Proceedings of the Institute of Acoustics

LEARNING TO RECOGNISE SPEECH FROM PARTIAL DESCRIPTIONS

Martin Cooke, Malcolm Crawford and Phil Green

University of Sheffield, Department of Computer Science, Sheffield, England.

1. INTRODUCTION

Robust speech recognition remains an elusive goal for speech technologists interested in the engineering of devices which can function in hostile acoustic environments. Various techniques, ranging from spectral subtraction (e.g. Lockwood & Boudy, 1991) to parallel model combination (Gales & Young, 1993) have been proposed to handle speech corrupted by noise (see Furui, 1992, for a recent review). Some of these approaches have been very successful at improving recogniser performance in the presence of certain types of noise. Kadirkamanathan (1992) presents a wide-ranging study of the hidden Markov model (HMM) decomposition technique, in which noisy speech is explained by a combination of HMMs for the speech and noise.

In spite of this progress, current approaches are difficult to generalise to the range of conditions in which listeners function adequately. For example, "noise" reduction algorithms generally assume stationary noise with known spectral properties. HMM combination techniques assume that exactly two sources are present and that models exist for noise sources as well as for the speech. By contrast, listeners appear to process speech in a robust fashion without such *a priori* constraints. In particular, when faced with an 'auditory scene', listeners are generally able to attend selectively to a single source.

Our recent work on computational auditory scene analysis (Cooke, 1993; Cooke & Brown, 1994; Brown & Cooke, in press) has attempted to apply principles of auditory organisation, derived from decades of psychoacoustic research (Bregman, 1990; see Darwin & Carlyon, in press, for an extensive recent review) to the problem of separating acoustic mixtures into groups of components which appear to derive from the same source. Figure 1 illustrates the separation of speech from a siren. As illustrated in this figure, the results of

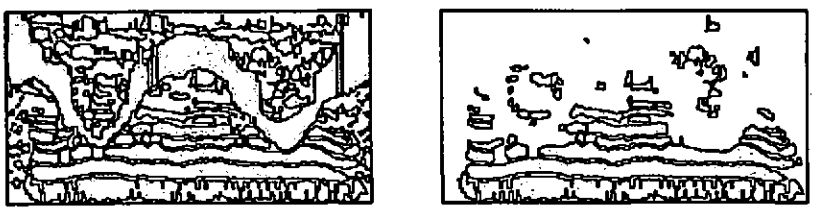


FIGURE 1. *Left*: Auditory time-frequency representation of an utterance mixed with a siren. *Right*: speech 'stream' produced by a model of ASA using the principle of grouping by pitch contour similarity.

sound source segregation may be far from perfect – for a variety of reasons, it may not be possible to recover complete sources from a mixture. The fact that we hear complete utterances rather than disjoint fragments, for example, is partly an illusion: it is well known (e.g. Warren, 1970; Warren *et al.*, 1994) that listeners can perceptually induce missing information under certain conditions which are guaranteed to hold in normal conditions. We have exploited these principles of perceptual continuity in our recent modelling work (Cooke & Brown, 1993) to obtain a certain degree of restoration. However, we have not used any information about specific sources such as speech, and there may be a limit to how much reconstruction can be achieved without such source-specific knowledge.

LEARNING TO RECOGNISE SPEECH FROM PARTIAL DESCRIPTIONS

If we are to develop auditory scene analysis followed by speech recognition as a new paradigm for robust ASR, it is necessary to address the question of how recognisers can be modified to handle fragmentary spectro-temporal descriptions of the sort illustrated in figure 1. There are two aspects to this question — the problem of *training* recognisers using incomplete patterns, and the issue of using such devices to *recognise* such "occluded" material. We have recently adapted the prevailing stochastic approach of continuous density hidden Markov models to tackle the latter issue (Cooke, Green, Anderson & Abberley, 1994; Cooke, Green & Crawford, 1994). In this paper, we also deal with the problem of training a recogniser based on partial information.

A method for training a self-organising Kohonen network (Kohonen, 1984) using incomplete data has recently been described by Samad & Harp (1992). This procedure is described in section 2. Section 3 demonstrates the results of applying the modified training procedure to a network which self-organises spectral vectors with missing components. Section 4 describes the task of recognition from partial data. Finally, we show (in section 5) how a constraint derived from auditory induction can be used to produce yet more robust performance.

