COMPUTATIONAL AUDITORY SCENE ANALYSIS: EXPLOITING THE CONTINUITY ILLUSION

M. P. Cooke and G.J. Brown

Department of Computer Science,
University of Sheffield, England.

ABSTRACT - Acoustic sources are often occluded by other sounds, yet the strategies employed by the auditory system are remarkably robust against these intrusions. There are often sufficient cues which allow the auditory system to determine whether sound components continue through such occlusions. In this paper, we review the situations where an assumption of continuity is warranted and show how the so-called *continuity illusion* can be modelled within a computational system for auditory scene analysis. We present results which demonstrate the practical effectiveness of the model in improving the performance of a system for segregating speech from other acoustic sources.

BACKGROUND

Acoustic sources such as speech are generally perceived against a background of other sounds, yet listeners are often able to recover sufficient information to allow interpretation of the individual sources. A recent theoretical account - *auditory scene analysis* (Bregman, 1990) - based on two decades of psychoacoustic investigation, proposes that the auditory system uses a combination of primitive (bottom-up) organisational principles and schema-driven (top-down or learnt) processes to determine which parts of the mixture are likely to have arisen from the same external event. Recent years have witnessed a number of attempts (Weintraub, 1985; Cooke, 1991; Mellinger, 1991; Brown, 1992; Ellis, 1992) to build computational systems for source segregation inspired by theoretical accounts of auditory scene analysis. Our own work has demonstrated the value of these principles in segregating speech from a variety of other sources. In a recent evaluation (Brown & Cooke, 1992), we showed that very large improvements in signal-to-noise ratios are possible using a system based primarily on grouping components which have sufficiently similar pitch contours.

Figure 1 presents a simplified view of the segregation process for a mixture consisting of an artificial siren and a voiced utterance. The map of auditory nerve firing rates is segregated to produce a 'mask' showing those time-frequency regions in the mixture that are believed to have arisen from the utterance. For evaluation purposes, the mask can be used in conjunction the utterance and the intrusion to produce a measure of the SNR after segregation. For the example shown in fig. 1, the SNR improves from .29 to .98 on a scale where 0=all intrusion and 1=all utterance.

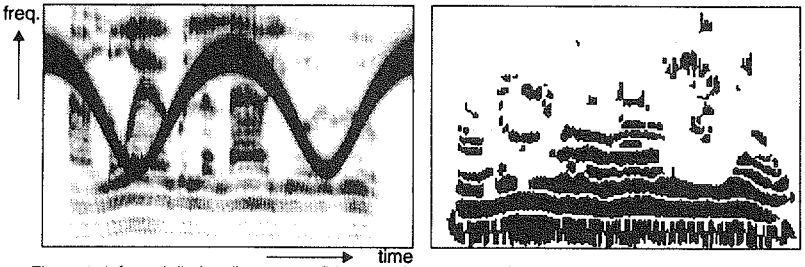

Figure 1. *left*: modelled auditory-nerve firing rates for a mixture of an artificial siren and the utterance "Our lawyer will allow your rule"; *right*: a 'mask' of time-frequency regions allocated to the utterance.

However, whilst SNR improvement is necessary, it is not sufficient. It is clear that significant chunks of the utterance have been removed by the mask. Little energy in the higher frequencies has been assigned to the utterance, and large sections of the utterance have been 'carved out' by the (louder) siren. In order to quantify the performance of the system in assigning energy to the utterance, we have developed a 'characterisation' metric. For the siren+speech example, we find that only 35% of the energy in the speech signal has been assigned by the segregation process (although the resynthesised speech remains intelligible).

Why is characterisation so poor? Part of an explanation lies in the fact that we are following a 'synthetic' approach to grouping (Ellis, 1992), in which organisational principles such as onset synchrony are used to piece together components of an acoustic source. We might say that the default is to segregate com-

ponents unless there is some reason to fuse them. Thus, a low characterisation performance points to the paucity of organisational principles in the model; certainly, the presence of speech-specific schema could help to fuse rather more of the auditory scene than the sole use of primitive principles as is the case in our current model. A possible solution is suggested by inspection of the firing rate map in fig. 1. Visually, components of the utterance appear to continue under the siren. There is an auditory equivalent to this which has received extensive experimental attention - the *continuity illusion*.

