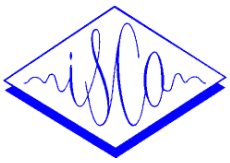

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

DECODING SPEECH IN THE PRESENCE OF OTHER SOUND SOURCES

Jon Barker, Martin Cooke

Department of Computer Science
University of Sheffield, UK
j.barker, m.cooke@dcs.shef.ac.uk

Dan Ellis

International Computer Science Insitute
Berkeley CA USA
dpwe@icsi.berkeley.edu

ABSTRACT

Conventional speech recognition is notoriously vulnerable to additive noise, and even the best compensation methods are defeated if the noise is nonstationary. To address this problem, we propose a new integration of bottom-up techniques to identify 'coherent fragments' of spectro-temporal energy (based on local features), with the top-down hypothesis search of conventional speech recognition, extended to search also across possible assignments of each fragment as speech or interference. Initial tests demonstrate the feasibility of this approach, and achieve a reduction in word error rate of more than 25% relative at 5 dB SNR over stationary noise missing data recognition.

1. INTRODUCTION

Recognition of speech in its natural, noisy, setting remains an important unsolved problem in a world increasingly dominated by mobile communication devices. While techniques for ameliorating the effects of stationary or slowly-changing acoustic backgrounds have been partially successful – albeit still some way short of human performance – little progress has been made towards handling nonstationary sources.

Approaches to the latter problem fall into two broad categories. *Bottom-up* (BU) techniques are based essentially on exploiting statistical regularities possessed by components emanating from a common sound source. The different perspectives of primitive computational auditory scene analysis (see review in [3]) and blind source separation/independent component analysis [2] fall into this category, as do mainstream signal processing approaches such as [5]. *Top-down* (TD) approaches utilise models of acoustic sources and attempt to find combinations which jointly explain the observation sequence. HMM decomposition[8] and parallel model combination (PMC) [6] are the prime examples of the model-based approach.

Neither bottom-up nor top-down approaches have been particularly successful at tackling real-world acoustic mixtures. BU algorithms, such as grouping by common fundamental frequency [5], tend to produce reasonable local results but fail to deliver complete separation. TD systems work well, but only when adequate models for all sound sources present exist, and when the number of sources is small (typically 2) and known in advance.

In this paper, we show how BU and TD approaches can be combined to exploit the locally reasonable behaviour of BU techniques and the globally-consistent decoding abilities of TD systems. Rather than relying on perfect BU organization, better BU performance helps by leading to reduced TD search. Our approach also does away with the requirement that prior models exist for all sources. At the heart of the system is a multi-source Viterbi decoder which pieces together a subset of evidence fragments delivered by BU processes. The next section describes the architecture of the multi-source decoder. Section 3 presents example behaviour on several artificial noise intrusions and recognition results on a noisy digit sequence recognition task. Section 4 discusses performance refinements and theoretical foundations.

This work was supported by the EU LTR project RESPITE (28149).

2. THE MULTI-SOURCE DECODER

Conventional speech recognition divides into two main pieces: an acoustic model, estimating the probability that observed features correspond to certain speech classes, and a hidden Markov model (HMM) decoder, which searches for a word-sequence hypothesis matching a highly likely sequence of speech-class states.

In contrast to conventional decoding, where all the observations are assumed to belong to the source being recognized, the task of a multi-source (MS) decoder is to determine the most likely model state sequence at the same time as deciding *which* observations to use, and which to ignore as 'background'. We assume that we have models for the speech source, but in contrast with approaches such as PMC and HMM decomposition we do not require models for the acoustic background.

The input to the MS decoder is a set of *coherent source fragments*. Such fragments consist of parameters such as energy estimates in some arbitrary time-frequency region. These fragments have been marked as belonging to a single source by an earlier BU process. Due to speech energy dynamics, it is feasible to find regions with favourable local SNR even if the global SNR is low. Example BU processes include forming time-frequency elements into tracks, or tracks into groups of harmonics with a common fundamental, or exploiting common onset or location in space. In this study we use extremely simple BU processing as described in section 3.

