Recent advances in speech fragment decoding techniques

Jon Barker, André Coy, Ning Ma and Martin Cooke

Department of Computer Science
University of Sheffield, Sheffield, UK
{j.barker, a.coy, n.ma, m.cooke}@dcs.shef.ac.uk

Abstract

This paper addresses the problem of recognising speech in the presence of a competing speaker. We employ 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 spectro-temporal 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 reports recent advances in this technique, and presents an evaluation based on artificially mixed speech utterances. The fragment decoder produces significantly lower error rates than a conventional recogniser, and mimics the pattern of human performance whereby performance *increases* as the target-masker ratio is *reduced* below -3 dB.

Index Terms: speech recognition, speech separation, simultaneous speech, auditory scene analysis, noise robustness.

1. Introduction

Often when we are listening to someone speak there are one or more competing speakers talking ‘in the background.’ With little effort we can effectively tune out the background speakers and understand the speaker of interest with little or no loss in intelligibility. How is this separation performed? Although localisation cues help in this task, they are not necessary. Listeners are able to effectively separate speech sources even when they are presented in a single channel (e.g. consider the situation where someone telephones from a noisy party).

Traditional robust automatic speech recognition (ASR) approaches perform poorly in simultaneous speech conditions. Most approaches rely on the speech and the noise having very different characteristics. For example, spectral subtraction assumes that the noise is more stationary than the speech signal [1]. RASTA processing assumes that the speech dominates the noise across a narrow range of modulation frequencies [2]. For such systems the simultaneous speech problem is pathological because the noise *is* speech.

In this paper we present a recognition system based on a technique called speech fragment decoding (SFD) [3]. This technique exploits the non-stationarity of the speech signal. Speech is sparsely encoded with most of the energy concentrated in compact time-frequency regions. This sparsity means that when viewed in the spectral-temporal domain a mixture of two speech sources can be approximated as a series of interleaved fragments of the two unmixed signals. The SFD technique works by first segmenting the spectro-temporal plain into a set of source fragments, and then collecting together the subset of fragments that best fits a model of the target speech.

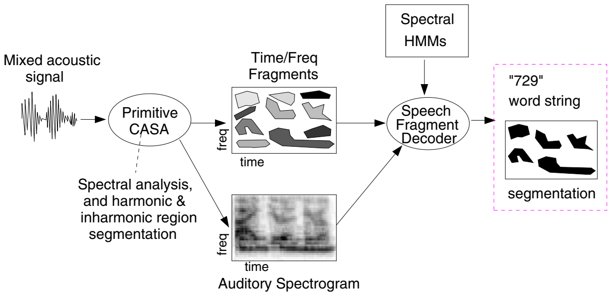

Figure 1: *A overview of the speech fragment decoding system.*

The paper evaluates the SFD approach using a small vocabulary simultaneous speech task, and compares the performance of the technique against that of listeners. Section 2 presents an overview of the basic SFD system. Section 3 describes some recent advances to the SFD approach. Section 4 presents the experimental results and draws comparison with human data. Finally, Section 5 concludes and presents some possible future research directions.

2. The speech fragment decoding technique

2.1. Overview

Figure 1 shows an overview of the speech fragment decoding system. There are two important components: i) the fragment generation process, which in our current implementation exploits models of primitive auditory scene analysis (ASA) and ideas taken from image processing; ii) the fragment decoding algorithm, which is based on an extension of the standard probabilistic theory for ASR to cater for the presence of multiple sound sources.

2.2. Fragment Generation

A spectro-temporal ‘ratemap’ representation of the acoustic signal is formed by first passing it through a 64 channel gammatone filter bank. The filter output is halfwave rectified, smoothed using a first order filter with a time constant of 8 ms, sampled at 10 ms intervals, and then log compressed [4]. We then analyse the ratemap to locate ‘coherent’ fragments - that is, spectro-temporal regions that are dominated by a single source (see sketch in Figure 1). The fragment generation proceeds in two passes. The first pass

Research supported by EPSRC grants GR/R47400 and GR/T04823.

identifies ‘harmonic fragments’. In brief, a technique exploiting the dendritic structure of the autocorrelogram is employed to form multiple pitch estimates for every 10 ms frame [5]. Then a multiple pitch tracking algorithm is employed to find smooth tracks through this raw data [6]. Each continuously voiced pitch track segment is converted into a spectro-temporal fragment by finding the frequency channels whose periodicity best matches the pitch estimate at each point in time. Once the time-frequency elements dominated by harmonic energy have been segmented in this way, the remaining elements - dominated by inharmonic energy - are segmented in a second pass using the Watershed algorithm. This is a technique commonly employed in image processing which acts to place segment boundaries between the spectro-temporal peaks in the energy surface [4].

