AUDITORY SCENE ANALYSIS AND HIDDEN MARKOV MODEL RECOGNITION OF SPEECH IN NOISE

P.D. Green, M.P. Cooke and M.D. Crawford
Speech and Hearing Research Group
Department of Computer Science, University of Sheffield, Sheffield S10 2TN, UK
{p.green,m.cooke,m.crawford}@dcs.shef.ac.uk

ABSTRACT

We describe a novel paradigm for automatic speech recognition in noisy environments in which an initial stage of auditory scene analysis separates out the evidence for the speech to be recognised from the evidence for other sounds. In general, this evidence will be incomplete, since intruding sound sources will dominate some spectro-temporal regions. We generalise continuous-density hidden Markov model recognition to this 'occluded speech' case. The technique is based on estimating the probability that a Gaussian mixture density distribution for an auditory firing rate map will generate an observation such that the separated components are at their observed values and the remaining components are not greater than their values in the acoustic mixture. Experiments on isolated digit recognition in noise demonstrate the potential of the new approach to yield performance comparable to that of listeners.

1. AUDITORY SCENE ANALYSIS AS A PREPROCESSOR FOR SPEECH RECOGNITION

Auditory scene analysis (ASA) describes the process by which listeners separate out and pay selective attention to individual sound sources within the mixture which reaches their ears [1]. Recent work at Sheffield [2,3] and elsewhere [4,5,6] has achieved some success in computational modelling of ASA based on grouping principles such as common onset, periodicity and good continuation of source components. If ASA depends on these unconditional, primitive processes, they may be viewed as a natural preprocessing stage for ASR. In contrast to most schemes for robust ASR (see Grenie & Junqua [7] for a review), this suggestion has the advantage that it does not require a model of the noise. Furthermore, there need be no assumption about how many sound sources are present, and the set of active sources may change with time.

Fig. 1 presents quantitative results from previous segregation studies (Cooke & Brown [8]) in terms of two metrics – SNR and characterisation. The latter measures the percentage of the speech signal recovered from a mixture. This figure illustrates that whilst we achieve significant SNR improvements in each case, current auditory scene analysis algorithms typically recover rather less than 40% of the energy associated with a target source. This is not surprising, since we proceed on the basis of finding reasons to group components. Some time-frequency regions will be masked to such an extent that purely data-driven grouping is unlikely to recruit them. We describe such data as *occluded speech*, although the analogy with visual occlusion should not be taken too far.

The work presented here explores the possibility that occluded speech might contain sufficient information for recognition, and proposes a two-stage approach to robust ASR: signal separation by auditory scene analysis followed by recognition of the (incomplete) segregated data. The main problem addressed is the modification of ASR techniques to handle such data. In [9] we showed that a straightforward adaptation of Kohonen nets maintains an encouragingly robust performance in a frame-by-frame phone labelling task when an increasing proportion of the input vector is unavailable (e.g. no significant deterioration up to 90% random removal using a filterbank representation). Such nets can also be *trained* on partial data. In this paper we report on recognition of occluded speech by hidden Markov models (HMMs). Section 2 describes modifications to the HMM probability computation for incomplete observation vectors. Section 3 demonstrates the results of an experiment in which *simulated* auditory scene analysis provides data for the modified Viterbi algorithm, comparing the results on digit recognition in multispeaker babble with listeners' performance. Section 4 extends the approach by exploiting an auditory induction constraint.

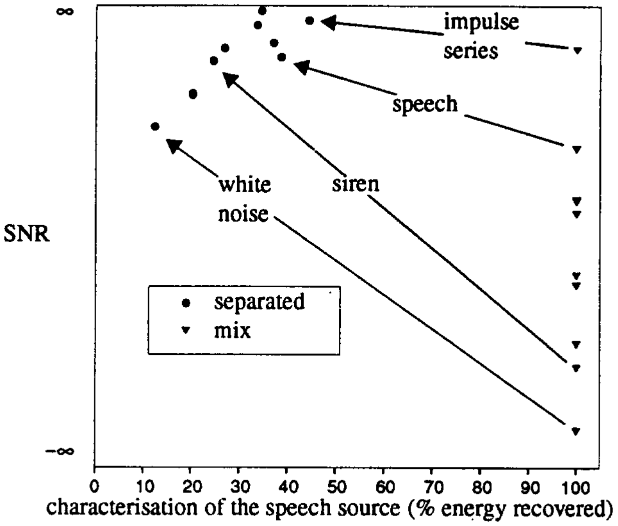

Fig. 1. *Results of computational ASA for speech mixed with 10 different noise sources (some of which are indicated on the figure). SNRs are measured before and after separation. Each point represents an average over 10 sentences.*

