFD technique is motivated by some very basic observations about the speech signal and the effects of mixing speech with competing sound sources. When speech is mixed with another sound source, certain spectro-temporal regions of the speech signal will be masked by energy from the competing source.

This causes two problems for a recogniser. First, it may not be straightforward to distinguish between regions that are masked and those that are not. Second, even if masked regions can be identified, there may be no good way to deal with them: the information in them has typically been degraded to the point of being lost, i.e. the original speech energy in these regions is essentially unknown.

However, the speech signal has three important qualities that the SFD technique can exploit to form a robust recognition hypothesis. First, the speech signal has a redundant encoding such that it remains intelligible even if large spectro-temporal regions are removed. For example, speech can remain intelligible even when passed through a narrow bandpass filter (Warren et al., 1995). Experiments using mutually exclusive frequency bands demonstrate that there is no one frequency region that is essential for recognition. Hence, a certain degree of information lost due to masking will not necessarily impact on intelligibility. Second, speech energy is concentrated in local spectro-temporal regions. For example, the energy in a frequency channel close to a formant peak will be many times greater than the average energy in that frequency channel. This means that when speech is mixed with a masker there will typically be regions that are totally swamped by the masker and other regions where the amount of masker energy is insignificant compared to the speech. In these latter regions, 'fragments' of the clean speech signal will be visible. If the masker itself is non-stationary (e.g. another speech signal) then there is an increased likelihood of having even more fragments of uncorrupted speech. [1] Third, the speech signal possesses certain continuities that allow spectro-temporal fragments of speech to be identified, and segmented from fragments of competing sources (Darwin, 2001; Bregman, 1990). At the signal level there exist continuities in properties such as pitch, energy envelope and spatial location. Although these cues by themselves may not offer a complete segregation of sources – e.g. there are breaks in the pitch contour, and the energy envelope can have discontinuities – they may be sufficient to allow reliable clustering of time-frequency elements into *local* fragments. [2] A second stage employing statistical models of speech that capture higher-level structure, such as the continuity of voice qualities, e.g. vocal tract length, accent and gender can then group the local fragments to recover a description of the target speech source.

The current paper examines the performance of the SFD technique on the

[1] A recent 'glimpsing' account of speech perception has demonstrated that speech intelligibility can be well predicted by a model that detects regions of uncorrupted speech and exploits the information in them (Cooke, 2006).

[2] The term 'time-frequency element' is used throughout the paper to refer to the 1-frequency band/1-time frame 'pixels' that compose the time-frequency representation of the acoustic signal. A 'fragment' is made up of a collection of these elements (see Figure 4).

Speech Separation Challenge task described in the introduction to this special issue. Briefly, the Challenge task requires systems to report keywords embedded in a target utterance spoken by one of a closed-set of 34 speakers. The level of a simultaneous masker utterance is varied over a range of SNRs from +6 dB down to -9 dB. In order to be able to perform well in this task, it is important that systems are able to identify the target speaker and selectively report the target speaker keywords rather than words spoken by the masker. In previous attempts to apply the SFD system to this challenge (Barker et al., 2006), a top-down approach to the speaker-identification component of the problem has been applied. Each speaker model is hypothesised as a potential target and used to generate a recognition hypothesis, and then the best overall scoring recognition hypothesis is selected. This paper reviews the performance of this approach and examines the similarities and difference between the results obtained using SFD and those obtained by listeners. One of the key observations of Barker et al. (2006) is that, in certain conditions, the SFD technique makes many errors due to failure to correctly identify the target speaker. The current paper examines the reasons for these speaker identification failures and as a result of this analysis proposes a novel attention-driven speaker identification technique that is applied to produce new and improved results on the Challenge problem.

The structure of the paper is as follows. Section 2 provides a review of the SFD technique, highlighting the relation between SFD and current ideas about human speech perception. Section 3 summarises experiments initially reported in Barker et al. (2006) in which the SFD technique was applied to the Challenge problem. Section 4 identifies why the SFD technique fails in some circumstances and proposes a novel fragment-based speaker identification technique based on a model of keyword attention. The section includes new results obtained when the speaker identification module is incorporated into the system. Section 5 concludes the paper with a discussion relating the SFD technique to model combination techniques that have also been shown to solve the simultaneous speech recognition problem. The discussion also consider plans for future SFD developments which aim to further close the gap between machine and human recognition performance.

2 The speech fragment decoding system

Figure 1 provides an overview of the speech fragment decoding system. The components represented in solid lines have been evaluated in Barker et al. (2006) and Ma et al. (2007) and are reviewed below. This paper focuses on the novel speaker identification module (dotted oval).

The system has been inspired in large part by the Auditory Scene Analysis

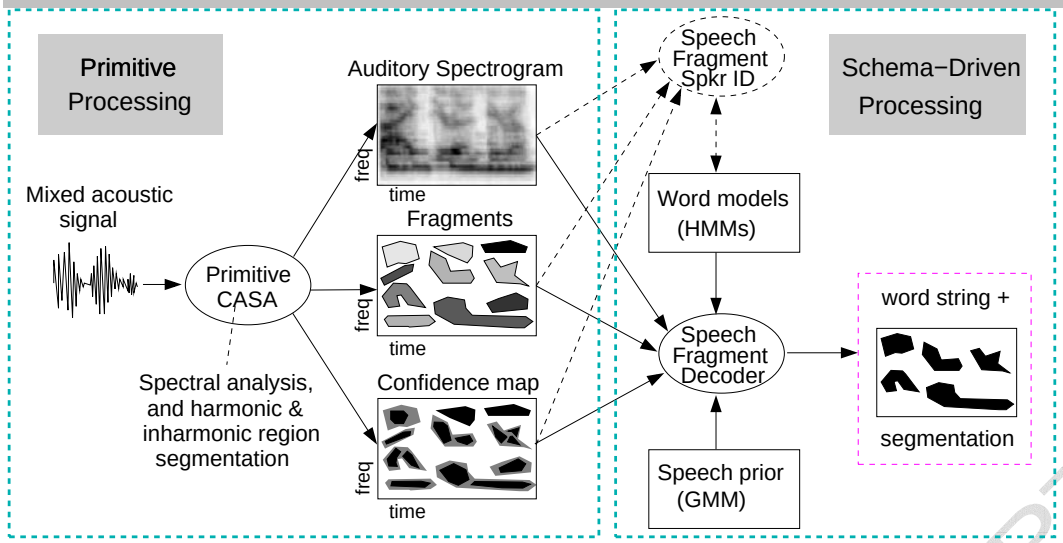


Fig. 1. *An overview of the speech fragment decoding system: Primitive processing identifies fragments and schema-driven processes match fragments to models of clean speech. The new fragment-based speaker identification module identifies the target speaker and hence selects the speaker-dependent HMMs to be used by the fragment decoder.*

(ASA) account of auditory organisation that describes how the perception of separated sound sources is derived from the mixed acoustic signal arriving at the ears (Bregman, 1990). Like ASA, the SFD system combines two separate processing stages: i) primitive processing – deterministic signal processing techniques that act to segment the spectro-temporal plane into a number of sound source fragment; ii) schema-driven processing – statistical model-driven processes that locate the most likely foreground/background segmentation and word sequence given the noise mixture and the set of fragments.

2.1 Primitive processing

The fragment generation component models aspects of the primitive processing stage of auditory scene analysis (ASA). The goal of this stage is to group time-frequency elements of a spectro-temporal representation of the acoustic signal in order to form a set of spectro-temporal fragments in which each fragment 'belongs' entirely to one sound source. A fragment is said to 'belong' to a sound source if the energy contributed to each time-frequency element of the fragment is dominated by energy from that source.