2. A MODIFIED TRAINING PROCEDURE FOR SELF-ORGANISING KOHONEN NETS

Self organising networks of the form introduced by Kohonen (1984) have the property of mapping similarities in input space to topological proximity in a suitably chosen output space — typically, a two-dimensional grid. Such networks are trained in an unsupervised manner and have found application in many domains, including feature extraction for speech recognition (e.g. Kangas, Torkkola & Kokkonen, 1992; Patterson, Anderson & Allerhand, 1994).

The structure and operation of the Kohonen net is straightforward. The network consists of a two-dimensional grid of cells, each of which has an associated weight vector whose dimensionality matches that of the input patterns. Initially weights are set to small random values. During training, input vectors are presented to the network in turn, each cell computing its match to the input vector. Typically, the match is computed as the inner product of the input vector and the cell's weight vector, or, alternatively, as the smallest Euclidean distance. The cell with the best match is chosen as the winning node, and its weights, *and those of its neighbours*, are modified to bring them closer to the input patterns. The fact that updating takes place over a neighbourhood is the key to the spatial ordering property of Kohonen nets — that similar input patterns tend to produce responses in neighbouring cells. Typically, weight updates are made in accordance with the following formula[1]:

$$\Delta w_{ij}(t) = \eta(t)\exp\left[-d^2(j,c)/2\sigma^2(t)\right]\left(x_i(t) - w_{ij}(t)\right)$$

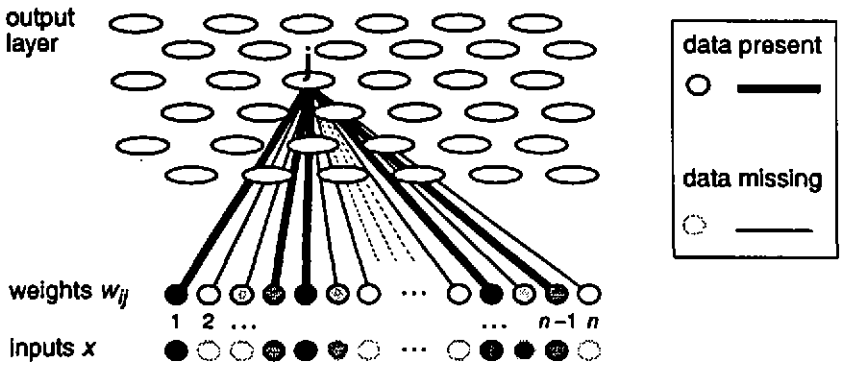
Here, w_{ij} is the weight between input component i and output unit j, $d(j,c)$ is the distance (in the output grid) between unit j and the winning unit c, and η and σ are parameters whose values are decreased geometrically during training. Progressively smaller weight changes are made to nodes further away from the winning cell.

For the case of input vectors with missing values, Samad & Harp (1992) presented the following modified training procedure. First, the winning node is computed using the subspace of available values. For example, using the Euclidean metric, instead of

$$m^j = \sum_{i=1}^{n} (w_{ij} - x_i)^2$$

where m^j is the distance between j's weight vector w_j and the input vector, x, and n is the dimensionality of

1. The development and notation used here is essentially that of Samad & Harp, and is presented for completeness.



	distance metric
Normal (Kohonen, 1984)	$\sum_{i=1}^{n}(w_{ij}-x_i)^2$
Incomplete vector (Samad & Harp, 1992)	$\sum_{i \in present}(w_{ij}-x_i)^2$
Incomplete vector with auditory induction constraint (see text)	$\sum_{i \in present}(w_{ij}-x_i)^2 + \sum_{i \notin present} max(0, w_{ij}-x_i)^2$

FIGURE 2. *Upper.* Input vector is compared with net weight vectors: the winning unit is chosen as that whose weight vector most closely matches the input vector, as determined by a distance metric. *Lower.* Distance metrics used in various experimental conditions. See text for details.

the vector we would compute:

$$m^j = \sum_{i \in present}(w_{ij}-x_i)^2$$

where *present* is the set of input units, i, for which values x_i are available at time t. This is illustrated in the table in figure 2, which also gives formulae for a further distance metric used in experiments detailed later.

