

COMBINING BOTTOM-UP AND TOP-DOWN CONSTRAINTS FOR ROBUST ASR: THE MULTISOURCE DECODER

Jon Barker, Martin Cooke

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

Dan Ellis

Department of Electrical Engineering
Columbia University, NY, USA
dpwe@ee.columbia.edu

ABSTRACT

Recognising speech in the presence of non-stationary noise presents a great challenge. Missing data techniques allow recognition based on a subset of features which reflect the speech and not the interference, but identifying these valid features is difficult. Rather than relying only on low-level signal features to locate the target (such as energy relative to an estimated noise floor), we can also employ the top-down constraints of the speech models to eliminate candidate target fragments that have a low likelihood of resembling the training set. The multisource decoder makes a simultaneous search in fragment-labelling space (target or interference) and word-string space, to find the most likely overall solution. When testing on the Aurora 2 task, this algorithm achieves up to 20% relative word error rate reduction in nonstationary noise conditions at low SNR.

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, little progress has been made towards handling non-stationary noise.

There are two broad categories of approaches to dealing with interference for which a stationarity assumption is inadequate: *Bottom-up* (BU) techniques exploit common characteristics to identify components emanating from a single source. Primitive computational auditory scene analysis (see review in [7]) and blind source separation/independent component analysis [4] fall into this category, as do mainstream signal processing approaches such as [9]. *Top-down* (TD) approaches utilise models of acoustic sources to find combinations which jointly explain the observation sequence. HMM decomposition [12] and parallel model combination (PMC) [10] are the prime examples of top-down approaches.

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 [9], 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 and known in advance.

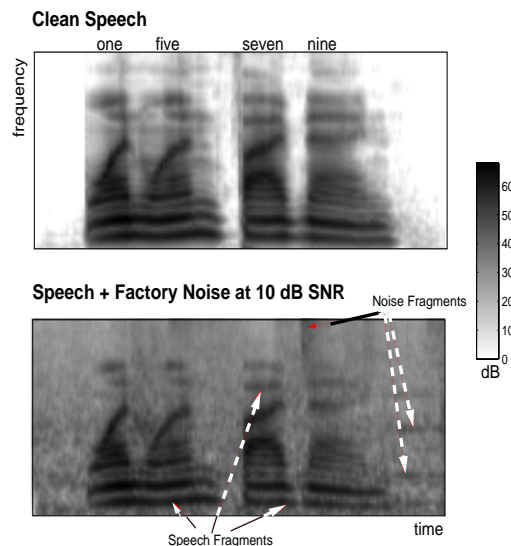


Figure 1: When noise is added to speech, low energy regions are masked, but prominent features, such as formants and resolved harmonics, may be largely unaffected.

2. THE MULTISOURCE DECODER

Consider the effect of additive noise on a time-frequency representation of speech. Figure 1 compares a speech utterance recorded in noise-free conditions with that of the same utterance with noise added at an average level of 10 dB SNR. Although the noise masks much of the speech signal, there are local regions corresponding to prominent features of the speech which are little affected by the noise. Identifying these regions would give a *partial* description of the underlying speech. Missing data techniques can be employed to recognise speech from these partial descriptions—experiments have shown that perfect identification of the reliable speech regions recognition systems restores performance close to noise-free levels [8]. However, reliably identifying the speech-dominated regions is a challenging problem, especially when the noise includes prominent energy that is easily confused with speech (e.g. as marked in the lower panel of Figure 1).

Basic missing data recognition consists of two separate steps performed in sequence: first a ‘present-data’ mask is calculated, based, for instance, on estimates of the background noise level.

This research is supported by the EC ESPRIT long term research project RESPITE. Thanks to Ljubomir Josifovski who provided the implementation for the adaptive noise estimation.

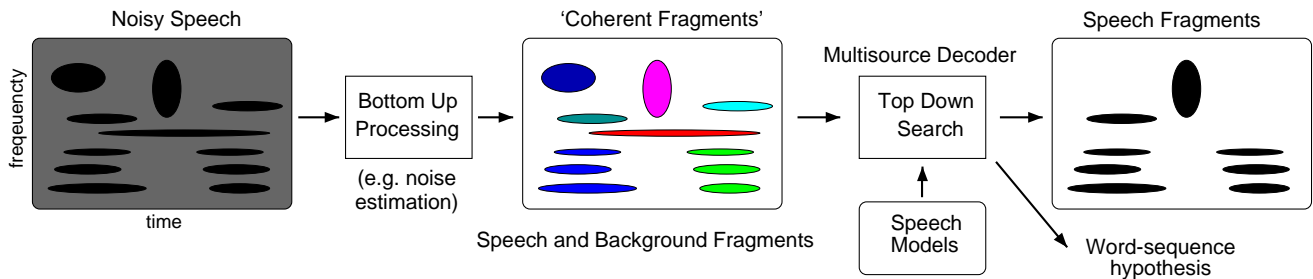


Figure 2: Bottom-up processes are employed to locate ‘coherent fragments’ (regions of representation that are due entirely to one source) and then a top-down search with access to speech models is used to search for the most likely combination of fragment labelling and speech model sequence.