The techniques employed in the current system exploit the continuity of the pitch that occurs in regions of voiced speech. [3] The algorithm, which is il-

[3] In general, cues other than pitch could be used. For example, binaural data could be segmented using interaural time and intensity cues.

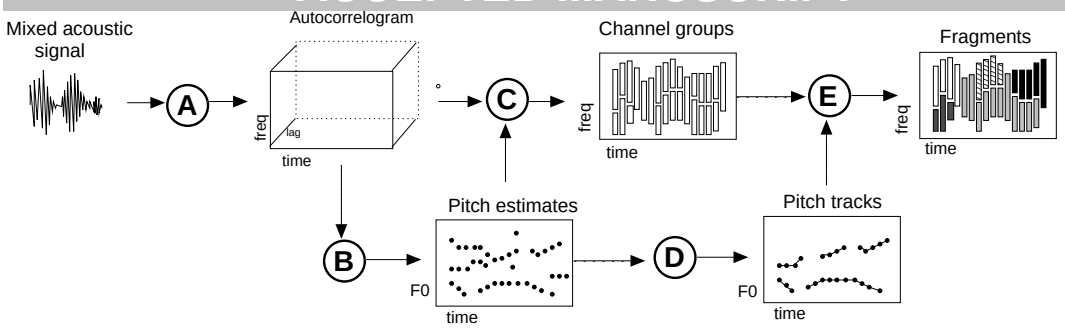


Fig. 2. *An overview of the CASA processing employed to generate the set of fragments (see text for details).*

lustrated in Figure 2, is briefly reviewed below and presented in full in Ma et al. (2007). It is based on three main steps, i) pitch estimation (**A**+**B**), ii) cross-frequency grouping (**C**) and iii) temporal integration (**D**+**E**).

First, an autocorrelogram representation of the mixed acoustic signal is computed (**A**); this involves first passing the acoustic signal through a gammatone filterbank (with filters spaced to mimic the non-linear frequency resolution of the ear) and then, at regular time intervals (here, 10 ms), performing an autocorrelation on the signals in each channel. For each time frame this produces a two-dimensional representation with axes of frequency and autocorrelation delay (lag). Ma et al. (2007) shows how these representations can then be processed to produce robust pitch estimates for multiple simultaneous sources (**B**).

Second, at each time frame, filter channels are grouped across-frequency into 'channel groups' if they appear to be dominated by the same harmonic source (**C**). This is done by comparing the periodicity of the signal in each filter channel with the harmonic source pitches that were estimated by stage B. Hence, each channel group is uniquely associated with a single pitch estimate. Filter channels in which the signal is not sufficiently periodic are left unassigned to channel groups and are later grouped to form separate *inharmonic fragments*.

Third, channel groups are integrated through time using a multi-pitch tracking algorithm (**D**+**E**). Coy and Barker (2007) describe a novel multi-pitch tracking algorithm that uses an HMM to model the change of voicing-state of a speech source, and a simple model of pitch dynamics within voiced segments. Independent HMMs are used to model each speech source, and a separate noise process is used to model the spurious pitch estimates generated by the pitch detection algorithm. Viterbi decoding is then able to form the most likely description of the data in terms of a number of potentially overlapping pitch track segments (**D**). Channel groups are then integrated into the same spectro-temporal fragment if their pitch estimates lie on the same pitch track segment (**E**).

A final stage, not shown in Figure 2, is used to capture time-frequency elements corresponding to regions of the signal that are not sufficiently periodic to be included in a harmonic channel group. These time-frequency elements are grouped into 'inharmonic fragments' by segmenting concentrations of inharmonic energy in the time-frequency representation using techniques commonly employed in image segmentation (Roerdink and Meijster, 2001).

The entire process results in the segmentation of the time-frequency plain into a set of fragments (some harmonic and some inharmonic) as depicted by the box labelled 'Fragments' in Figure 1.

2.2 Schema-driven processing

The schema-driven component of ASA is modelled by a hypothesis-driven fragment decoding process (see Figure 3). The goal of this process is to identify which of the source fragments 'belong' to the acoustic foreground (i.e. the target speaker) and which belong to the acoustic background (i.e. the masker speaker). Given a set of fragments F, a valid foreground/background segmentation can be constructed by taking a subset of fragments F_f and labelling them as foreground (shaded black in Figure 3), and taking the complement set of fragments F'_f and labelling them as background (shaded white). For any given segmentation hypothesis, the most likely word sequence (i.e. the speech recognition output) can be evaluated using HMMs trained on clean speech in conjunction with missing data ASR techniques (Cooke et al., 2001). In missing data ASR, a standard Viterbi decoding process is employed but the computation of the state likelihoods is adapted to consider whether or not features are part of the foreground or part of the background: spectral features that are part of the foreground (i.e. shaded black in the figure) are matched to the clean speech models. The remaining spectral features (shaded white) are regions where the speech has been energetically masked by the background. In these regions the speech energy is unobserved but is known to be less than that of the observed energy of the mixed speech and masker. The models are matched to the missing data by integrating over all possible values that the speech energy could have given this constraint.

The missing data ASR technique can be used to find a the word sequence with the highest likelihood for a given segmentation. The speech fragment decoder then effectively compares the likelihoods of the best word sequence *for all possible segmentations*. The segmentation and word sequence that is jointly the most likely is selected (right hand side of figure). While the number of possible segmentations is very large Barker et al. (2005) demonstrate how they can be evaluated in an efficient manner by sharing computation between separate segmentation decodings.

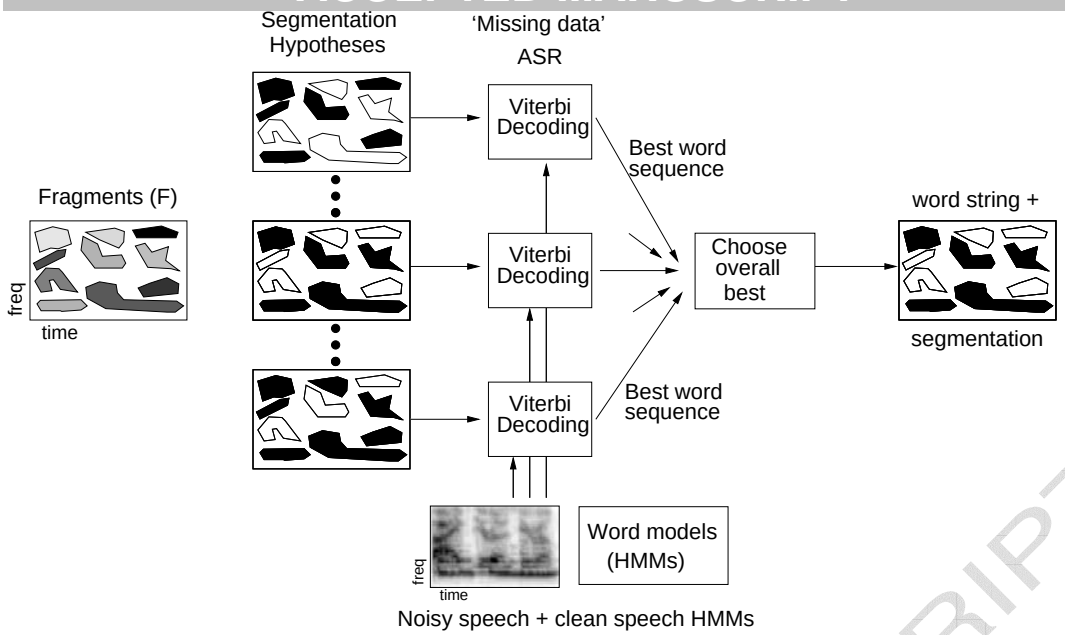


Fig. 3. *An overview of the top-down 'schema-driven' component of the SFD system. See text for details.*

The missing data systems on which SFD is based can employ either *hard* or *soft* segmentations. In a hard segmentation, every time-frequency element is treated as being wholly part of the background or wholly part of the foreground. When such systems are tested using segmentations that are known to be correct (e.g. that have been generated using prior knowledge of the unmixed sources), it is found that they are capable of producing very high recognition accuracies. However, when estimating the foreground/background segmentation from the data, the correct labelling of some time-frequency elements can be uncertain. It has been demonstrated that the problems of segmentation estimation can be reduced by labelling uncertain spectral-temporal elements with soft values (between 0 and 1) which express the degree of belief that the element belongs to either foreground or background (with 0.5 meaning that either interpretation is equally likely) (Barker et al., 2000). These soft values are then used to form a weighted interpolation between the foreground and background interpretations in the state-likelihood calculations.