2. HMM RECOGNITION OF OCCLUDED SPEECH VIA MARGINAL DISTRIBUTIONS

In ASR using continuous density HMMs, each model state is associated with a probability distribution for the p-dimensional observation vector $\mathbf{x}$ modelled as a finite mixture of multivariate Gaussian distributions, so that the probability density function (pdf) of $\mathbf{x}$ when the model is in state j has the form:

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \mathrm{N}(\mathbf{x}, \mu_{jk}, \mathbf{U}_{jk}) \qquad (1)$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

Here, for each j, the c_{jk} for $k = 1, \ldots, M$ are mixture coefficients and $\mathrm{N}(\mathbf{x}, \mu_{jk}, \mathbf{U}_{jk})$ is the pdf of the p-dimensional Gaussian distribution with mean vector μ_{jk} and variance-covariance matrix $\mathbf{U}_{jk}$:

$$\mathrm{N}(\mathbf{x}, \mu_{jk}, \mathbf{U}_{jk}) =$$

$$\frac{1}{(2\pi)^{p/2} |\mathbf{U}_{jk}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_{jk})' \mathbf{U}_{jk}^{-1} (\mathbf{x} - \mu_{jk})\right) \quad (2)$$

where $|\mathbf{U}_{jk}|$ is the determinant of $\mathbf{U}_{jk}$ and $'$ denotes transpose.

At the heart of the Viterbi recognition algorithm it is necessary to compute the probability that a given distribution could generate a given observation vector. For occluded speech, this probability must be computed for a *partial observation vector.* If some of the components of $\mathbf{x}$ are not observed, calculation of the probability density can be based on the following distributional fact:

Suppose $q < p$ of the components of $\mathbf{x}$ are not observed. Then, the *marginal* distribution of the remaining components, $\mathbf{x}^*$ say, is also of mixed Gaussian form, and the corresponding pdf is

$$b_j^*(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} \mathrm{N}(\mathbf{x}^*, \mu_{jk}^*, \mathbf{U}_{jk}^*) \quad (3)$$

A formal proof is presented in Cooke, Green, Anderson & Abberley [10]. Informally, the marginal is simply derived from the full mixture density function by *striking out the rows and columns from the mean vector and covariance matrix corresponding to the missing components.* This is equivalent to integrating the distribution over the missing components.

To investigate the marginal distribution technique for occluded ASR we have performed a number of experiments in which varying proportions of the observation vector are randomly masked out. In [10] we demonstrated that in a phone recognition task using the TIMIT database and MFCC-derived parameter vectors, more than half the available components can be deleted without appreciable deterioration in performance. This holds both for single and mixture densities, and for diagonal and full covariance matrices.

In this paper we report on more realistic experiments, in which the parameterisation is in the auditory rather than the cepstral domain, the recogniser is presented with a mixture of speech and noise and the occlusion is derived from simulated auditory scene analysis.

3. EXPERIMENTS — LISTENERS AND MODELS

3.1 The NOISEX corpus

NOISEX [11] is a corpus intended to facilitate comparative studies of digit recognition in noise of various kinds and for a range of SNRs. It includes both single-digit and digit-triple material: here we have used only single digits – both training and test data consists of 10 repetitions of each digit by a single speaker. Our digit models have 8 emitting states with 'straight-through' topology and diagonal covariance matrices. The results presented here are for noise type 06 (multispeaker babble) but similar patterns of results have been found for other NOISEX sources. The HTK toolkit [12] was adapted for this study.

3.2 Signal representation

Mixture signals were transformed into an auditory rate map – a series of average firing rate vectors produced by a model of the auditory periphery. The model consisted of a 64 channel gammatone filterbank [13] covering the range 50 to 6500 Hz in equal steps of ERB-rate. The output of each filter was further transformed by a model of inner hair cell function [14].

3.3 Simulating auditory scene analysis

In the work reported below we have simulated the effect of ASA rather than construct a complete data path from our segregation algorithms. We have done this in order to explore the *potential* of ASA as an ASR front end. The ASA grouping principles which we have researched so far work well only for voiced speech. As a consequence, the results presented are conditional on a future scene analysis front-end being able to deliver a similar level of performance as our simulation.