2.3. Fragment Decoding

The speech fragment decoder extends standard ASR decoding to account for the presence of competing sources. HMMs are trained on the spectro-temporal representation of the clean speech signal. The state-likelihoods are a function of the observed data and a foreground/background segmentation hypothesis. They are computed using missing data techniques. Potential segmentation hypotheses are derived by considering all possible combinations of foreground/background labelling for the set of coherent fragments. The full set of labellings can be evaluated in an efficient manner using a splitting and merging graph: When a new fragment starts all HMM states are duplicated and the ‘fragment is foreground’ interpretation is sent to one copy of the HMM, and the ‘fragment is background’ interpretation to the other. When a fragment ends, pairs of HMM states are remerged and the best scoring token of each pair is maintained. Crucially, independent foreground versus background decisions are made within each state of the HMM, so the labelling of the fragment is a function of the word string hypothesis. By coupling segmentation and recognition in this way, the segmentation exploits constraints that are implicit in the acoustic and language models. For example, in a digit recognition task, a fragment containing the unvoiced fricative energy of /s/ and an adjacent voiced fragment /evən/ may both be labelled as foreground because they match the word ‘seven’. However, if the fragment /s/ is followed by a fragment /i:/ from the competing speaker, then the fragments will be identified as not belonging together as they do not fit well to any digit model. Note, that in a different task with a different vocabulary, an /s/ fragment and an adjacent /i:/ fragment may been seen as making the word ‘see’. The language model influences the segmentation. A full description of the decoding algorithm is presented in [3].

3. Advances in the decoder framework

The system employed in the current work improves over that presented in Barker, Cooke and Ellis [3] in a number of ways. The most significant advances are described in the sections that follow.

3.1. Soft fragments

In the original SFD system each fragment labelling hypothesis is represented as a discrete missing data mask - spectro-temporal elements within background fragments are treated as wholly missing (0) and those within foreground fragments are wholly present (1). However, it has been shown that missing data systems perform better when using ‘soft’ masks [7]. The missing data mask is allowed to hold values between 0 and 1 (indicating a degree of ‘presentness’). The acoustic match score is then computed in a way that blends between the missing and present interpretations. The current SFD system converts each fragment labelling hypotheses into a soft masks. Within harmonic fragments, spectral-temporal points that have a single clear periodicity are given a fragment value close to 1 (i.e. these points are either clearly foreground or background). Points where there is evidence of multiple pitches are given a value closer to 0.5 (for details see [5]). Inharmonic fragments remain discrete. During decoding, a soft mask is constructed from the fragment values, x, by either using x directly if the fragment is being hypothesised as foreground, or by taking $1 - x$ if the fragment is hypothesised as background. Note, mask values close to 1 will alternate between 1 and 0 (present and missing), more ambiguous values close to 0.5 will alternate between being a little big greater than 0.5 (probably present) to a little bit less than 0.5 (probably missing).

3.2. Delta features

The current system extends [3] through the use of temporal difference features. Some care has to be taken to employ these feature effectively. Consider first the standard missing data approach: when the static features are judged to be present, delta features can also be computed and employed. For *missing* static features the probability calculation integrates over the range of energy that the missing feature could have had. However, because it is not possible to compute a meaningful bound on the value of unknown *delta* features, missing delta features are simply ignored. This causes no problems in missing data systems where a single missing data mask is evaluated. However, in the SFD system each segmentation hypothesis is represented by a different missing data mask, so hypotheses are compared in which different amounts of data are missing. If the probability computation includes terms for the present deltas but not the missing deltas then these hypotheses will not be comparable. This problem can be corrected by using the present deltas in a different way. In the current system if a delta feature is present, the likelihood of the corresponding static feature and the delta feature are averaged, i.e. in all probability calculations the term $p(x_i|q)$ is replaced with $\frac{1}{2}(p(x_i) + p(x'_i|q))$ where x_i is a static feature and x'_i is the corresponding delta, and q is the HMM model state. This is akin to treating the static feature and its delta as two observations of the same process rather than as treating them as independent observations. This treatment leaves the number of terms in the probability calculation unaffected by the number of present features in the missing data mask.

3.3. Speech prior

Early implementations of the SFD technique scaled the bounded marginal term computed for missing features by dividing by the observed energy. This gives the term a similar range to the observed data likelihood terms. It is shown in Barker, Cooke and Ellis [3] that it is the correct way to treat the statistics under the assumption that the missing features have a uniform prior distribution. However, a uniform distribution is a poor model of the speech log energy values used in the current work. Instead a GMM-based speech prior model is trained using clean speech (see Section 4 for details), and this model is used in the manner described in Section 2.3 of [3]. The use of an appropriate speech prior has been shown to considerably improve performance.