Second, missing data recognition is performed by searching for the most likely speech model sequence consistent with this evidence. By contrast, the multisource decoding approach integrates these two steps, so that the search includes building the present-data mask to find the subset of features most likely to correspond to a single voice, as well as the corresponding word sequence. However, to search across all possible time-frequency masks would be computationally prohibitive. Instead, search is limited to combinations of larger time-frequency *fragments*, generated by a first stage of BU processing. The entire multisource decoding process is illustrated in Figure 2.

It is preferable that the BU processes dissect the time-frequency plane into a small number of larger pieces, rather than many small ones. A representation consisting of N fragments presents the multisource decoder with a search space of 2^N possible assignments; hence, offering fewer (and hence larger) fragments results in a more efficient search.

At the same time, it must be possible to find a set of fragments that distinguishes between the true speech energy and other interference. If a fragment contains elements of both the speech and noise, then the decoder will be forced to either reject good speech data, or to introduce noise into its speech source hypothesis. The goal of the BU processes is to generate ‘coherent fragments’ – local regions of the spectro-temporal representation that are dominated by just one source.

In [1], we presented a dynamic-programming algorithm, which, by combining equivalent hypotheses at the conclusion of each fragment, reduces the search complexity to 2^M , where M is the maximum number of *simultaneous* fragments. Crucially, although N increases with utterance length, M remains essentially constant.

3. FRAGMENT GENERATION

As explained above, the performance of the multisource decoder, both in the sense of computational cost and in terms of its ability to separate speech from other sounds, is strongly dependent on the BU mechanisms that generate the initial fragments. Our current system first estimates and excludes stationary ‘background’ noise through adaptive noise estimation, then attempts further division of the remaining energy to improve fragment coherency.

3.1. Adaptive noise estimation

The per-channel mean and variance of the slowly-varying component of the background noise are estimated from the first 10 frames of data, then updated with subsequent frames which are identified as noise-dominated.

For each time-frequency ‘pixel’, $P(SNR > 0)$ is calculated, i.e. the probability that the data is *not* masked by the stationary noise background. If $P(SNR > 0)$ is low (i.e. < 0.5) the pixel is attributed to the background. The remaining pixels whose energy indicates they are above the background may be due to either i) speech or ii) a high-energy nonstationary noise component. The correct label for these points will be found by the top-down multisource decoder search.

Figure 4 (A) shows the spectrogram of the utterance “seven five”, to which a stationary background noise and a series of broadband high-energy noise bursts have been added. Adaptive noise estimation identifies the stationary component, leaving the unmasked speech energy and the nonstationary noise bursts as candidate ‘present data’, as shown in panel C. This however must be broken up into a set of fragments to permit searching by the multisource decoder.

In order to confirm that the top-down process in the decoder is able to identify the valid speech fragments, we may examine its performance given a small set of ideal coherent fragments. These can be generated by applying *a priori* knowledge of the clean speech, i.e. comparing the clean and noisy spectrograms to mark out regions where either the speech or the noise dominate. Panel D of Figure 4 shows the foreground mask from the noise estimator divided up this way. Given these fragments, the decoder is able to correctly recognise the utterance as “seven five”, using the fragments in panel E as evidence of the speech. The correct speech/noise fragment labelling is shown in panel F. Comparing E and F, it can be seen that the decoder has accepted all the speech fragments, while correctly rejecting all the larger fragments of noise. (Some small noise regions have been included in the speech, implying their level was consistent with the models.)

In practice we need to generate a set of fragments similar to those shown in panel D without knowledge of the clean signal. The current work takes a very basic approach: we start with the regions that are unaccounted for by the background noise model (e.g Figure 4 panel C) and form a separate fragment from each contiguous region. These fragments are then further divided in an attempt to split harmonic energy from inharmonic energy: Fragments are cut into separate pieces when they cross temporal boundaries marked by a voicing detector which is based on the height of the first peak in the summary autocorrelogram (see [2] for details of this technique).

3.2. Multisource decoding with soft decisions

Better missing data ASR results are obtained by softening the speech/noise decisions and assigning pixels with a *probability* of being speech rather than a binary speech/noise label [3]. Although the multisource decoder makes hard speech/noise assignments at the level

of fragments, the current system uses the actual $P(SNR > 0)$ probabilities *within* each fragment when calculating the likelihood of matches to the clean speech models.