The simulation procedure is illustrated in Fig. 2. We preserve those time-frequency regions in the rate map for each mixture in which the local SNR exceeds a threshold – these are the regions which are most likely to survive masking and be available for grouping by ASA. The value of the threshold can be used to determine a trade-off between the number of regions deemed to be part of the signal and the SNR represented by this selection. For instance, if we were to choose only those time-frequency regions where the local SNR is greater than 3 dB, a large number of regions would be removed, but the rate values would more accurately reflect those in the target signal. Conversely, an assumption that the auditory system could recover those regions with an SNR greater than -3 db, say, would lead to larger numbers of regions assigned to the speech source, but with values dominated by the noise.

For comparison with human performance on the same task, listeners were presented with the noisy digits across a range of SNRs. Their results are shown as the bold lines in figures 3-5, demonstrating effectively perfect recognition at 0 dB SNR, falling off towards chance level at -15 dB SNR.

Fig. 3 (left panel) shows recognition performance as a function of global SNR for various local SNR thresholds used in the ASA simulation. At a local SNR threshold of -10 dB (i.e. the signal is deemed to survive even if it is 10 dB below the noise), the model shows a poor match to listeners' performance except at very low SNRs. It is unreasonable to expect to recover signals via ASA (or any other technique) which are locally 10 dB down on the background. At the other extreme, a local SNR of 10 dB (i.e. the signal is selected if it is locally 10 dB above the noise) serves to recruit a very small number of time-frequency regions. It appears that these are inadequate to characterise the signal judging by the recognition performance. The best match to listeners' data is achieved at a local SNR of 0 dB (i.e. retain regions where the signal is locally more energetic than the noise). However, a gap remains between simulated ASA and human performance on this task.

Several explanations for this deficit can be hypothesised. Better performance might be achievable through an auditory representation based on neural synchrony rather than a rate representation. Paucity of training data is another factor. However, a more interesting possi-

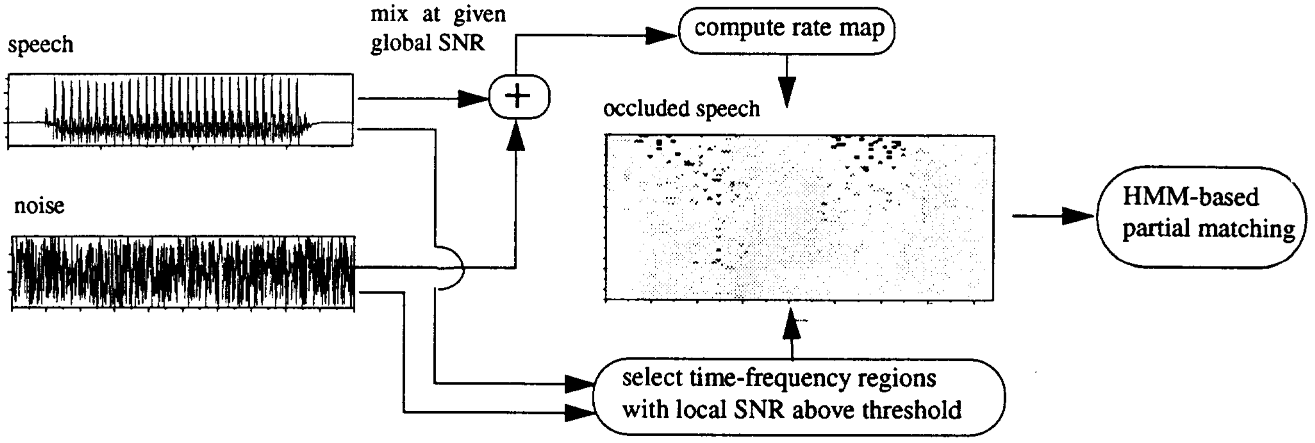
speech
mix at given global SNR
+
compute rate map
occluded speech
noise
HMM-based partial matching
select time-frequency regions with local SNR above threshold

Fig. 2. *Procedure for auditory scene analysis simulation.*

bility is that listeners may not only use an estimate of time-frequency regions which are deemed to be dominated by the target source, but could also verify their hypotheses against the mix as a whole using constraints on auditory induction.

4. EXPLOITING AUDITORY INDUCTION

It is well known that listeners can perceptually restore or *induce* missing acoustic elements so long as the missing regions are replaced by a suitable occluding source. When applied to speech perception, this is known as the *phoneme restoration effect* (Warren [15]). Since occlusion is a natural consequence of overlapping acoustic sources, auditory induction may be a powerful source of constraint in normal speech perception.