Having found the winner, weight modifications are applied *only* to those elements of the weight vector corresponding to input values which are present:

$$\Delta w_{ij}(t) = \begin{cases} \eta(t)\exp[-d^2(j,c)/2\sigma^2(t)](x_i - w_{ij}(t)) & \text{if } i \in \textit{present} \\ 0 & \text{otherwise} \end{cases}$$

In the experiments described in sections 3 and 4, we modified the public-domain SOM_PAK software for Kohonen net simulation (Kohonen, Kangas & Laaksonen, 1992) in accordance with the foregoing, to handle partial data.

3. LEARNING FROM PARTIAL DATA

A net of dimensions 19x13 was used: recognition tests were performed after the nets had been trained using two epochs of 39000 and 195000 presentations and with learning rates of 0.01 and 0.05 respectively[1]. The input representation was produced by a 64-channel gammatone filterbank, with channel centre-frequencies equally spaced on an ERB-rate scale between 200 and 5500 Hz, and the output of each filter processed by a model of inner hair cell transduction (Meddis, 1988), smoothed over a 10 ms window. Training and test data was generated from utterances produced by a single male Japanese speaker from the ATR large-scale speech database (Kurematsu *et al.*, 1990)[2]. A balanced set of 250 frames of each of the 27 labels was randomly selected for use as testing data; from the remaining labels, 1000 were randomly selected as training data (when there were insufficient instances of a particular label, repetitions were made)[3].

Recognition performance of nets trained with varying proportions of data deleted from the input vectors was investigated in a series of 10 different conditions, in which *during training* input vector components were deleted at random with a probability which varied in from 0.0 (no deletion) to 0.9 (90% deletion). The trained nets were calibrated (i.e. one of 27 phone labels was attached to each output node) using the training set, and their recognition performance measured in terms of recognition accuracy — simply the percentage of labels in the test set correctly identified. Results of these experiments are given in the *left* panel of figure 3, which shows recognition accuracy as a function of deletion probability during training.

Performance is remarkably robust (albeit at a low baseline level) across all deletion conditions. Recognition accuracy barely drops, even when 90% of the frame is deleted at random. A useful visualisation of the whole trained net can be obtained by displaying the network weights in a spectrogram-like fashion, as shown in figure 4 which illustrates the similarity between nets trained using complete data vectors and those trained with 85% of the components deleted at random[4].

4. RECOGNITION FROM PARTIAL DATA

A second series of experiments was conducted to determine the effect of deletion of data *during recognition*. The procedure followed the same pattern as for the training experiments, but with a second dimension: during the recognition process, components in the input vectors were randomly deleted with a probability of between

1. Pilot studies indicated that nets of size 17x11 and smaller were unable to adequately encode the label-set, whereas increasing net dimensions above 19x13 gave diminishing improvements in performance for escalating processing requirements. Similarly, increasing the length of training runs did not significantly improve performance, whilst again lengthening the time required for training.
2. Experiments have also been conducted using data from multiple speakers from the TIMIT database, and using a PLP representation (Hermansky, 1990) as input to the nets, with similar results.
3. In a series of pilot studies it was found that non-balanced training data gave worse results than for balanced training (this might be expected, since the net then tended to be dominated by frequently-occurring labels, often to the exclusion of less-frequent ones). Furthermore, for this application, increasing the size of the training set did not appreciably improve performance.
4. This corresponds to approximately the proportion of deletions through simulated auditory scene analysis as described later.