Given a set of source fragments, MS decoding is based on two key ideas. First is the ability from missing data recognition [4, 1] to evaluate the match of a speech model to an observation whose elements are variously tagged as 'present' or 'missing' (i.e. masked behind an interfering sound source). The second idea is that this missing-data match can be compared across alternative possible interpretations of which data is indeed valid; finding the best match should establish both the correct word sequence and the optimal present/missing labelling.

2.1. Decoding evidence fragments

One approach to recognising speech from a set of evidence fragments is to evaluate every possible combination of fragments over an entire utterance. Unfortunately, there are 2^N subsets of N fragments, and N could typically become rather large. An alternative approach is to merge decoder hypotheses every time a fragment ends. The complexity then reduces to 2^M where M is the maximum number of *simultaneous* fragments. This is tractable if BU processes deliver evidence fragments above some minimum granularity, say over some tens of milliseconds duration. Crucially, although N increases with utterance length, M remains essentially constant. Based on the examples in section 3, N can exceed 40 even for short utterances, while M rarely exceeds 6.

The resulting decoder is based on the standard token-passing Viterbi algorithm with the following modifications:

- 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*.
- 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.
- When a fragment ends we compare, for each state, pairs of tokens that differ only in the label of the fragment that is ending. The less likely token is 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.

The scheme may also be seen as a parallel set of normal Viterbi decoders (i.e. with one token for each state) but when a new fragment starts each decoder is duplicated, and when a fragment ends pairs of decoders are merged.

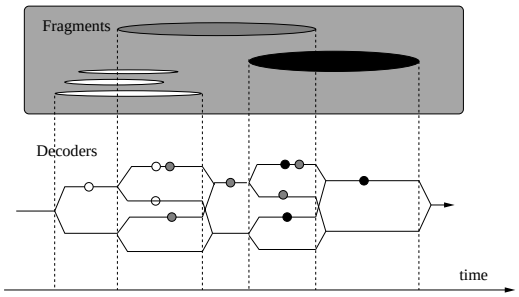

Figure 1: The evolution of a set of parallel decoders. Each parallel path represents a separate decoder, with the shaded dots indicating which ongoing fragments are being considered as speech.

Figure 1 illustrates the evolution of a set of parallel decoders 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, the decoder is duplicated. In one decoder all tokens are assigned the "white is speech" labelling and in the other they are assigned "white is background". When the grey fragment starts the decoders are again duplicated, each pair covering both possible labellings for the grey fragment. When the white fragment ends, pairs of decoders are merged if their labelling only differs for the white fragment, and so on until the end of the utterance. Note that there are at most 4 active decoders, not the 8 required to decode every possible subset of 3 fragments.

2.2. The merging problem

When a fragment ends and decoders are merged, tokens from each decoder are paired up and their likelihoods are compared. However, there is a problem inherent in this comparison: these tokens have arisen from decoders with different speech/background labellings, and as such are calculating missing-data fits based on different patterns of present and missing data. The missing data framework treats the two types of data somewhat differently: the match score for present data is the likelihood calculated by marginalising the full model probability density function (pdf) over the missing features. However, for the missing data we calculate the 'bounded' probability of the speech being less energetic than the observed background - a true probability rather than a likelihood, and as such not directly comparable. This difficulty has been overcome in previous missing data work, where the amount of present and missing data is the same for each competing hypothesis, by the simple expedient of a scaling factor. However, when comparing decoder hypotheses with differing foreground/-background interpretations, a better solution is required.

As one solution, the results in section 3 scale the missing data probabilities by dividing them by the integration range over which they were computed e.g. if the observed value of the background is X, then the speech energy is assumed to lie between 0 and X, and the model probability is computed by integrating the pdf between 0 and X. An 'average likelihood' (i.e. the average value of the pdf over the integration range) is formed by dividing the probability by X, something comparable (when scaled by a fixed constant) with the likelihood values constituting the match score for present data.

Further discussion of this issue along with a more principled solution are presented in section 4.1.