THE CONTINUITY ILLUSION

If parts of a signal are deleted and replaced with some other, louder, signal with the right characteristics, the softer sound is often heard as continuing through the louder intrusion. It has been demonstrated using simple tonal signals and with more complex signals such as speech (Warren, 1970). Bregman (1990) suggests an explanation in terms of auditory scene analysis, the role of which is to find the simplest explanation for some auditory configuration *which is not contradicted by any of the sensory evidence*. Consider the possible explanations for the configuration of fig. 2. It is possible that A1 and A2 are perceived as having separate explanations in the environment. Alternatively, A1, B and A2 may have arisen from the same source. A third explanation is that A1 and A2 arose from the same source, and that B is the manifestation of something else. Which explanation should the auditory system adopt? The answer depends on the precise nature of A1, A2 and B. But, if there is no evidence that A1 and A2 have arisen from separate events, the scene analysis explanation will adopt the simpler interpretation on the basis that it is more likely to reflect the sort of sound configurations that exist in the environment.

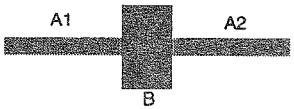

Figure 2. Components of the continuity illusion: B is a loud sound, whilst A1 and A2 are softer sounds with 'similar' properties. (Redrawn from Bregman (1990), fig. 3.22.)

Bregman goes on to separate out the questions of **whether** a sound has continued through an interruption and **what** the occluded sound is. Since it appears that any type of sound can undergo perceptual restoration, the 'what' question involves factors such as sentential context (e.g. Warren, 1970). How can we tell 'whether' a sound has continued through an interruption? Bregman suggests 4 rules which are outlined here. A detailed discussion of the psychoacoustic evidence for each is presented in his book (Bregman, 1990: p. 345-382).

1. **No discontinuity in A**: There should be no evidence that A1 stopped at the onset of B. Similarly, there should be no reason to suppose that A2 started just as B ended.

2. **Sufficiency of evidence for occlusion**: There must be enough neural activity during B - more than would have been generated by a continuation of A.

3. **A1-A2 grouping**: A1 and A2 must have similar properties. In fact, there should be reason to suppose that they have arisen from the same event.

4. **A is not B**: It should be possible to interpret B as the result of a separate event rather than as a continuation of A1.

One consequence of rules 3 and 4 is that perceptual restoration might occur *after* components have been grouped.

In the remainder of the paper we show how these rules can be used to improve the characterisation performance of the model. A general method for answering the 'whether' question is presented, followed by a specific solution to the 'what' issue which exploits harmonic relationships between components. Before describing how the rules governing perceptual restoration are built into a model, the representations and algorithms used as the basis for auditory scene analysis are outlined.

COMPUTATIONAL AUDITORY SCENE ANALYSIS

Figure 3 illustrates the representations used in the computational auditory scene analysis system (Brown, 1992) employed here. The model uses a series of representational abstractions called *computational maps*, motivated by the discovery of auditory maps which appear to place-code acoustic parameters in orderly two-dimensional arrays of neurons. Maps for periodicity, intensity, frequency transition, spectral shape and interaural time and intensity differences appear to be encoded in this fashion. In addition to the raw firing rate representation produced at the output of a model of the auditory periphery, Brown's model employs maps for onsets, offsets, frequency transitions and periodicity.

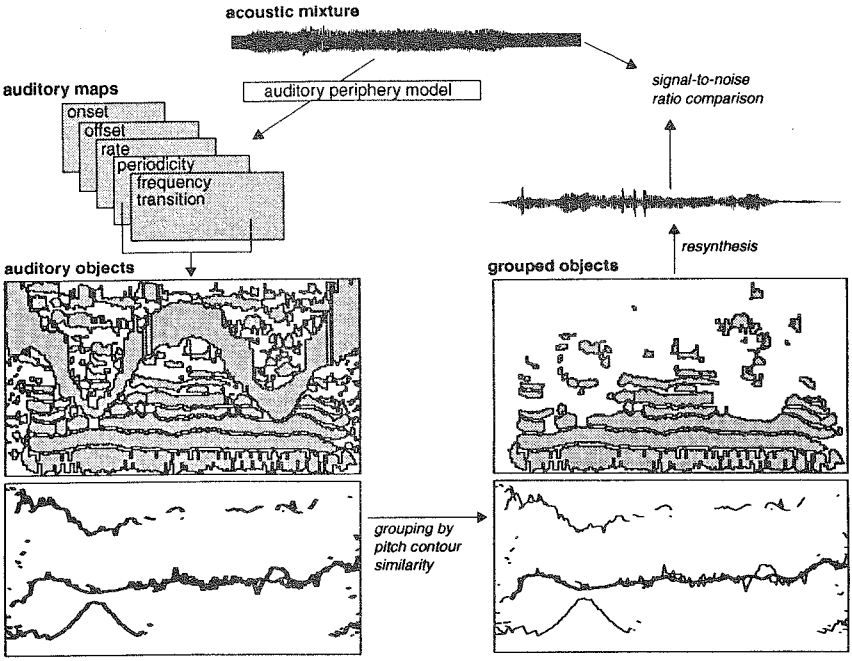

Figure 3. Auditory scene analysis: outline of representations and strategy in Brown's model. The lower panels show pitch contours for each auditory object, where the vertical axis corresponds to periodicity delay. Objects grouped by pitch contour similarity are shown on the right and have such contours highlighted.