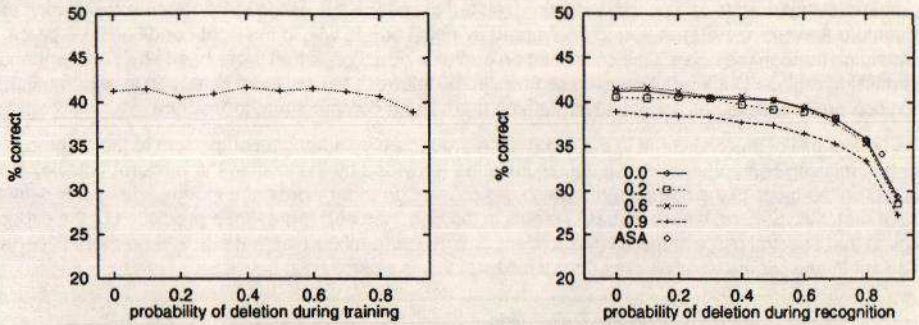


FIGURE 3. *Left*: Recognition accuracy vs. probability of component deletion during training. *Right*: Recognition accuracy vs. probability of component deletion during recognition, for various probabilities of deletion during training (given by the inset legend); see text for details of ASA condition.

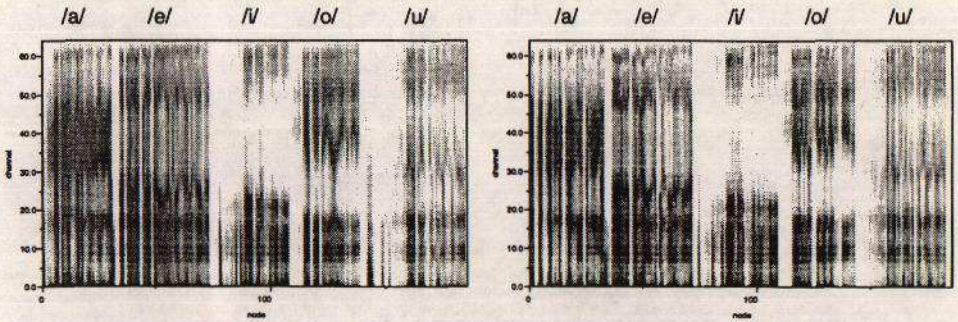


FIGURE 4. Spectrogram-like plots of net weight vectors (1 frame = 1 node) sorted by label (within-label order is insignificant) for nets trained using (*left*) complete data vectors and (*right*) data vectors with 85% of the components randomly deleted.

0 and 0.9, again in steps of 0.1. This gave a total of 100 results for each representation (10 deletion conditions during training by 10 deletions conditions during recognition), which are selectively summarised in the *right* panel of figure 3: again, recognition performance is encouragingly robust for nets trained in all conditions.

Of course, the distribution of deletions as a result of auditory scene analysis will be anything but random. In order to simulate the effects of ASA-deletions, a "mask" was created which denoted positions of spectral peaks in each frame of the test data. Spectral peak information was derived from Cooke's *synchrony strand* (Cooke, 1993) representation, which has been used as the basis for a model of auditory scene analysis. These peak positions were used to indicate channels in the input vector for which data was "present" (and corresponds to a deletion probability of around 85%). Using these "correlated" deletions improves recognition accuracy, as can be seen from the "ASA" point on the graph in figure 3.

5. EXPLOITING AN AUDITORY CONTINUITY CONSTRAINT

As was noted earlier, under certain circumstances the auditory system induces missing information. This

phenomenon is referred to as auditory induction; related to speech it is called the *phoneme restoration effect* (*cf.* Bashford & Warren, 1987). A sound interrupted by noise bursts will, in the right conditions, be perceived as continuing through the noise. One constraint on auditory induction is that there has to be sufficient energy in the missing regions to allow the missing segment to be inferred; hence, there is reason to assume that the recognition process itself has access to something more like a complete auditory scene.

We can make use of this constraint in the recognition procedure by adding a component to the distance metric for any missing components whose maximum value (provided by the level in the mixture) falls below that expected on the basis of the components which are present. In other words, the incomplete vector defines a possible matching pattern, whose values (weights in the Kohonen net) represent a prediction of the expected energy at that spectral place. If there is insufficient energy in the mixture at that place, then there is certainly insufficient in any source which makes up the mixture: this is illustrated in figure 5.