In Coy and Barker (2007) these ideas were extended to the SFD system. The 'soft' decoder takes a spectro-temporal *confidence map*, c, as an additional input. The confidence map encodes the degree of belief that each element is a true member of the fragment. Confidence map values range from 0.5 (no confidence) to 1 (high confidence). The confidence map values can now be combined with the fragment labels to make a soft segmentation (often referred to as a 'soft missing data mask'). In regions covered by foreground fragments, the soft missing data mask takes values directly from the confidence map, c. In regions covered by background fragments, the mask takes the values $1 - c$.

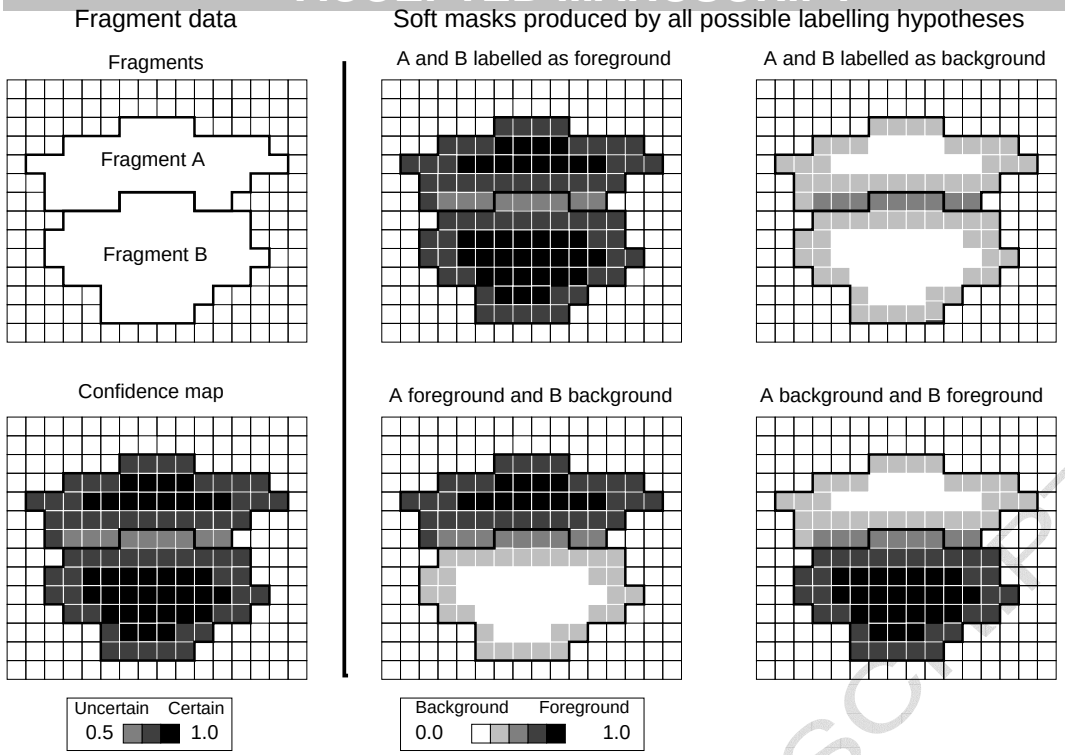


Fig. 4. *Left: Fragments A and B and a corresponding confidence map. Right: The soft masks that are produced for all four possible segmentation hypotheses that can be generated from two fragments.*

Hence, in regions where the confidence values are close to 1, the soft masks will have values close to either 1 or 0, and will therefore be similar to the discrete masks. In regions where the confidence values are low, (i.e. close to 0.5), the soft mask will have values close to 0.5 for both interpretations (see Figure 4). The result is that time-frequency elements with high confidence will have more influence on determining the fragment labelling than elements of low confidence.

3 Initial SFD Recognition Experiments

This section presents the simultaneous speech recognition evaluation task and reviews the SFD performance previously reported in Ma et al. (2007) and Barker et al. (2006). An analysis of the weaknesses of these systems motivates the novel development reported in Section 4.

3.1 Simultaneous speech data

The SFD system has been evaluated using simultaneous speech utterances constructed from the Grid corpus (Cooke et al., 2006) and in accordance with rules dictated by the Pascal Speech Separation Challenge.[4] The Grid corpus consists of utterances of the form indicated in Table 1 spoken by 34 speakers. In the present study, pairs of endpointed utterances are artificially added at a range of target-masker ratios (TMR). The 'colour' for the target utterance is always 'white', while the 'colour' of the masking utterance is never 'white', i.e. the word 'white' acts as a label which uniquely identifies the target speaker (it will henceforth be referred to as the 'identifier-word'). The task is to recognise the letter and digit spoken by the target speaker (i.e. by the person who utters the identifier-word 'white'). A full description of the preparation of the two speaker speech data is presented in Cooke et al. (2008b). The test set has 600 utterance pairs at each TMR. The 600 utterance test set splits into three roughly equal-sized sub-sets in which target and masker are i) the same talker (ST), ii) the same gender but different speakers (SG), and iii) of opposing gender (DG).

Table 1
Structure of the sentences in the Grid corpus.

VERB	COLOUR	PREP.	LETTER	DIGIT	ADVERB
bin	blue	at	A–Z	1–9	again
lay	green	by	(no 'W')	and 0 (zero)	now
place	red	on			please
set	white	with			soon

3.2 Recogniser configuration

A 64-channel log-scaled auditory spectrogram representation was employed. Briefly, the signal is filtered using a bank of 64 gammatone filters with centre frequencies spread evenly on an equivalent rectangular bandwidth (ERB) scale between 50 and 8000 Hz with filter bandwidths matched to the ERB of human auditory filters. The instantaneous Hilbert envelope of each gammatone filter output is computed. This envelope is smoothed by a first-order low-pass filter with an 8-ms time constant, sampled at 10 ms intervals and log-compressed to produce a series of 64-dimensional spectral feature vectors. A 128-dimensional feature vector is constructed consisting of 64 log-energy and 64 delta log-energy terms. Speaker-dependent word-level HMMs are trained

[4] http://www.dcs.shef.ac.uk/∼martin/SpeechSeparationChallenge.html

using 500 utterances from each of the 34 Grid speakers. Each word is modelled using 2 states per phoneme in a left-to-right model topology with no skips, and with 7 diagonal-covariance Gaussian mixture components per state. In order to normalise the missing data likelihood computation, the SFD needs a prior model for speech (i.e. the distribution of the observed spectral feature vector prior to conditioning on the HMM state) – use of this prior is mandated by Equations 18 and 21 of Barker et al. (2005). In Barker et al. (2005) a uniform prior was used. However, Coy and Barker (2007) has shown that better results are achieved by using a more accurate prior trained directly on the speech data. In the current work a GMM speech prior was constructed by training a set of speaker-dependent HMMs with a single mixture per state, and then pooling the Gaussians from all HMM states with weights scaled to correct for the differing prior probabilities of each HMM state. Tests on development data showed this prior to be more effective than priors constructed from HMMs with a greater number of Gaussians.