3. EXPERIMENTS

Experiments to test the new decoder architecture build on previous missing data work on robust recognition of connected digits. The current experiments evaluate the new decoding algorithm while using a naive technique to perform the dissection of the spectrogram. This establishes a baseline against which to compare future work that will employ more sophisticated auditory scene analysis techniques.

The system employed in these experiments has the following steps:

1. The first 10 frames (assumed to contain only noise) are averaged to estimate a stationary noise spectrum.
2. The noise spectrum is used to estimate the local SNR.
3. A present data mask is made by thresholding the local SNR estimate at some minimum SNR.

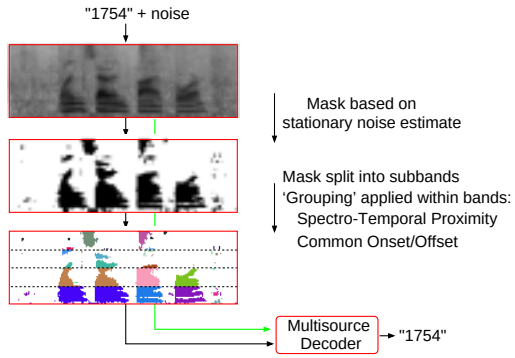

Figure 2: An overview of the multi-source recognition system.

(These 3 steps match the standard missing data approach.)

4. The present data mask is dissected by first dividing it into four frequency bands and then labelling contiguous regions within each subband as the separate fragments.
5. The set of fragments and the noisy speech representation are passed to the MS decoder.
6. Spectro-temporal regions that are not contained in any fragment are assigned a fixed *background* label.

If the actual noise is non-stationary the noise spectrum estimates, and hence the local SNR estimates, are often 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 spurious region in the present data mask, usually causing poor recognition. In the new approach, the MS decoder should reject these fragments and label them as background, thereby producing a better recognition hypothesis. This effect is illustrated in figure 3, 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 MS decoder. The final panel shows a *backtrace* of the fragments that the MS 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 4 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.

The examples discussed so far are artificial and the non-speech intrusions in the data mask are very distinct. To test the technique on real noise, TIDigits utterances [7] were mixed with NOISEX factory noise [9] 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 recognizers.

Recognition was performed on a 240 utterance test set. The missing data systems were based on a 24 channel filter bank representation and 8 state, 10 mixture HMMs (as described in [4]). The results in figure 5 compare standard missing data with the new

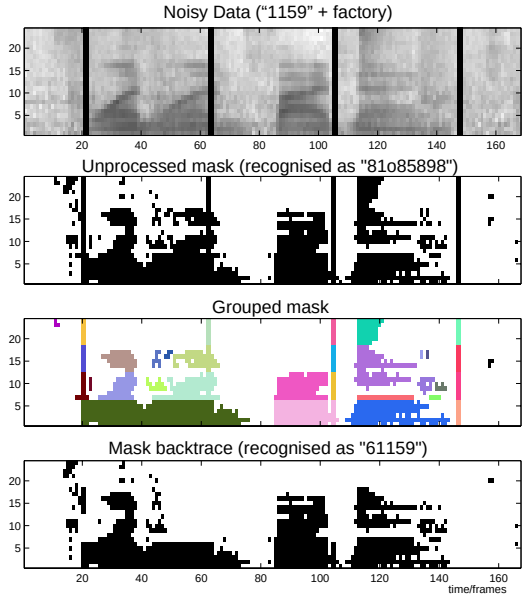

Figure 3: An example of the multi-source system performance when applied to data corrupted by artificial transients (see text).

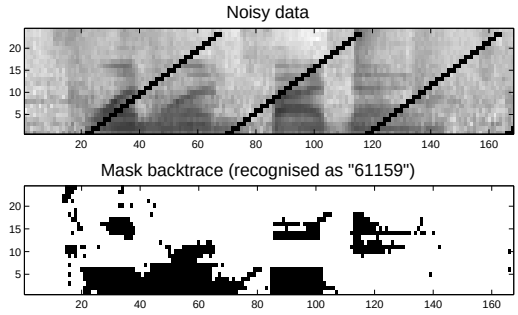

Figure 4: Another example of the multi-source decoding for data corrupted with artificial chirps.

multi-source technique. The scaling constant required to balance missing and present data scores as described in 2.2 was optimally tuned, but a single value sufficed for all noise levels.