4. Experiments with simultaneous speech

4.1. The Grid data

Experiments were performed using simultaneous speaker data constructed from the Grid corpus [8] and in accordance with rules dictated by the Interspeech 2006 Speech Separation Challenge.[1] The Grid corpus consists of utterances of the form indicated in Table 1 spoken by 34 speakers. In the present study pairs of end-pointed utterances have been 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'. The task is to recognise the letter and digit spoken by the target speaker (i.e. by the person who says 'white'). A full description of the preparation of the two talker speaker data is presented in Cooke et al. (submitted) [9]. The test set has 600 utterance pairs at each TMR; 200 pairs in which target and masker are the same speaker, 200 pairs of the same gender (but different speakers), and 200 pairs of mixed gender.

Table 1: *Structure of the sentences in the GRID corpus.*

VERB	COLOUR	PREP.	LETTER	DIGIT	ADVERB
bin lay place set	blue green red white	at by on with	a-z (no 'w')	1-9 and zero	again now please soon

4.2. The recogniser configuration

A 64-channel log-scaled ratemap representation was employed (see Section 2.2). The 128-dimensional feature vector consisted of 64 log-energies and 64 delta log-energies terms. Speaker-dependent word-level HMMs were trained using 500 utterances from each of the 34 Grid speakers. Each word was 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. A GMM speech prior was employed (see Section 3.3). This 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 is 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 is decoded using the HMMs corresponding to the target speaker, ii) 'unknown speaker' - the utterance is decoded using HMMs corresponding to each of the 34 speakers and the overall best scoring hypothesis is selected (this can be implemented as an extended grammar in which 34 speaker dependent grammars are placed in parallel).

Adaptive beam-pruning was implemented to reduce the computational cost of decoding the 'unknown speaker' configuration. The beam width was adapted in such a way as to prune a fixed percentage of the partial hypotheses at each frame. This percentage was tuned using a small development set (150 mixtures at 0 dB). It was found that 90% of the hypotheses could be pruned with little impact on the recognition result, and with a resulting reduction in decoding time of over 75%.

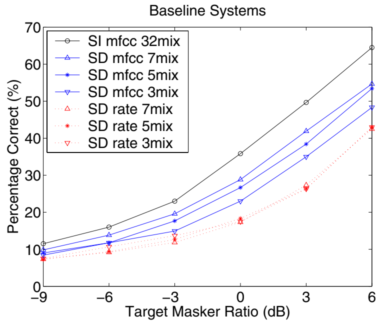

Figure 2: *Results for the baseline system using speaker dependent (SD) or speaker independent (SI) models trained on either ratemap or MFCC features.*

Results have also been obtained for a conventional HMM system using models with an identical topology trained on either the ratemap representation or 13 MFCC features along with their deltas and accelerations. This baseline has been tested using speaker dependent models employing up to 7 Gaussian mixtures per state and running in the known speaker configuration. It has also been tested using a single set of speaker independent (SI) HMMs employing 32 mixtures per state and trained using the combined training data from all 34 Grid speakers.

4.3. Results

Figure 2 shows results for the baseline system. Performance scores (throughout the paper) indicate the percentage of the total number of *letter* and *digit* tokens that were recognised correctly. MFCCs outperform the ratemap representation. HMMs trained on all speakers provide better performance than a speaker dependent HMM matched to the target speaker, i.e. the speaker dependent HMMs were less tolerant to noise. However, the performance of all systems degrades quickly as the TMR decreases. At 0dB the recognition score is 32%, and at -9 dB performance is down to chance levels, 7%.

Recognition results for the SFD system are shown in Figure 3 (and Table 2), plotted alongside results from a group of listeners [9], and the best conventional system from Figure 2. The SFD clearly outperforms the baseline across all TMRs and across all mixture conditions. Unlike the conventional system the SFD is able to exploit knowledge of the target speaker identity (compare the 'known' and 'unknown' speaker curves). Prior knowledge of the speaker identity only fails to confer an advantage in the 'same speaker' condition. Although no recogniser performance matches that of listeners, the shape of the listener data is matched remarkably closely by the SFD in the 'unknown' speaker condition. The characteristic pattern in which performance *increases* as the TMR *decreases* below about -3 dB, occurs because the negative effects of increased *energetic masking* are more than offset by decreased *informational masking* [10]. Informational masking is at a peak at around 0 dB where fragments of the target and masker are most confusable. As the TMR moves away from 0 dB the fragments become more easily separable on the basis of their level. For the unknown speaker configuration the extra errors at 0 dB, where target and masker are most confusable, result when hypotheses pass-