The recogniser employed a grammar representing all allowable Grid utterances in which the colour spoken is 'white'. In all experiments it was assumed that the target speaker is one of the speakers encountered in the training set, but two different configurations were employed: i) 'known speaker' - the utterance was decoded using the HMMs corresponding to the target speaker, ii) 'unknown speaker' - the utterance was decoded using HMMs corresponding to each of the 34 speakers and the overall best scoring hypothesis is selected. The 'unknown speaker' system is compliant with the rules of the Challenge. The 'known speaker' system breaks the rules by assuming knowledge of the target speaker identity. However, it provides an interesting control allowing the extent to which the 'unknown speaker' system errors are due to a failure to correctly identify the target speaker to be judged.

A global beam pruning algorithm (Noll et al., 1992) was employed to reduce the computational cost of decoding the 'unknown speaker' configuration. At each frame the highest scoring token across all speech states and all segmentation hypotheses is located. Then the percentage of tokens falling within a specified beamwidth is computed. When this percentage is too small the beamwidth is successively widened by a small fraction, else if it is too large the beam is successively narrowed. Tokens outside the beam are not propagated in the next frame. A small development set of 150 mixtures at 0 dB was used to select the target percentage of hypotheses to prune. It was found that this target could be raised to 90% without significant impact on the recognition result, and with a resulting reduction in decoding time of over 75%.

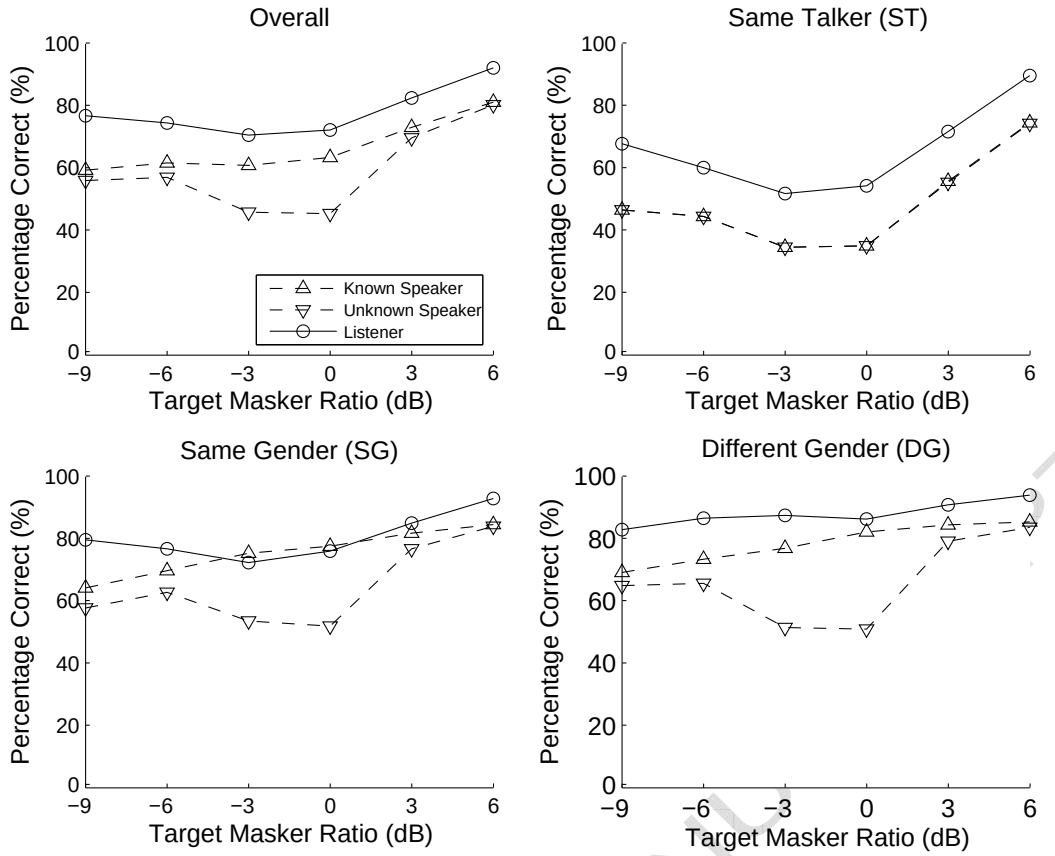


Fig. 5. *Results for the speech fragment decoder in known speaker and unknown speaker configurations compared against average listener results.*

3.3 Results

Previously reported recognition results for the SFD system are reviewed in Figure 5 (and Tables 2 and 3). [5] The results are plotted alongside those from a group of listeners (Cooke et al., 2008a).

The pattern of results can be broadly explained in terms of the combined effects of two types of masking: energetic masking, which occurs in time-frequency regions where energy due to the masker utterance dominates that of the target utterance preventing the extraction of reliable features; and informational masking, which is a masking effect that can occur after feature extraction and is partly related to the foreground-background confusion caused by potential similarity between fragments of the target and masker signals (see Durlach et al. (2003) for a detailed discussion of informational masking effects).

[5] These recognition scores are slightly higher than those reported in Barker et al. (2006) due to correction of an error in the application of the pitch tracker. See Section 4.1.1(I) of Ma et al. (2007) for details.

Consider first the listener data ('o'). In the DG condition there is a general decrease in scores from around 95% at 6 dB to around 80% at -9 dB. As the gender difference provides consistent cues for source separation, the performance is largely determined by the degree of energetic masking. This is in clear contrast to the ST condition in which, as the TMR drops from 6 dB down to 0 dB the performance falls very sharply to less than 60%. As the target and masker utterances are from the same speaker, and hence have the same F0 range and identical vocal tract length, the level difference is the only consistent cue for distinguishing the competing utterances. Therefore, at 0 dB, where even the level cue is removed, performance is very poor. Interestingly, as the TMR descends below 0 dB, the recognition performance starts to improve again despite increased energetic masking. This pattern is consistent with the effects of informational masking reported by Brungart et al. (2001). Note also that performance at -9 dB in the ST condition (about 65%) is significantly worse than in the DG and SG conditions. This is likely to be an energetic masking effect. In the ST condition the speech energy of the target and masker tend to be located in the same frequency regions (e.g. around the speaker's average formant positions). Finally, looking at human performance in the SG condition, in which both target and masker have a similar F0 range and a similar (but not identical) vocal tract length, it can be seen that the performance curve lies somewhere between that for the ST and the DG cases. There is again a dip at around 0 dB but in this case it is less pronounced.

Now consider the performance of the SFD system (triangles). The unknown speaker configuration, although having approximately double the error-rate of the human listeners, shows a remarkably similar pattern of overall performance. As was observed with listeners, scores decrease as TMR falls towards 0 dB and recovers again as the TMR descends further. At 0 dB the SFD system has a similar problem to that of listeners. Fragments of both the target and masker are at the same level as the fragments of the speaker identifier-word, 'white'. This results in errors when hypotheses passing through the masker HMM have a higher likelihood than that of the target. However, the target will generally be favoured because the grammar forces the decoding through the colour 'white' which is known to be spoken by the target speaker.