The frequency transition and periodicity maps can be combined to produce a collection of 'auditory objects', each of which represents the evolution in time of collections of channels which have similar responses. Auditory objects resulting from the siren+speech mixture are shown in fig. 3.

Segregation of acoustic sources in this model is currently achieved by grouping objects which have sufficiently similar predicted pitch contours. The subset of objects grouped in this way for the siren+speech mixture is shown on the right of fig. 3. Other grouping rules based on onset and offset synchrony are present in the model but are not employed in this study.

MODELLING THE CONTINUITY ILLUSION

The strategy for exploiting the continuity illusion consists of two parts. First we determine whether perceptual restoration is required for each component of the auditory scene, then we decide what that restoration should consist of. To answer the 'whether' question, our model draws on auditory maps for rate, onsets and offsets. Bregman's 4 rules are transformed into cost functions in a dynamic programming search which attempts to link auditory objects grouped by Brown's model.

Specific rules are handled as follows:

1. **No discontinuity in A**: A cost is computed based on the evidence for an offset at the end of A1, and on the evidence for an onset at the start of A2 (using offset and onset maps respectively).

2. **Sufficiency of evidence for occlusion**: An extension of A1 should be supported by the presence of sufficient activity in the map of firing rate. A cost is incurred if the rate drops below that registered at the end of A1.

3. **A1-A2 grouping**: Since the search for perceptual restoration takes place after grouping, it ought to be the case that pairs of objects linked are parts of the same source. However, it is still possible for objects to link with inappropriate objects (e.g. a 2nd harmonic track linking with a 3rd harmonic; both are part of the same organisation but it would be an error to link them as part of a continuous track). Dealing with this effectively appears to rely on source-specific knowledge. A general principle which might be

used is frequency proximity: a cost could be associated with a transition in frequency. Here, we only attempt to restore similarly labelled harmonics, as described in the next section.

4. **A is not B**: If A1 is transformed gradually into B, little or no activity in the onset map at the transition would be expected (similarly for the offset map at the B-A2 interface). Conversely, any evidence for an onset at these points can be rewarded by introducing a negative cost based on the degree of onset activity at the A1-B interface (and for offsets at the B-A2 interface).

The costs associated with rules 1,2 and 4 above can be combined to produce a dynamic programming search through the time-frequency region between any pair of objects we are seeking to link. In the general case, the system does not know which A2 is a candidate continuation of A1, so it is necessary to allow the search to proceed up to some specified limit in time (300 ms in the current implementation). Indeed, more than one such continuation object may be discovered.

Letting R be the firing rate at the end of object A1, and $r(t,f)$,$on(t,f)$ and $off(t,f)$ denote the activity in the rate, onset and offset maps at time t in channel f, the local cost function, $c(t,f)$ is defined:

$$c(t,f) = \begin{cases} 0 & R < r(t,f) \\ \mu_{\text{rate}}(R - r(t,f)) & R \geq r(t,f) \end{cases} \quad (1)$$

where μ_{rate} is a constant used to weight the contribution of this component in relation to that made by the onset and offset maps. Equation 1 implements rule 2 above.

Initial costs are based on rules 1 and 4. For each channel f in the search extent, we define:

$$c(t_0,f) = \mu_{\text{on/off}}(off(t_0,f) - on(t_0,f)) \quad (2)$$

where t_0 is the time at which A1 ends. This provides a positive penalty for any evidence of an offset at the end of A1, balanced by evidence for an onset. The constant $\mu_{\text{on/off}}$ is a weighting factor.