4. EXPERIMENTS EMPLOYING THE AURORA 2 CONNECTED DIGIT TASK

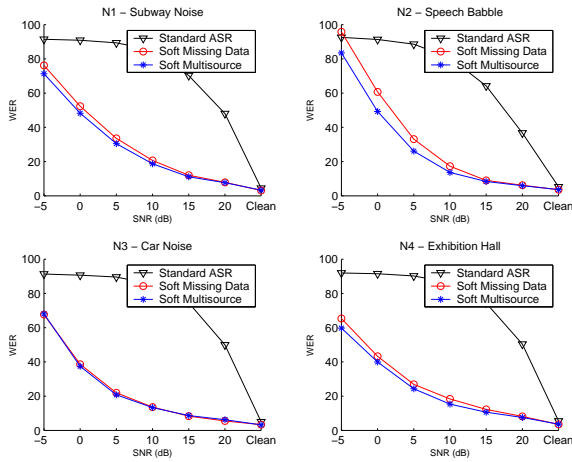


Figure 3: Results for the Aurora Test Set A (see text)

The system has been tested using the Aurora 2.0 speaker independent connected digit recognition task [11]. Acoustic vectors were obtained via a 32 channel auditory filter bank [6] with centre frequencies spaced linearly in ERB-rate from 50 to 3750 Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a first order filter with an 8 ms time constant, and sampled at a frame-rate of 10 ms. Finally, cube-root compression was applied to the energy values. Whole-word digit models were trained from the Aurora clean speech training set using a 16-state feed-forward topology, with seven diagonal Gaussians per state.

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

Results for the four noise conditions in the Aurora test set A are shown in Figure 3.¹ For three of the four noise conditions the multisource processing achieves a better performance than the standard missing data system. For the highly non-stationary speech babble noise the performance improvements at low SNRs are fairly large. The only noise for which no improvement is seen is the car noise (N3). Examination of the noises shows that the car noise is the most stationary of the four and is well modelled by the adaptive noise estimate. It is therefore not surprising that for this noise type the multisource decoding technique, which is designed to deal with non-stationary noise events, can do little to improve over the strong performance of the standard missing data technique.

5. DISCUSSION

Computational complexity

In the Aurora experiments, the number of fragments per utterance often exceeded 100. However, as illustrated in Figure 4 (G), the

¹The results presented here are for systems that are not employing temporal difference features and hence the baseline is somewhat lower than similar results published in previous papers e.g. [3]

maximum number of simultaneous fragments was never greater than 10 and the average number of hypotheses per frame computed over the full test set was less than 4. Although the decoder is evaluating on average roughly four times as many hypothesis as a standard missing data decoder, much of the probability calculation may be shared between hypotheses and hence the computational load is increased by a factor much smaller than four.

Fragment label priors

Study of the fragments generated in our current system shows that most of the noise has been successfully accounted for by the adaptive noise estimate, and the remaining regions which form the fragments are mainly speech. In the current system we model this prior knowledge by biasing the decoder towards favouring hypotheses in which fragments are labelled as speech (i.e hypotheses in which fragments are labelled as noise will only win where the data is a very poor fit to the speech models).

Primitive sequential and simultaneous grouping effects may be modelled in a similar fashion. For example, groups which onset at the same time are likely to come from the same source and therefore the decoder should favour hypotheses in which such fragments share the same label. Statistics for modelling these effects could potentially be learnt from *a priori* masks.

Three-way labelling of time-frequency cells

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

Decoding multiple sources

This technique is named ‘multisource decoding’, yet in the current incarnation we are only recognising a single source, the most likely fit to the speech models. A natural future extension would be to search for fits across multiple simultaneous models, possibly permitting the recognition of both voices in simultaneous speech. This again resembles the ideas of HMM decomposition [12, 10], but because each ‘coherent fragment’ is assumed to correspond to only a single source, the likelihood evaluation is greatly simplified. The arguments about the relationship between large, coherent fragments and search efficiency remain unchanged.

Improvements to fragment generation

The fragments in the current system rely on a very simple and crude model - mainly that energy below an estimate ‘noise floor’ is to be ignored, and the remainder can be divided up according to some simple heuristics. It is likely that more powerful fragmentation will result in significant improvement gains for the technique. For instance, within the regions currently marked as ‘voiced’, subband periodicity measures could indicate whether frequency channels appear to be excited by a single voice, or whether multiple pitches

suggest the division of the spectrum into multiple voices (as in [5]). Sudden increases in energy within a single fragment should also precipitate a division, on the basis that this is strong evidence of a new sound source appearing.