Although SFD seems to model the trends in listeners for the ST and SG condition, it fails in the DG condition. Here, at 0 dB, listeners have no problem in distinguishing target and masker while the SFD suffers a large drop in performance. This is possibly due to the failure of the SFD system to fully exploit pitch information. Although pitch is used as the basis for forming source fragments, it is not used explicitly to group fragments through time. Such sequential grouping is achieved by scoring each fragment's compatibility with the HMMs of the various speakers.

Finally, consider the differences between the 'known' and 'unknown' speaker

configuration of the SFD. As expected, the known speaker system has far better performance. When the model for the target speaker is provided, confusions between foreground and background that reduce performance at 0 dB are removed. The performance curves for the DG and SG cases now seem to be subject only to the effects of energetic masking. At 0 dB, although masker fragments are at the correct level, they do not match well to the speaker-dependent models of the target speaker. Note, however, that informational masking effects are again present in the ST condition. In this case, although speaker-dependent models for the target speaker are provided, confusion occurs because these models can also match the fragments due to the masker. Again, the decoder can be drawn into incorrectly reporting the masker utterance.

Although the 'known speaker' system performs better than the 'unknown speaker' system, the latter system is a perhaps a better model of the situation faced by listeners when performing the task. Listeners are only asked to report the letter and digit of the utterance containing the word 'white'. They are not given advanced knowledge of the identity of the target speaker. As an alternative experiment, that would be closer in spirit to the 'known speaker' configuration, subjects could be presented with several utterances of the target speaker immediately prior to being presented with the mixture. If listeners are able to use the target speaker utterances to prime speaker-dependent models, then it might be predicted that listener results in these conditions would show a much smaller effect of informational masking. This may be especially the case if the target speakers are familiar to the listener.

The 'unknown speaker' SFD system has a result pattern that can be explained by the interplay of effects similar to energetic and informational masking in the HSR results. However, there are a number of differences in the ASR and HSR set-up that make it inappropriate to draw strong conclusions from this similarity. The ASR system has two particular advantages over the listeners. First, the ASR system uses a closed-set of speaker-specific models. Although listeners may have become familiar with the speakers in the corpus throughout the course of the listening experiment, they were not given the opportunity to learn speaker characteristics in advance. However, listeners may be able to adapt rapidly to the target speaker which may mean that speaker-dependent models are a better model of listener performance than, say, a fully speaker-independent model. Second, in the ASR experiments, the target speech was presented at a fixed level that matches the level of the training data. Consequently, the ASR system was able to use absolute level to identify fragments as belonging to the target. In contrast, the HSR experiments employed a roving level with the explicit aim of disrupting such a strategy (Cooke et al., 2008a). Hence, in the HSR experiments, although *relative* level may be used to sequentially group fragments that appear at the same level, *absolute* level cannot be used to directly label fragments as foreground/background.

Table 2
Keyword recognition correct percentage (%) in known speaker configuration.

	TMR (dB)					
Condition	-9	-6	-3	0	3	6
Overall	59.42	61.75	60.92	63.50	73.17	81.25
ST	46.61	44.57	34.62	35.07	55.88	74.66
SG	64.25	69.83	75.42	77.65	81.84	84.64
DG	69.25	73.50	77.00	82.25	84.50	85.50

Table 3
Keyword recognition correct percentage (%) in unknown speaker configuration.

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

4 Attention-driven fragment-based speaker identification

4.1 Introduction

The previous section demonstrated that when the SFD system is run in the 'known speaker' configuration, results in most conditions are effectively free from the effects of informational masking. However, when the target speaker identity is *unknown*, the decoder is subject to large informational masking effects. In the TMR range -3 dB to 0 dB the decoder is essentially unable to distinguish fragments of the target and masker, resulting in a large drop in performance in all target/masker gender conditions. It appears that, unlike listeners, the decoder is little able to use speaker differences alone to reliably follow the correct source. In particular, this shortcoming persists even in the different gender condition. This may appear surprising considering the large acoustic differences that exist between the speaker-dependent models. However, this outcome is understandable given the nature of the recognition task. At TMRs close to 0 dB the target is only identified by the fact that the target speaker says the word 'white'. The system is, in effect, required to solve a speaker identification problem using a single word in the presence of substantial energetic masking.

Consider the demands of the speech separation challenge task. In order to identify the target speaker, listeners must pay specific attention to the identifier-word, 'white,' as it is the only word known a priori to be spoken uniquely by the target speaker. Some mechanism is then needed to associate the word 'white' with the grid reference occurring later in the utterance. This may be done in a bottom-up manner by tracking low-level properties such as pitch, or in a top-down manner by exploiting high-level invariances such as the vocal tract length or accent of the speaker of the identifier-word. In principle, the SFD can model both approaches. If pitch can be tracked from the identifier-word to the letter-number keyword combination, then energy from each region can be incorporated in the same fragment. In practice, this does not happen since discontinuities in voicing lead the primitive grouping process to segment the mixture into shorter segments, typically of the duration of a syllable. Fragments of the identifier-word and the letter-digit keywords must be linked by top-down mechanisms, i.e. because they both match well to a specific speaker-dependent model.

In the previous section, top-down tracking was implemented by decoding the fragment set using models for each potential target speaker and then selecting the decoding from the speaker that gave the highest overall utterance likelihood. However, top-down tracking can be unreliable. It is possible that the masker speaker is selected even if it is a poor local match to the colour identifier-word, since a good fit to masker fragments over the remainder of the utterance can cause it to score better than the target speaker model overall. This is particularly the case if the identifier-word forms only a short portion of the utterance.

Evidence that the SFD makes this type of error can be seen by examining the speaker identities associated with the ASR hypothesis generated in the unknown speaker configuration. This can be done by investigating the Viterbi backtrace to determine which of the parallel speaker HMMs the winning hypothesis had passed through. Table 4 presents target speaker identification accuracies computed in this way. For conditions where the target and masker are different speakers, the figures in brackets indicate the percentage of times that the target speaker was incorrectly identified as the masker speaker. At 0 dB the target is being confused for the masker in nearly 40% of cases. If these confusions could be reduced by a model that paid closer attention to the identity of the speaker of the word 'white' then the recognition result at 0 dB could be greatly improved.

Table 4
Target speaker identification accuracy (%) produced by current SFD system in unknown speaker configuration. Figures in brackets indicate the percentage of mixtures for which the target is misidentified as the masker.

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

4.2 Attention-driven speaker identification

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

$$utterance ::= W_b, W_i, W_e \tag{1}$$

where W_b (beginning), W_i (identifier) and W_e (end) are three non-terminal grammar items that generate word sequences that are concatenated to form the utterance. W_i generates a sequence of *identifier*-words that uniquely identifies the target speaker, i.e. sequences generated by the grammar segment W_i cannot occur in the masker utterance. It is also required that a set of speaker-dependent models exists for each candidate target speaker.

An utterance-level speaker-dependent HMM for each potential target speaker is constructed according to the above grammar. These speaker-dependent utterance models are placed in parallel as illustrated by the network in Figure 6. The speaker identification mechanism operates by examining token scores generated by the SFD during decoding of the noisy utterances. Let $S_e(t, seg, n)$ be the scores of tokens that arrive at time t in the non-emitting node at the end of W_i for each foreground/background segmentation hypothesis, seg, and for each speaker, n. As the identifier-words will match well to the target model around the time when they finish, tokens through the target model can be expected to have higher likelihoods than those of tokens that have passed through the other speaker models.