The dynamic programming iteration uses a 3-way decision function; correspondingly, the cumulative cost $C(t,f)$ of extending A1 to a cell (t,f) is given by:

$$C(t,f) = c(t,f) + min(C(t-1,f-1), C(t-1,f), C(t-1,f+1)) \quad (3)$$

For each cell containing a candidate A2 start point, a penalty for any onset evidence is added, again balanced by evidence for an offset. For all such cells, the additional cost is given by:

$$c(t,f) = \mu_{\text{on/off}}(on(t,f) - off(t,f)) \quad (4)$$

Finally, the lowest cost continuation is determined (of course, there may be no such continuation present within the search region). Its cost reflects the likelihood that A1 has a continuation. The lower the cost, the more likely that A1 can be continued. Ideally, any evidence that rules 1-4 are broken should result in a negative answer to the 'whether' question. In practice, costs which are less than a small positive tolerance indicate that restoration is admissible - the precise value of this tolerance should depend on how conservative a restoration strategy one wishes to adopt. A choice of values for μ_{rate} and $\mu_{\text{on/off}}$ is dependent on many factors, not least of which is the sensitivity of the onset and offset maps. Currently, we weight contributions from eqns. 1,2 and 4 equally.

SIGNAL RESTORATION

Having determined that a component is likely to have been occluded, the question of what form the restoration should take is raised. Whilst an answer to this may be thought to require source-specific knowledge, there may be cues provided by primitive grouping processes such as harmonicity. Here, since components are grouped according to a principle of pitch contour similarity, it makes sense to use harmonic expectations to determine appropriate time-frequency tracks. Specifically, each component receives a harmonic label in the grouping process, and the cost of linking successive fragments of each harmonic is calculated using the approach outlined in the previous section. In fact, it is unnecessary to employ a DP search in this case because we can predict the time-frequency regions occupied by the missing harmonic fragment. A further modification to the general algorithm is possible since the system knows which A2 matches each A1 - they will have the same harmonic number. Hence, it is possible to predict the harmonic energy in the occluded region. Rather than use a constant R in eqn. 1, a predicted firing rate $R(t,f)$ is calculated by linear interpolation of the rates between the end of A1 and the start of A2. More sophisticated prediction methods based on speech-specific knowledge might be employed.

It is inappropriate to assign the full amount of energy in any restored time-frequency region to the signal, since some of it will belong to the occluding source. We can achieve such an assignment of energy by filling in the mask with real-valued weights, $w(t,f)$ computed as:

$$w(t,f) = min(1, \frac{R(t,f)}{r(t,f)}) \quad (5)$$

Figure 4 shows the weighted mask (left panel) corresponding to the siren+speech mixture and the restored speech (right panel) produced by this mask. It is evident that the restoration has been partially successful in reconstructing occluded parts of the utterance. Precisely how successful? Using the metrics described in the next section, we find that characterisation has improved from 35% to 49%, whilst the SNR improvement has been reduced from .98 to .92. Hence, as one might expect, the mask has allowed through some of the siren in addition to more of the speech signal. The next section provides a quantitative analysis of a larger mixture corpus.

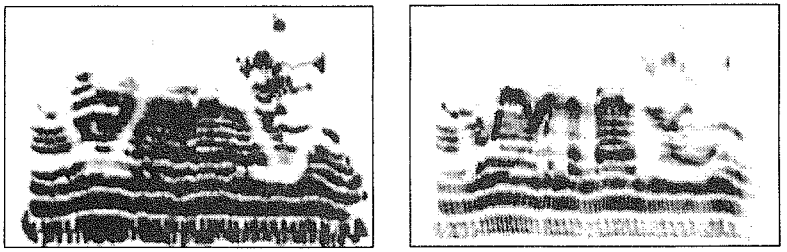

Figure 4. *left*: mask after exploiting the continuity illusion. *right*: restored speech formed by multiplying the mask with the rate map for the mixture.

PERFORMANCE METRICS - CHARACTERISATION AND SNR IMPROVEMENT

One way to assess the performance of a separation system is to measure some kind of distance between original and reconstructed signals. Whilst this is certainly possible, it does not necessarily lead to an intuitive metric which can be compared with listeners' ratings of intelligibility improvement. For example, a signal which contains very little intrusion (high SNR) after separation but poorly characterises the original signal may still be quite intelligible despite having a large objective distance score.

Instead, we prefer to separate out the two measures referred to earlier as characterisation and SNR improvement. The SNR for each 10 ms frame of a restored signal is measured using eqn. 6 below which varies from 0=all noise to 1=all signal. Quantities $s(t,f)$ and $n(t,f)$ denote the RMS energy in the signal and background, respectively, measured at time t and frequency f. Results thus obtained are averaged across the utterance. The arctangent compression is used since, for some frames, the denominator in eqn. 6 can equal zero.

$$\mathrm{SNR(t)} = \frac{2}{\pi}\mathrm{atan}\left(\frac{\sum_f \min[s(t,f), p(t,f)]}{\sum_f \max[0, p(t,f) - s(t,f)]}\right) \quad (6)$$

where $p(t,f) = w(t,f)\,[s(t,f) + n(t,f)]$ is the value of the signal predicted by the mask.