In the HMM framework, this constraint can be formulated in terms of the probability that components *in the mix* take on a value greater than the expected value. We have adopted the following procedure to incorporate this constraint into the calculation of $b_j(\mathbf{x})$:

1. Split observation **x** into subvectors **s** and **m** representing those deemed to be part of the target signal and those representing the remainder of the mix. For ease of presentation and without loss of generality, assume components are re-ordered such that **x** = (**s**,**m**).

2. Compute $b_j(\mathbf{s})$ using equation 3 as before.

3. Compute $b_j(\mathbf{m})$ as $\prod_i erf\left(\frac{m_i - \mu_i}{\sigma_i}\right)$

where *erf* represents the error function (computed using a Chebyshev approximation) and *i* ranges over subvector **m**.

4. Set $b_j(\mathbf{x}) = b_j(\mathbf{s}) \times b_j(\mathbf{m})$

Note that this is valid only for diagonal covariance matrices.

Fig. 3 (middle) shows the results obtained using this procedure. In general, performance is improved at a global SNR of 0 dB, though not greatly at lower SNRs. This is not altogether surprising since higher noise backgrounds will be compatible with a larger number of hypotheses, thus diminishing the value of the constraint.

Of interest here is the improvement at local SNRs of 0 and -5 dB, the latter suggesting that a more liberal regime which allows through time-frequency regions contaminated by noise may produce good performance at a range of global SNRs if used in conjunction with an auditory induction constraint. It may be possible to process the evidence from the two components, **s** and **m**, to give greater or lesser weight to auditory induction in order to further improve performance.

Fig. 3 (right) summarises the results at a local SNR of 0 dB, and compares the simulated ASA performance with a recent study using this database.

5. FURTHER WORK

This study demonstrates the *potential* for an ASA-based approach to robust ASR. Of course, it is necessary to complete the automatic path to determine whether computational ASA can deliver the separation required to support the simulated performance. This is a major focus of our current work, and will involve using ASA techniques devised for voiced speech unconditionally – relying on the induction constraint to fill in the gaps.

Another aspect of the new formalism involves training HMMs from incomplete patterns. We have speculated that the distributions learned in this way will be significantly different when using auditory representations from those derived from clean speech [16].

The underlying model of grouping which fits most closely with this work is one in which groups are signalled by some mechanism for highlighting which tonotopic channels are similar along some auditory dimension (e.g. common amplitude modulation, spatial location, onset or offset). One attractive scheme which achieves in-place grouping is that proposed by von der Malsburg & Schneider [17], in which channels deemed to be responding to components of the same acoustic source are forced to fire synchronously. We are currently investigating similar neural oscillator models for auditory scene analysis.

ACKNOWLEDGEMENTS

This work is supported by UK Science and Engineering Research Council Grant GR/H53174.