To eliminate the contribution to the token score that has been made by word models in the utterance prior to the identifier-words, (i.e. during W_b), each token maintains a record of its score, $S_b(t, seg, n)$, when it first entered the identifier-word sequence, W_i. Tokens also keep a record of the duration spent

traversing the identifier-word sequence model, $D(t, seg, n)$. The score S_b is then removed from the end score S_e in the logarithmic domain and normalised by dividing by $D(t, seg, n)$ to reveal the score accumulated during W_i alone. So for the token in the nth speaker model of segmentation, seg, arriving at the end of W_i at time t, the score is computed as,

$$S'(t, seg, n) = \frac{S_e(t, seg, n) - S_b(t, seg, n)}{D(t, seg, n)} \tag{2}$$

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

$$target = \underset{n}{\text{argmax}}[\max_{t,seg} S'(t, seg, n)] \tag{3}$$

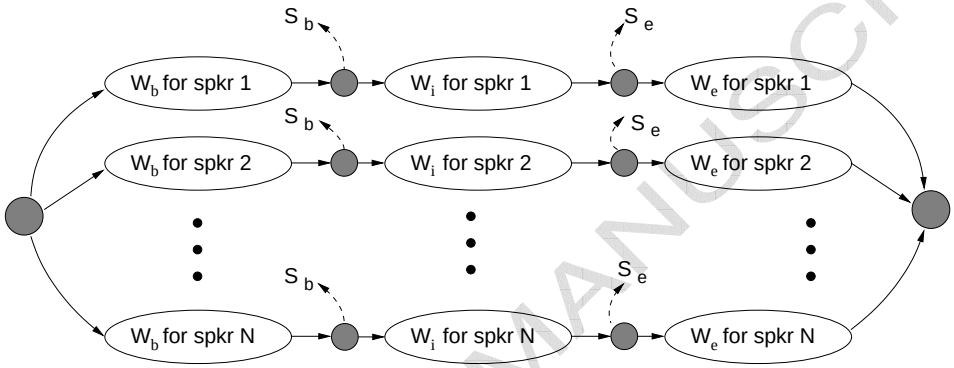


Fig. 6. *A decoding network for the current system. White ovals represent word-level HMMs, and grey circles are non-emitting nodes connecting the HMMs. In each frame, a record is made of the score ($S_e - S_b$) of each token arriving in the non-emitting states at the end of the identifier-word sequence, W_i, for each speaker and each segmentation (not shown). The score is normalised by dividing by the number of frames spent traversing W_i.*

An example of the scores generated when applying this task to the speech separation challenge data is shown in Figure 7. For the challenge task, the grammar for the identifier-word sequence, W_i, is simply the word 'white'. In this example, the target is a female speaker saying 'place white in G 6 please' and the masker is another female speaker saying 'lay green at Q 0 now'. The oracle segmentation is displayed in panel (b). The dashed vertical line indicates where the word 'white' finishes. In panel (c) the solid line shows token score for the best segmentation for the target speaker and the dashed lines are the token scores for the best segmentations of the remaining speakers. In the first 0.2 seconds no valid tokens have reached the non-emitting node at the end of 'white'. When valid tokens start to arrive at the non-emitting node, initially, the scores are low for all speakers because the observations do not fit well

to any model of 'white'. Around the time when 'white' finishes (indicated by the dotted vertical line at 0.7 seconds) the target speaker receives tokens with considerably better scores, causing a significant jump, while the scores of the other speakers are still relatively low. This is the time when the fragments forming the word 'white' match the target speaker model. For later frames, the scores of the tokens arriving in the non-emitting node become lower and lower.

Note, although in the current work the identifier-word is known to occur at a fixed position in the utterance, the speaker identification technique allows for more general situations. The grammar of the segment prior to the identifier-word sequence, W_b, does not need to represent a fixed number of words. For example, W_b may be an arbitrarily lengthened sequence of words taken from a vocabulary that excludes the keyword, in which case the speaker identification algorithm would essentially be co-occurring with a general keyword spotting task. The algorithm summarised by equations 2 and 3 would remain the same.

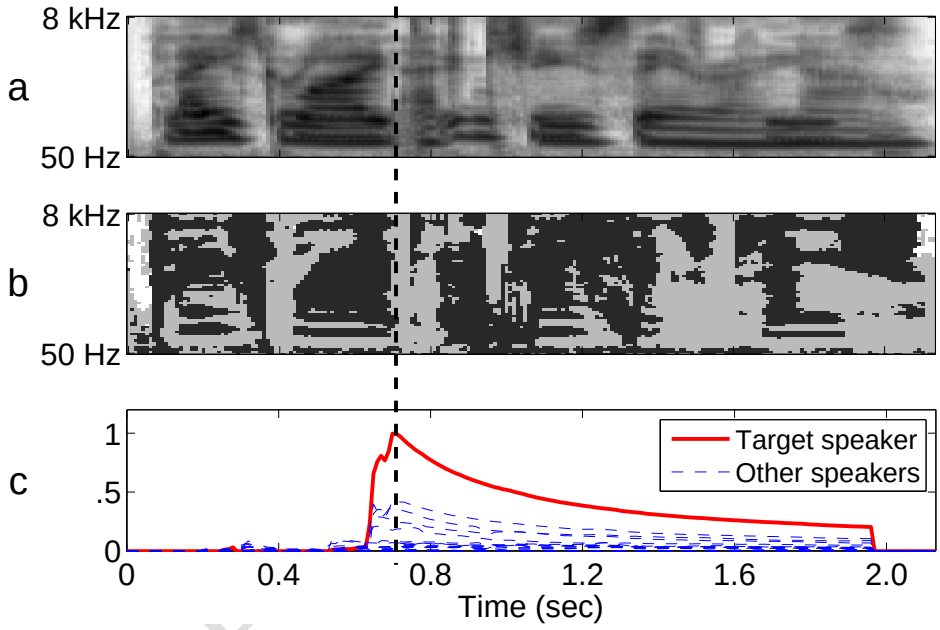


Fig. 7. *(a) A 'ratemap' representation of the mixture of 'place white in G 6 please' (target, female) plus 'lay green at Q 0 now' (masker, female) at a TMR of 0 dB. The dotted vertical line indicates where the word 'white' finishes. (b) The 'oracle' segmentation: dark grey – pixels where the value in the mixture is close to that in the target speech; light grey – the mixture value is close to that in the masker speech; white – low energy regions. (c) The score S' computed for each speaker at each frame in the mixture. The trace for the target speaker is shown as a solid line. The scores have been scaled by dividing by the peak S' value.*

Results for the target speaker identification approach are presented in Table 5. It can be seen that at TMRs above -6 dB the attention model offers significantly better results than the original SFD system. The 0 dB TMR

that previously caused problems now has a speaker identification accuracy of 96.3% and the masker speaker is very rarely selected. Note that the reason this technique is successful is largely due to the manner in which the speech mixture has been initially separated into spectro-temporal fragments dominated by individual sources. The token scores are generated by the speech fragment decoder based on the most likely fragment combination in each frame. Hence, the system is performing *fragment-based* speaker identification. If the system was constructed using a conventional Viterbi decoder, computing token scores based on full-band observation evidence, the token that passes through the model for the target speaker would not necessarily generate the peak S' score.

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

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

A second strategy that listeners could be using, and which is also not modelled by our system, is to infer that if they have not heard the word 'white' then the target is probably the quieter of the two speakers, and therefore they should focus attention on the quieter speaker when listening for the grid reference. These strategies could be modelled by using a parallel invocation of the attention-driven speaker identification mechanism that, by using an identifier-word grammar of $(blue|green|red)$, would identify the *masker* speaker rather than the target. Depending on whether it was the target speaker or the masker speaker that was more reliably identified, the utterance could either be decoded with the target speaker models, or a parallel combination of all speaker models except the masker respectively. Judging the reliability of the speaker identification would require a suitable confidence score; a possible candidate would be the difference of the scores, S', between the highest and second

highest scoring speaker.