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

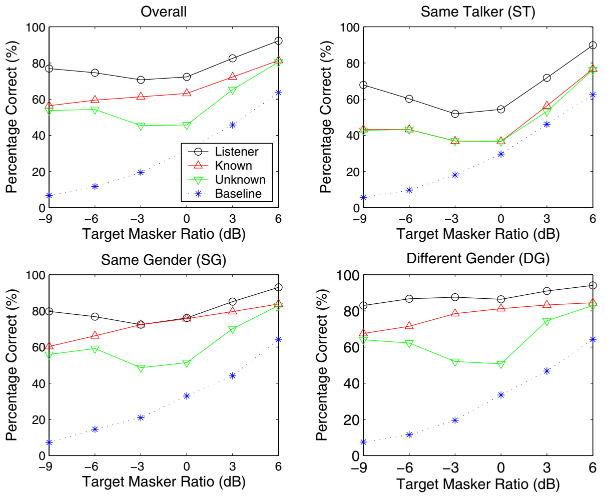

Figure 3: *Results for the speech fragment decoder in known speaker and unknown speaker configurations compared against the baseline system and average listener results.*

ing through the HMM for the masker wins over 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. An example of a typical decoding is shown in Figure 4.

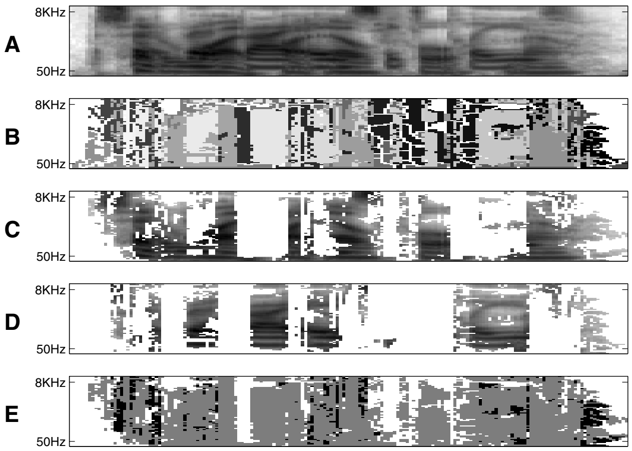

Figure 4: ***(A)*** *Ratemap representation of the mixture ‘set white by l 4 now’ (target) plus ‘set red by a six please’ (masker) at 0 dB.* ***(B)*** *The set of fragments where each is represented using a different shade of grey.* ***(C)*** *The regions assigned to the foreground in the ‘unknown speaker’ configuration. The target was recognised correctly.* ***(D)*** *The regions assigned to the foreground if decoder is given models of just the masking speaker. 5 out of 6 words in the masker utterance are now recognised correctly.* ***(E)*** *Demonstrating the complementarity of the the target and masker decodings: the grey regions occur in only* ***C*** *or* ***D****; the small black regions occur in both* ***C*** *and* ***D****.*

Table 2: *Result for unknown speaker configuration (%).*

	-9 dB	-6 dB	-3 dB	0 dB	3 dB	6 dB
Overall	53.8	54.3	45.4	45.8	65.3	80.5
ST	42.8	43.2	36.9	36.7	53.2	76.2
SG	55.9	59.2	48.6	51.4	70.11	83.0
DG	64.0	62.3	52.0	50.8	74.5	83.0

5. Conclusions

The paper has described a novel approach to robust ASR which works by coupling the problems of foreground/background segregation and speech recognition. Whereas most robust ASR technique 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, 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 same talker and same gender conditions. Future work will aim to develop a statistical model of primitive sequential grouping that will weight segmentation hypotheses according to continuity of primitive properties across fragments through time.

6. References

[1] P. Lockwood and J. Boudy, “Experiments with nonlinear spectral subtractor, Hidden Markov Models and the projection for robust speech recognition in cars,” *Speech Communication*, vol. 11, 1992.

[2] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–588, 1994.

[3] J.P. Barker, M.P. Cooke, and D.P.W. Ellis, “Decoding speech in the presence of other sources,” *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[4] A. Coy and J. Barker, “Recognising speech in the presence of a competing speaker using a ‘speech fragment decoder’,” in *Proc. ICASSP 2005*, Philidelphia, 2005, pp. 425–428.

[5] N. Ma, P. Green, and A. Coy, “Exploiting dendritic autocorrelogram structure.,” in *Proc. Interspeech 2006*, submitted.

[6] A. Coy and J. Barker, “A multipitch tracker for monaural speech segmentation,” in *Proc. Interspeech 2006*, submitted.

[7] 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 2000*, Beijing, 2000, vol. 1, pp. 373–376.

[8] M.P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *JASA*, submitted.

[9] M.P. Cooke, M.L. Garcia Lecumberri, and J. Barker, “The non-native cocktail party,” in preparation.

[10] D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott, “Informational and energetic masking effects in the perception of multiple simultaneous talkers,” *JASA*, vol. 100, pp. 2527–2538, 2001.