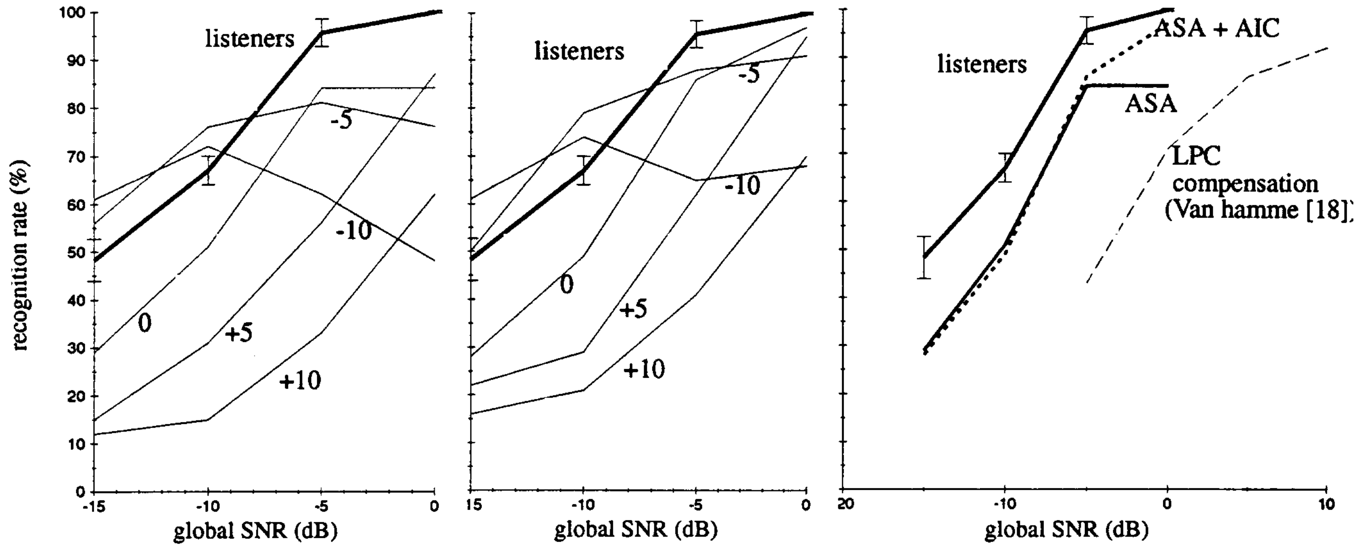

Fig. 3. *Isolated digit recognition in multispeaker babble as a function of global SNR. Performance at various indicated local SNRs for simulated ASA (left), augmented by an auditory induction constraint (middle) is compared to that of listeners (error bars represent +/- 2 standard errors). Results at a local SNR of 0 dB are redrawn in the right panel and compared to a recent NOISEX study.*

REFERENCES

[1] A.S. Bregman (1990), *Auditory Scene Analysis*, MIT Press.

[2] M.P. Cooke (1993), *Modelling Auditory Processing and Organisation*, Cambridge University Press.

[3] G.J. Brown & M.P. Cooke (1994), Computational auditory scene analysis, *Computer Speech & Language*, **8**, 297-336.

[4] T. Nakatani, T. Kawabata & H.G. Okuno (1993), 'Speech stream segregation by multi-agent system', *Technical Report of IEICE SP-97* (1993-11).

[5] D.P.W. Ellis (1993), 'A computer implementation of psychoacoustic grouping rules', *MIT Media Lab Perceptual Computing - Technical Report #224.*

[6] D.K. Mellinger (1991), *Event formation and separation in musical sound*, Ph.D. Thesis, Stanford University.

[7] M. Grenie & J.-C. Junqua (eds) (1992), *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes (ISSN 1018-4554).

[8] M.P. Cooke & G.J. Brown (1993), 'Computational auditory scene analysis: Exploiting principles of perceived continuity', *Speech Communication*, **13**, 391-399.

[9] M.P. Cooke, P.D. Green & M.D. Crawford (1994), 'Handling missing data in speech recognition', *Proc. ICSLP-94*, Yokohama, 1555-1558.

[10] M.P. Cooke, P.D. Green, C.W. Anderson & D.C. Abberley (1994), 'Recognition of occluded speech from hidden Markov models', *Dept. of Computer Science Research Report 94-0501, University of Sheffield.*

[11] A. Varga, H.J.M. Steeneken, M.J. Tomlinson & D. Jones (1992), 'The NOISEX-92 study on the effect of additive noise on automatic speech recognition', CD-ROM available from the Speech Research Unit, DRA Malvern, UK.

[12] S.J. Young (1992), *HTK Version 1.4: User, Reference and Programmer Manual*, Cambridge University Engineering Department, Speech Group.

[13] R.D.Patterson & J. Holdsworth (1992), 'A functional model of neural activity patterns and auditory images', In: *Advances in Speech, Hearing and Language Processing Vol. 3* (ed. W. A. Ainsworth), JAI Press, London.

[14] R. Meddis (1988), 'Simulation of auditory-neural transduction: further studies', *Journal of the Acoustical Society of America*, **83** (3), 1056-1063.

[15] R.M. Warren (1970), 'Perceptual restoration of missing speech sounds', *Science*, **167**, 392-393.

[16] M.P. Cooke, M.D. Crawford & P.D. Green (1994), 'Learning to recognise speech in noisy environments', *Proc. ATR Workshop on Biological Foundations of Speech Perception and Production*, Osaka (published as ATR Technical Report).

[17] C. von der Malsburg & W. Schneider (1986), 'A neural cocktail-party processor, *Biol. Cybern.*, **54**, 29-40.

[18] H. Van hamme, 'Ardoss: autoregressive domain spectral substraction for robust speech recognition in additive noise', *Proc. ICSLP-94*, Yokohama, 1019-1022.