It can be seen that multi-source decoding provides a significant improvement at the lower SNRs, e.g. at 5db 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 multi-source technique is providing an improvement over a missing data system that is already robust by the standards of traditional techniques.

4. DISCUSSION

The results in figure 5 labelled "a priori" show the performance achieved using missing data techniques if prior knowledge of the

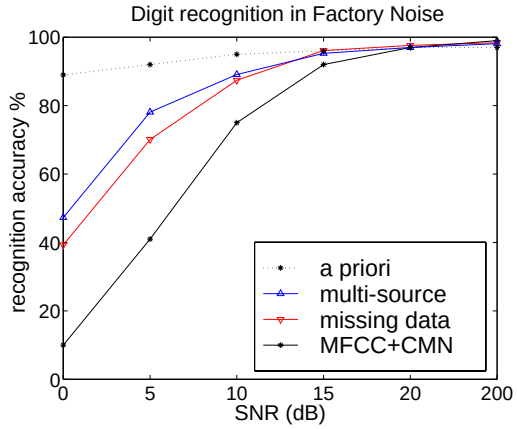

Figure 5: Recognition results for a baseline MFCC system, a missing data system, and the multi-source system.

noise is used to create a perfect local SNR mask. Even using the multi-source technique results fall far short of this upper limit as the noise level rises above 10dB SNR.

One possible cause of this this significant performance gap is that the fragments supplied to the multi-source 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 adjoining spectro temporal regions this technique will not be able to separate them. This is evident is figures 3 and 4 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.

4.1. Posterior probability formulation

Traditional speech recognition is formalized as a search for the most likely model state sequence Q^* (and hence wordstring) given the observed data X i.e.

$$Q^* = \underset{Q}{\operatorname{argmax}}\ p(Q|X)$$

The term being maximised is rearranged through Bayes' rule to be $p(X|Q)p(Q)$, where the data prior $p(X)$ is ignored since it does not depend on Q. In missing data recognition, earlier processing supplies a mask, and we maximise

$$p(Q|mask, X)$$

As discussed above, the subsidiary calculation of $p(X|Q, mask)$ presents problems when combining model likelihoods for present data with bounded integrals for the missing dimensions. This can be avoided, at some compuational expense, by evaluating the full posterior above including the data prior $p(X|mask) = \sum_Q p(X|Q, mask)p(Q)$.

The multi-source approach is different in that maximisation occurs over the mask too (constrained by a-priori BU fragment assignment), so we are now finding the Q that maximises

$$p(Q, mask|X)$$

This can be expanded by Bayes' rule into:

$$p(Q|mask, X)p(mask|X)$$

i.e. the missing data term plus a mask-specific weighting independent of the state sequence. This term could represent a prior on different mask patterns (i.e. $p(mask)$), perhaps reflecting BU 'good continuation' rules applied to the underlying fragments. The dependence on X suggests an integration of fragment formation into the search – perhaps weighting alternate fragmentations by their likelihoods.

5. CONCLUSIONS

We have presented an approach to exploit bottom-up organization within the top-down hypothesis-search framework of conventional speech recognition. This approach, in conjunction with missing data techniques, allows speech recognition that also searches across possible interpretations of fragments as speech or background. Initial experiments, based on crude BU techniques and using a partially-heuristic probabilistic formulation, show the approach as tractable and able to offer significant improvements in high-noise situations over static missing data recognition. Future work to integrate more sophisticated BU algorithms and a rigorous probabilistic evaluation holds great promise.

6. REFERENCES

[1] 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. to appear.

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

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

[4] M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*. Accepted for publication.

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

[6] 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.

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

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

[9] 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.