Table 5
Target speaker identification accuracy (%) based on token scores. Figures in brackets indicate the percentage of mixtures for which the target is misidentified as the masker.

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

4.3 Employing speaker identification in the SFD ASR system

In the final set of recognition experiments, the SFD system was evaluated in the 'known speaker' configuration using the model of the speaker that has been identified by the speaker identification technique described above. If a failure of speaker identification was automatically to lead to recognition errors, then the performance of this complete system could be estimated by combining the speaker identification scores and the results of the previous 'known speaker' system. However, this assumption cannot be made. For example, if the corpus contains a pair of speakers (A and B) who have accents that are sufficiently close, then speaker identification errors resulting from A/B confusions would be likely, but it may still be possible to recognise speaker A's utterance correctly using speaker B's models, and vice versa. If speaker-dependent models overlap it may even be that using the speaker model that is selected by the data produces better results than using the target speaker's true model.

Figure 8 shows the new recognition results (Table 6) plotted against those previously shown for the SFD system in both 'known speaker' and 'unknown speaker' configurations. Listener results have also been plotted for the sake of comparison. At all TMRs, except -9 dB, the SFD system employing speaker identification produces recognition performance that is just a little less than the 'known speaker' configuration. In fact, the results obtained can be nearly precisely modelled by taking the 'known speaker' result in cases where the speaker identification has been correct, and taking chance level performance where speaker identification is incorrect (chance performance on this task is 7%).

Note that both the unknown speaker system and the system using the speaker identification step are compliant with the rules of the Pascal Speech Separa-

tion Challenge, as in both the identity of the target speaker has been inferred from the data rather than explicitly provided. However, the system using the speaker identification stage has considerably better performance. Improvements are especially significant at -3 dB and 0 dB where previously informational masking effects frequently led the system to incorrectly report the masker utterance. The speaker identification module is not particularly sensitive to informational masking. Although at 0 and -3 dB a greater proportion of the identification errors are due to the target being misidentified by the masker, the number of errors at these TMRs is not significantly higher than may be expected from energetic masking alone. Informational masking effects are presumably less significant in this task because top-down knowledge is highly constrained. If target/masker confusions do occur it must be because the masker's model for the word 'white' is a better fit to the observed data than the target's. This is quite different to the target/masker confusions that occur in the identification of the grid reference, which happen when the ASR system (or listener) may have some notion of who the target speaker is (i.e. who said 'white') but does not known what letter and digit are to be spoken.

Finally, comparing the SFD system with human listeners, the system has an overall error rate that is almost twice as large at most TMRs. This may be due to the use of HMM representations of speech that are known to be inadequate in many respects. However, there are a few details of the HSR/ASR comparison that deserve specific note.

In the SG condition, in contrast to listener results, there is no dip in ASR performance at -3 and 0 dB, and the system achieves the same level of performance as humans at these TMRs. Table 4 shows that even in the SG condition the system is able to reliably identify the target speaker using the identity-word fragments. Once the target speaker identity is known the system will reliably report the target speaker's keywords. Listeners on the other hand will often report the keywords spoken by the masker speaker. It is possible that listeners experience increased difficulty identifying and tracking the target speaker because, unlike the SFD system, they do not have access to pre-trained models of the potential speakers. It would be instructive to repeat the listener experiments after having allowed the listeners to familiarise themselves with the speakers in the Grid corpus. This may result in listeners exhibiting greatly reduced informational masking effects in the SG condition.

In the ST condition the SFD system performs relatively poorly compared to humans. In this condition, although the identity of the target is clear (i.e. the speaker identification module makes few errors), target identification alone is not sufficient to avoid foreground/background confusions. It is necessary for the system to track subtle acoustic cues in order to group fragments of the identifier-word 'white' with the fragments of the Grid reference spoken by the *target* speaker. Improvements to the primitive grouping stage to allow pitch

to be reliably tracked over longer segments would no doubt help close the gap. However, listeners may also be using more subtle top-down cues. They are potentially aware of long term inter-dependencies in speech that arise from variations in the manner of speaking even between pairs of utterances of the same speaker. The speaker-dependent models used by the SFD system, which average over all inter-speaker variability, have no access to such cues.

The final notable HSR/ASR difference is that the SFD system decreases in performance between -6 and -9 dB whereas human results show an improvement. This is partly because, as explained earlier, the SFD system has no strategy to cope with the case where the word 'white' is fully masked. However, the SFD system shows a decrease at -9 dB even in the 'known speaker' system. It appears that the SFD system is more susceptible to the effects of energetic masking. Despite speech energy being concentrated in local regions, at small TMRs the effect of energetic masking starts to become very significant, and fragments of the target get rapidly smaller. The letter-digit recognition task needs access to subtle phonetic cues to distinguish many of the letters (e.g. 'm' and 'n'; 'p' and 'b'). It is likely that listeners have both better source separation algorithms and acoustic models more sophisticated than the HMMs employed in the SFD. Even small improvements to either component may make a big difference in extreme conditions.

Table 6
Keyword recognition correct percentage (%) with attention-driven speaker identification.

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

5 Discussion and conclusions

The paper has reviewed a fragment-based approach to robust ASR which works by coupling the problems of foreground/background segregation and speech recognition. Whereas most robust ASR techniques have problems in non-stationary noise conditions, the SFD system mimics listeners in that it is able to take advantage of the fact that non-stationary noises provide unmasked glimpses of the target speech source. Recognition performance is significantly above that of a conventional HMM ASR system (see the introduction to this

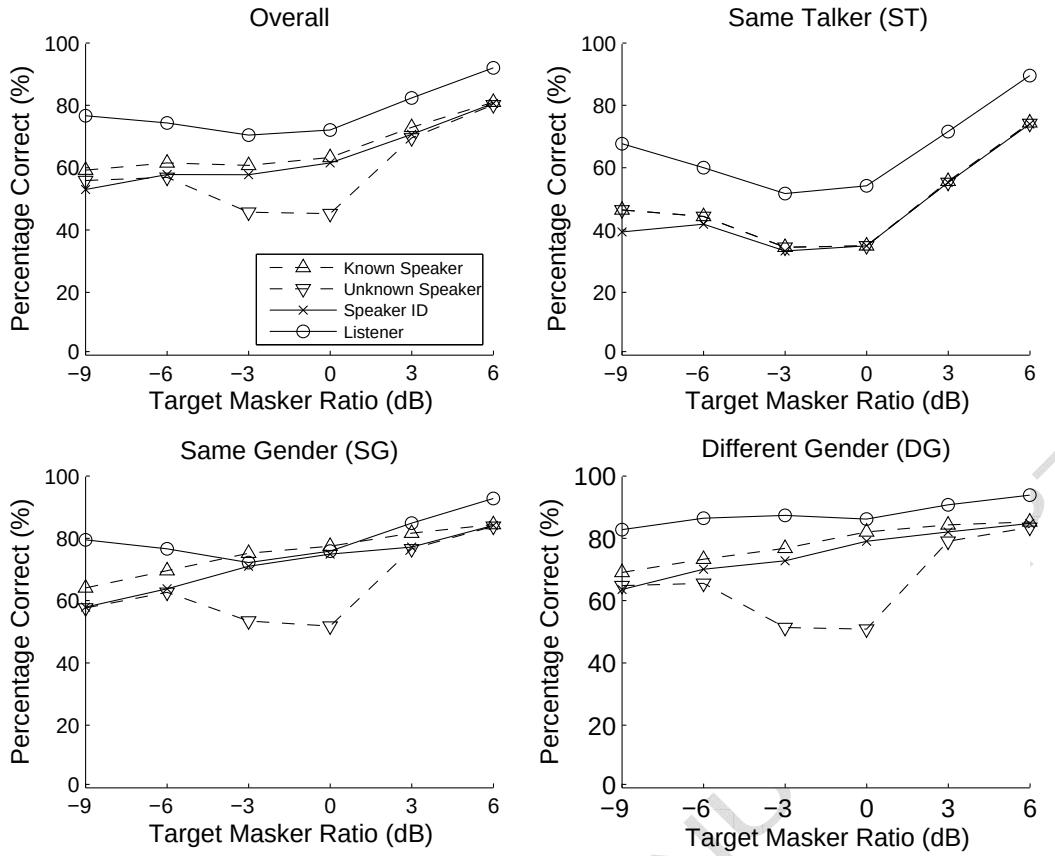


Fig. 8. *Keyword recognition results for the SFD system incorporating the attention-driven speaker identification model compared against listener data and the known speaker and unknown speaker SFD configurations previously reported in Barker et al. (2006).*