Characterisation is measured using eqn. 7. A value of 1 corresponds to a time frame in which the reconstructed signal completely characterises the utterance. Note in eqn. 7 that the role of the max function is to allow the signal to 'overcharacterise' the utterance i.e., to allocate more energy than required in some time-frequency regions. Of course, such an over-allocation will result in a degradation of the SNR as measured by eqn. 6.

$$\mathrm{CHAR(t)} = 1 - \frac{\sum_f \max[0, s(t,f) - p(t,f)]}{\sum_f s(t,f)} \quad (7)$$

Figure 5 shows the value taken by SNR and CHAR over a database of 100 mixtures (Cooke, 1991) consisting of 10 voiced utterances mixed with each of 10 various acoustic sources. Results in each case are averaged over the set of voiced utterances. The figure also shows the original SNR in the mixture

and the SNR after grouping but prior to exploiting the continuity illusion. Characterisation performance

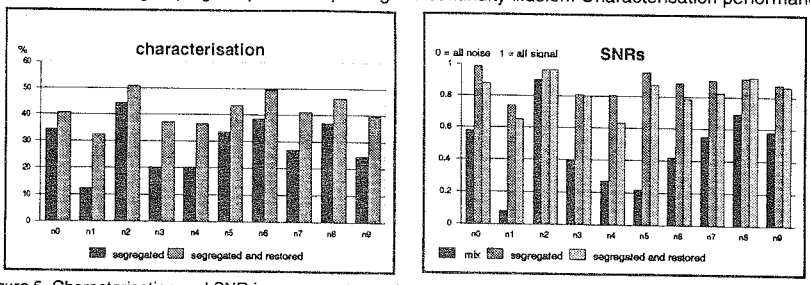

Figure 5. Characterisation and SNR improvement over the 100 mixtures. The noise sources are: n0=1kHz tone; n1=white noise; n2=impulse series; n3=lab. noise; n4=rock music; n5=siren; n6=telephone; n7=female speech; n8=male speech; n9=female speech. SNRs are measured before segregation, and before and after the continuity illusion is exploited.

increases in each case, sometimes to double its original level. SNR improvement generally falls, though its level is still well above that of the mixture itself. Informally, speech resynthesised from masks to which the continuity illusion has been applied generally sounds more natural than tokens resynthesised prior to application of this effect.

DISCUSSION

We have presented what we believe to be the first model to exploit the continuity illusion in the segregation of real acoustic mixtures. The results measured over a fairly large corpus indicate that real benefits are achievable in terms of reconstructing some target signal from the mixture, without much loss in SNR improvement. The general method, based on a dynamic programming search, can be used in any situation where evidence for occlusion is sought. Precisely how this evidence is used depends on such things as source-specific knowledge.

The fact that characterisation improvement leads to some loss in SNR points to deficiencies in estimating the level of each harmonic component during occlusion. Our simple strategy used here, based on linear interpolation, could certainly be improved upon. In fact, use of the rate map (rather than RMS energy) is probably inappropriate since the compressive effect of the hair-cell model feeding such maps will result in overestimated weights in the mask.

Improvements in characterisation beyond those reported here appear to require the development of more extensive auditory scene analysis strategies and the employment of further grouping principles. Future work will attempt to widen the range of cues used within the model to include, for example, those based on source location. We also intend to integrate source knowledge with primitive organisational principles.

ACKNOWLEDGMENT: MPC thanks the Royal Society for a study visit grant.

REFERENCES

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

Brown, G.J. (1992) *Computational Auditory Scene Analysis: A Representational Approach*, Ph.D. Thesis, University of Sheffield.

Brown, G.J. & Cooke, M.P. (1992) "Computational auditory scene analysis: grouping sound sources using common pitch contours", Proc. Inst. of. Acoustics, Windermere, November.

Cooke, M.P. (1991) *Modelling Auditory Processing and Organisation*, Ph.D. Thesis, University of Sheffield, to be published by Cambridge University Press.

D. Ellis (1992) *A Perceptual Representation of Audio*, M.S. Thesis, MIT.

Mellinger, D.J. (1991), *Event Formation and Separation in Musical Sound*, Ph.D. Thesis, Stanford University.

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

Weintraub, M. (1985) *A Theory and Computational Model of Monaural Auditory Sound Separation*, Ph.D. Thesis, Stanford University.