6. CONCLUSION

We have presented a two stage technique for recognising speech in the presence of other sound sources: i) a bottom up processing stage is employed to produce a set of source fragments, ii) a top-down search which, given models of clean speech, uses missing data recognition techniques to most likely combination of source speech/background labelling and speech model sequence. Preliminary ASR experiments show that the system can produce recognition performance improvements even with simple bottom-up processing. We believe that through the application of more sophisticated CASA-style bottom-up processing we will be able to improve the quality of the fragments fed to the top-down search and further improve the performance of the system.

7. REFERENCES

- [1] J.P. Barker, M.P. Cooke, and D.P.W. Ellis. Decoding speech in the presence of other sound sources. In *Proc. ICSLP '00*, Beijing, China, October 2000.
- [2] J.P. Barker, P. Green, and M.P. Cooke. Linking auditory scene analysis and robust ASR by missing data techniques. In *Proc. WISP '01*, Stratford-upon-Avon, UK, 2001.
- [3] 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.
- [4] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1004–1034, 1995.
- [5] G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336, 1994.
- [6] M. P. Cooke. *Modelling auditory processing and organisation*. PhD thesis, Department of Computer Science, University of Sheffield, 1991.
- [7] 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.
- [8] M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, jun 2001.
- [9] P.N. Denbigh and J. Zhao. Pitch extraction and separation of overlapping speech. *Speech Communication*, 11:119–125, 1992.
- [10] M. J. F. Gales and S. J. Young. HMM recognition in noise using parallel model combination. In *Eurospeech'93*, volume 2, 837–840, 1993.
- [11] D. Pearce and H.-G. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP '00*, volume 4, 29–32, Beijing, China, October 2000.
- [12] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *ICASSP'90*, 845–848, 1990.

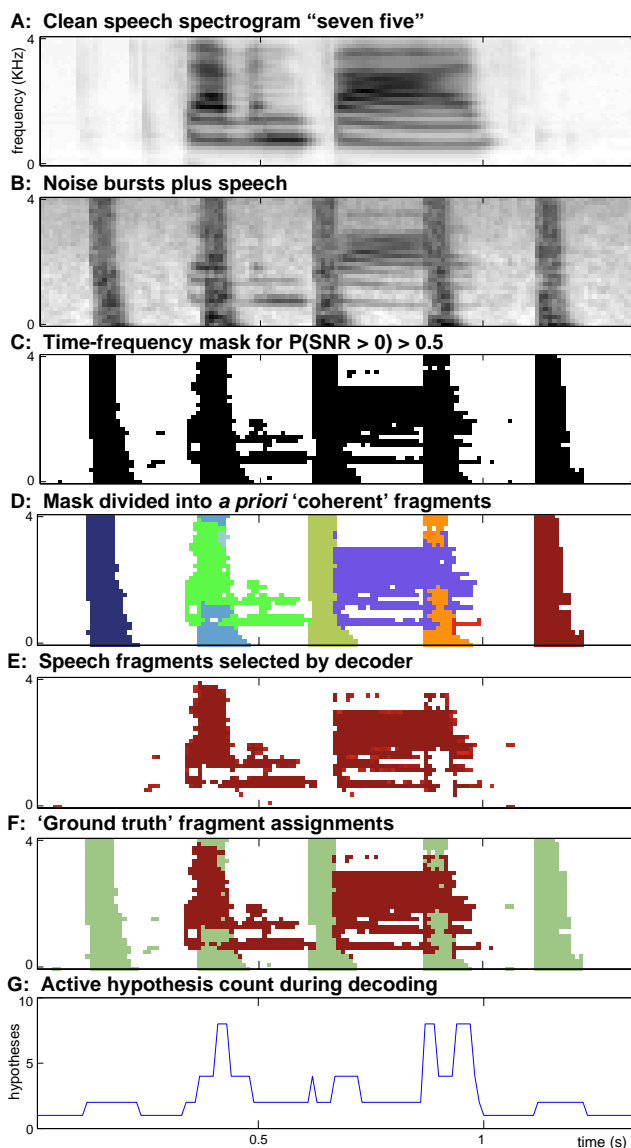


Figure 4: Panel A shows a spectrogram of the utterance “seven five”. Panel B shows the same signal but after adding a two state noise source. Panel C shows the components of the mixture that are not accounted for by the adaptive background noise model. Panel D displays a test set of perfectly coherent fragments generated using *a priori* knowledge of the clean signal. Panel E shows the groups that the multisource decoder identifies as being speech groups. The correct assignment is shown in panel F. Panel G plots the number of grouping hypotheses that are being considered at each time frame.