special issue for baseline results), and is relatively insensitive to the noise level over a broad range of TMRs. The system has performance curves similar to those of listeners with characteristic dips around 0 to -3 dB TMR in the ST and SG conditions. The paper has also shown how the fragment-based approach can be adapted to reliably identify the target speaker and eliminate nearly all of the target/masker confusions that occurred in the original systems of Barker et al. (2006) and Ma et al. (2007). Results of the system around 0 dB are greatly improved, especially when considering mixtures of speakers of identical gender, where the SFD performance at 0 dB is not significantly different from that of listeners.

A key strength claimed for the solution is that it does not need tailoring to the specific details of the additive noise environment. Considering the current implementation, notwithstanding the assumption made by the pitch estimator that there are at most two harmonic sources with significant energy at any time instant, the fragment generation stages restrict themselves to using only very general properties of the way that sounds combine. They do not

make assumptions about the specifics of either the foreground or background sources. Fragments of the target speech source will be grouped irrespective of the nature of the background. For example, the same technique has been applied in the past with both speech (Coy and Barker, 2007) and non-speech backgrounds (Barker et al., 2005). (Clearly though, some backgrounds may be more disruptive than others). The top-down component only requires a statistical model for the target speech. There is no statistical model representing the background.

The fragment decoding approach contrasts with other techniques that rely on detailed models of both foreground and background. For example, a common approach is to train HMMs for both the target and the masker speaker and then to model how pairs of acoustic states combine. Combining detailed models of this type can lead to much better results than those reported here for this problem (e.g. Kristjansson et al., 2006; Virtanen, 2006). However, there is an assumption here that one has access to models of the background speakers. In the Challenge data the masker utterance was chosen from the same closed-set of speakers that provided the target utterance, and for which training data was available. Hence, background models were readily available. If the target utterances had been mixed with maskers from a different set then speaker-dependent model composition strategies would not be available. An alternative strategy would be to combine the speaker-dependent foreground models with a speaker-independent background model. The lack of specificity of the background model would presumably lead to poorer results. If the separation task was further generalised so that the background contained one *or more* masker speakers, then the model combination approach would become even more difficult to apply. The SFD approach described here, however, could be applied with essentially no change, and providing good quality fragments could be located, it would be expected to produce a good quality recognition result.

Obviously, if knowledge of the background does happen to be available then it should not be ignored. The fragment generating front-end employed here lessens the need for a strong background model but it does not make such a model obsolete. Fortunately, the fragment-based front-end is not incompatible with background modelling. If an HMM for the background source were available then fragments could still be assigned to either foreground or background, and likelihoods could still be maximised jointly over both fragment assignment and state sequences for both the foreground and background models. The probability calculations would require some modification and the state-space would be larger but the technique would be otherwise similar to the standard SFD approach presented here. Further work is needed to examine ways in which background knowledge can be smoothly integrated into the fragment-based architecture, and to examine how the fragment constraints can be exploited to adapt and elaborate existing models of both the foreground

and background.

Further work is also needed to extend the technique to more realistic situations where speaker-dependent models are not available a priori. If the speaker-dependent models are simply replaced with speaker-independent speech models the speech fragment technique would still be expected to work in situations where the noise background is sufficiently non-speech-like. However, it would clearly experience difficulties in the speech separation condition studied in this paper where both target and masker fragments would match well to the non-speaker-specific models. A potential solution lies in the on-line adaptation of speaker independent models toward the target speaker. Techniques need to be developed that exploit the partial segmentation produced by the fragment generation process to allow robust adaptation of the clean speech models using noisy speech data. The adaptation and training of clean speech models from noisy speech data remain challenging problems in need of further research effort.

References

Assmann, P., Summerfield, Q., 1990. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. Journal of the Acoustical Society of America 88 (2), 680–697.

Barker, J., Cooke, M., Ellis, D., 2005. Decoding speech in the presence of other sources. Speech Communication 45 (1), 5–25.

Barker, J., Coy, A., Ma, N., Cooke, M., 2006. Recent advances in speech fragment decoding techniques. In: Proc. Interspeech 2006. Pittsburgh, pp. 85–88.

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. ICSLP 2000. Beijing, China, pp. 373–376.

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

Brungart, D. S., Simpson, B. D., Ericson, M. A., Scott, K. R., 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. Journal of the Acoustical Society of America 100, 2527–2538.

Cooke, M., 2006. A glimpsing model of speech perception in noise. Journal of the Acoustical Society of America 119, 1562–1573.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America 120, 2421–2424.

Cooke, M., Garcia Lecumberri, M., Barker, J., 2008a. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. Journal of the Acoustical Society of America 123, 414–427.

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

Cooke, M., Hershey, J., Rennie, S., 2008b. The speech separation and recognition challenge. Computer Speech and Language (this issue).

Coy, A., Barker, J., 2007. An automatic speech recognition system based on the scene analysis account of auditory perception. Speech Communication 49 (5), 384–401.

Darwin, C., 2001. Auditory grouping and attention to speech (keynote paper). In: Proceedings of the Institute of Acoustics. Vol. 23. pp. 165–172.

de Cheveigné, A., 1993. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. Journal of the Acoustical Society of America 93 (6), 3271–3290.

Durlach, N., Mason, C., Kidd, Jr., G., Arbogast, T., Colburn, H., Shinn-Cunningham, B., 2003. Note on informational masking. Journal of the Acoustical Society of America 113 (6), 2984–2987.

Hirsch, H., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ICSLP 2000. Vol. 4. pp. 29–32.

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., Gopinath, R., 2006. Super-human multi-talker speech recognition: The IBM 2006 Speech Separation Challenge system. In: Proc. Interspeech 2006. Pittsburgh.

Ma, N., Green, P., Barker, J., Coy, A., 2007. Exploiting correlogram structure for robust speech recognition with multiple speech sources. Speech Communication 49, 874–891.

Meddis, R., Hewitt, M., 1992. Modeling the identification of concurrent vowels with different fundamental frequencies. Journal of the Acoustical Society of America 91 (1), 233–245.

Noll, A., Ney, H., Mergel, D., Paeseler, A., 1992. Data driven search organization for continuous speech recognition. IEEE Trans. Speech and Audio Processing 40, 272–281.

Roerdink, J., Meijster, A., 2001. The watershed transform: Definitions, algorithms and parallelization strategies. Fundamenta Informaticae 41 (12), 187–228.

Virtanen, T., 2006. Speech recognition using factorial hidden markov models for separation in the feature space. In: Proc. Interspeech 2006. Pittsburgh.

Warren, R., Riener, K., Bashford, J., Brubaker, B., 1995. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. Perception and pyschophysics 57 (2), 175–182.