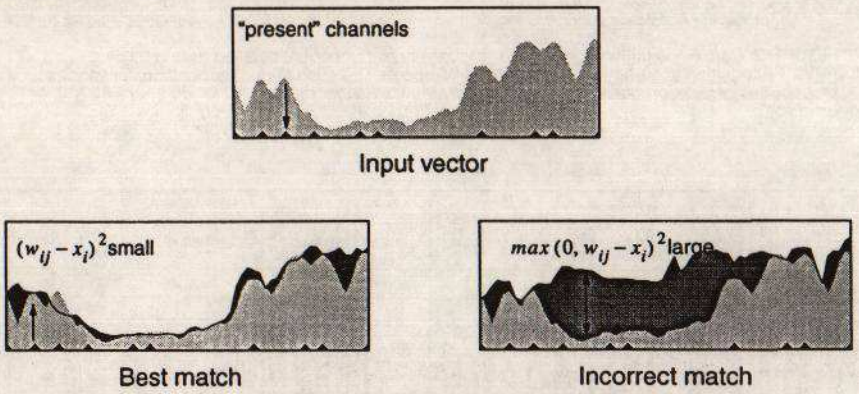


FIGURE 5. Auditory induction constraint: the "*best match*" panel shows a correct match between the input vector (light grey) and the net weight vector (dark grey). The distance measure is determined primarily by standard metric. The "*incorrect match*" panel shows a match between the same input vector and a net weight vector which "expects" more energy than is present in the signal.

The implementation of this constraint is shown in the third row of the table in figure 2. Figure 6 presents the results of adding this constraint to the recognition algorithm, and clearly shows a further flattening of the recognition curve as the probability of deletion increases. This demonstrates that even simple-minded application of auditory induction (better estimates of the missing components could be obtained using greater temporal context, for instance) prove beneficial in this architecture.

5. DISCUSSION

The computational studies reported in this paper show that speech recognisers do not necessarily require information from all spectral regions in order to function adequately. More important is the demonstration that it is still possible to learn, despite missing elements. We regard these results as sufficiently encouraging to pursue a new approach to robust ASR, based on an initial stage of auditory scene analysis, followed from recognition from partial data.

Of course, these studies are limited in extent and have presented a low baseline recognition performance, well below that which it is possible to achieve in more sophisticated recognition architectures. In other studies

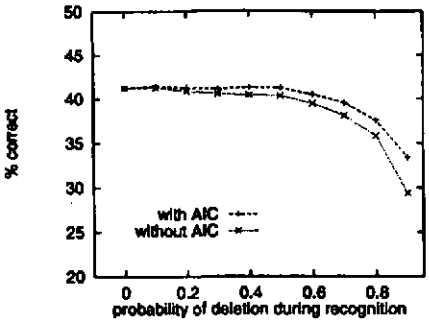


FIGURE 6. Recognition accuracy vs. probability of deletion for 2 recognition algorithms. The "without AIC" curve employs the standard distance measure whilst "with AIC" includes modifications suggested by the continuity effect as described in the text.

(reported in Cooke, Green, Anderson & Abberley, 1994), we show that it is possible to modify the power stochastic framework of HMMs to handle incomplete observation vectors. We have yet to demonstrate training within the HMM approach, but believe both that aspect, and the use of an auditory induction constraint, can be incorporated into the Viterbi search.

A further limitation is the lack of temporal context in the training/recognition process. In real listening situations, we would expect deletions to be correlated from instant to instant (due to correlations in both foreground and background sources). It is not obvious whether this will improve or degrade results; however, early results using the HMM approach — in which context does play a part — indicate that selective retention of spectral peaks produces enhanced performance.

The potential of ASA as a basis for robust ASR is clear — the new approach makes no assumptions about the number of acoustic sources present at any time, their prominence, or their spectro-temporal content. ASA therefore holds the promise of providing a *general* answer to the problem of speech recognition in a wide range of unpredictable acoustic conditions, free from the constraints afflicting most other proposed solutions.

What of the implications for speech perception? This study provides the first constructive demonstration of a process which infants might use to form auditory-phonological representations in normal listening conditions. The process is based on an unconditional stage of scene analysis, followed by organisation of the resulting fragmentary evidence. It is unlikely that the details are anything like those in our model, but the simulations presented here at least show one mechanism which can explain why infants need not be confined to an anechoic chamber for the first few months of life (unlike most ASR devices!).

It is possible to speculate further. Suppose speech perception is forcibly conditioned on the results of primitive scene analysis. One possibility is that access to 'speech schemas' — stored representations of speech sounds — is primarily via fragmentary descriptions. Further, top-down interaction could occur in a verificatory mode, as suggested by the auditory induction constraint (for further elaboration of these ideas see Cooke, Crawford & Green, 1994; also see Traunmüller, 1994).

ACKNOWLEDGEMENTS

This work was supported by SERC Image Interpretation Initiative Research Grant GR/H53174 and a study visit grant to ATR, Kyoto to Malcolm Crawford. Kohonen net simulations adapted the public domain SOM_PAK code (Kohonen, Kangas and Laaksonen, 1992), for which the authors express their thanks.

REFERENCES

J.A. Bashford Jr. & R.M. Warren (1987), "Multiple phonemic restorations follow the rules for auditory induction", *Perception & Psychophysics*, 42 (2), 114-121.

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

G.J. Brown & M.P. Cooke (in press), "Computational auditory scene analysis", *Computer Speech & Language*.

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

M.P. Cooke, P.D. Green, C. Anderson & D. Abberley (1994), "Recognition of occluded speech by hidden Markov models", University of Sheffield Department of Computer Science Technical Report TR-94-05-01 (submitted to *Computer Speech & Language*).

M.P. Cooke, P.D. Green & M.D. Crawford (1994), "Handling missing data in speech recognition", *International Conference on Speech and Language Processing*, Yokohama.

M.P. Cooke, M.D. Crawford & P.D. Green (1994), "Learning to recognise speech in noisy environments", ATR Workshop on *A biological framework for speech production and perception*, Kyoto, September (to be published as an ATR technical report).

M.P. Cooke & G.J. Brown (1994), "Separating simultaneous sound sources: issues, challenges and models"; In: *Fundamentals of Speech Synthesis and Speech Recognition*, E. Keller (ed.), John Wiley & Sons, 295-312.

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

C.J. Darwin & R.P. Carlyon (in press), "Auditory Grouping"; In: *Handbook of Perception and Cognition, Volume 6: Hearing*, B.C.J. Moore (ed.), Academic, Orlando, Florida.

S. Furui (1992), "Towards robust speech recognition under adverse conditions", *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes, 31-42.

M.J.F. Gales & S.J. Young (1993), "HMM recognition in noise using parallel model combination", *Proc. EUROSPEECH '93*, 837-840.

H. Hermansky (1990), "Perceptual linear predictive (PLP) analysis of speech", JASA, 87 (4), 1990 (April), 1738-1752

M. Kadirkamanathan (1992), "HMM decomposition recognition of speech in noise: a comprehensive experimental study", Proc. ESCA Workshop on Speech Processing in Adverse Conditions, Cannes, Nov. 92., 187-190.

A. Kurematsu *et al.* (1990), "ATR Japanese speech database as a tool of speech recognition and synthesis", *Speech Communication*, 9, 357-363.

T. E. Kohonen (1984), *Self-Organisation and Associative Memory*, Springer, Berlin.

T. E. Kohonen, J. Kangas & J. Laaksonen (1992), "SOM_PAK, The Self-Organizing Map Program Package, Version 1.2", SOM Programming Team, Laboratory of Computer and Information Science, Helsinki University of Technology, Finland.

P. Lockwood & J. Boudy (1991), "Experiments with a non-linear spectral subtractor, hidden Markov models & the projection, for robust speech recognition in cars", *Proc. EUROSPEECH '91*, Genoa.

R. Meddis (1988), "Simulation of auditory-neural transduction: further studies", JASA, 83 (3), 1056-1063.

R. D. Patterson, T. R. Anderson & M. Allerhand (1994), "The auditory image model as a preprocessor for spoken language", *Proc. ICSLP*, Yokohama, 1395-1398.

T. Samad & S.A. Harp (1992), "Self-organisation with partial data," Network, 3, 205-212.

H. Traunmüller (1994), "Conventional, biological and environmental factors in speech communication: a modulation theory", *Phonetica*, 51, 170-183.

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

R.M. Warren, J.A. Bashford, E.W. Healy & B.S. Brubaker (1994), "Auditory induction: reciprocal changes in alternating sounds", *Perception & Psychophysics*, 55(3), 313-322.