Modelling Auditory Scene Analysis: A Representational Approach

Guy J. Brown and Martin P. Cooke

Department of Computer Science, University of Sheffield

P.O. Box 600, Mappin Street, Sheffield S10 2TN, U.K.

Introduction

Speech is normally heard in the presence of other interfering sounds, a fact which has plagued speech technology research. In this paper, a technique for segregating speech from an arbitrary noise source is described. The approach is based on a model of human auditory processing. The auditory system has an extraordinary ability to group together acoustic components that belong to the same sound source, a phenomenon named *auditory scene analysis* by Bregman [1]. Models of auditory scene analysis could provide a robust front-end for speech recognition in noisy environments [4], and may also have applications in automatic music transcription [9]. Additionally, we hope that models of this type will contribute to the understanding of hearing and hearing impairment.

Auditory Maps

In analogy with the work of Marr [7] in vision, the modelling approach adopted here views audition as a series of representational transforms, each of which makes some aspect of the preceding representation explicit. A possible criticism of this functionalist philosophy is that the choice of representation is somewhat arbitrary. Therefore, we have attempted to justify our model by using representations that are motivated by physiological studies of the higher auditory system. Important acoustic parameters appear to be place-coded in the higher auditory system within orderly arrays of neurons, called *auditory maps*. The maps are two-dimensional, with frequency and some other parameter represented on orthogonal axes. The value of the parameter at a particular frequency is indicated by the firing rate of a cell at the corresponding position in the neural array. Parameters that appear to be coded in this manner include periodicity, intensity, frequency transition, spectral shape, interaural time difference and interaural intensity difference (see [2] for a review).

Physiological maps of this type provide a good basis for deriving useful representations of acoustic events. Hence, our approach has been to model auditory maps that provide the primitives needed for auditory scene analysis, and to demonstrate an algorithm for segregating concurrent sounds which exploits these primitives in an effective way.

Representations For Auditory Scene Analysis

Bregman [1] has noted that mechanisms of auditory scene analysis can be broadly classified as *simultaneous* or *sequential*. Simultaneous grouping processes segregate concurrent sounds into different perceptual streams, whereas sequential grouping processes segregate acoustic components that have arisen from the same source over time. Here, primitives for simultaneous grouping are provided by auditory maps coding *periodicity, onsets* and *offsets*. Primitives for sequential grouping are provided by a map of *frequency transition.*

Auditory Nerve. The input to each auditory map is provided by a simulation of firing activity in the auditory nerve. Mechanical filtering in the cochlea is modelled by a bank of gammatone filters, with centre frequencies spaced linearly on an ERB-rate scale. Transduction by inner hair cells is simulated by the Meddis [8] hair cell model, which provides a probabilistic representation of auditory nerve activity.

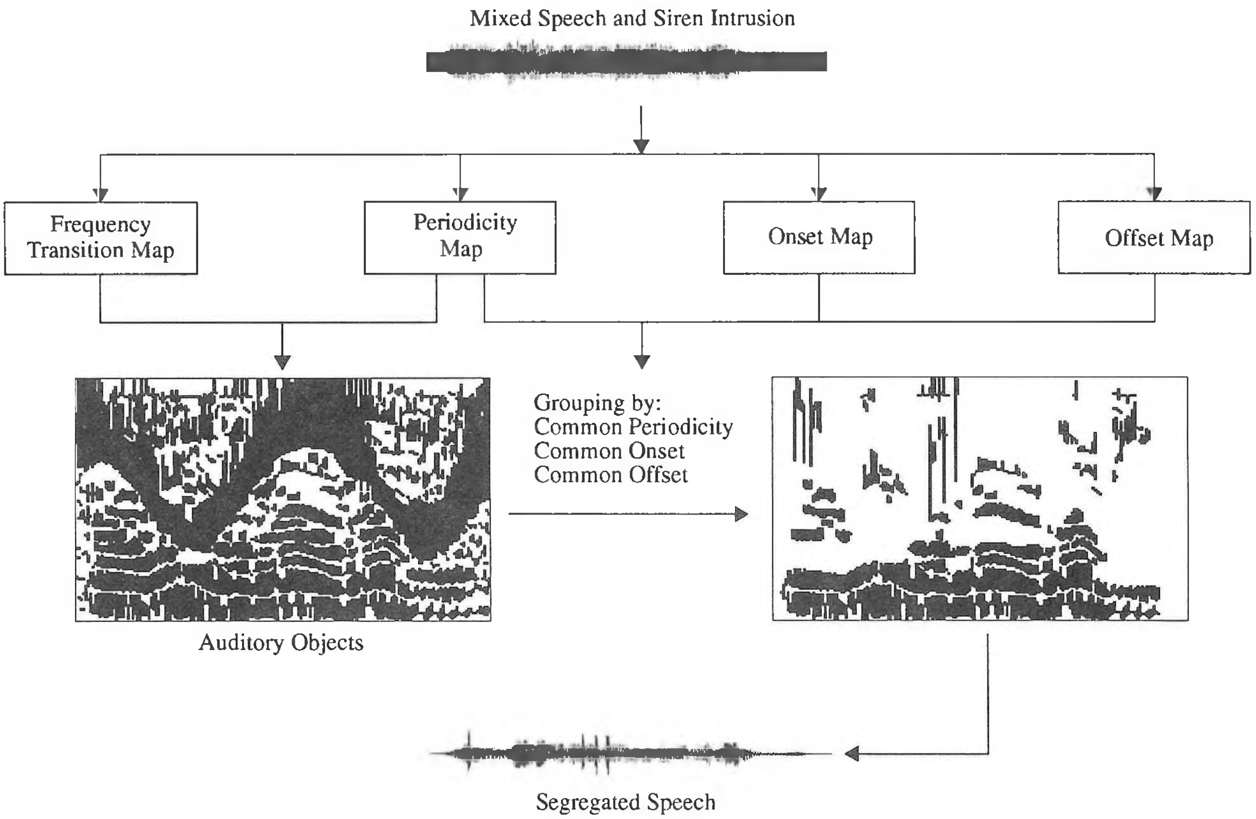
Periodicity Map. Periodicity information is extracted from the auditory nerve by a running autocorrelation analysis at each characteristic frequency, as suggested by Licklider [6] in his "duplex" model of pitch perception. Auditory filters that are responding to the same spectral dominance give rise to a similar pattern of response in the periodicity map, providing an early opportunity for grouping channels which belong to the same acoustic source. The temporal fine structure in adjacent channels is compared by a cross-correlation metric, and channels with a similar pattern of response are combined into *periodicity groups.*

Onset and Offset Maps. Onsets and offsets are identified by maps that have been developed in analogy with ON-IN cells in the cochlear nucleus, which receive an excitatory input and a delayed inhibitory input. The delay before inhibition is a variable parameter which is represented along one axis of the onset map, and determines the rate of amplitude onset that the cell is most responsive to. Offset cells are modelled in a similar manner, except that excitation is delayed relative to inhibition.

Frequency Transition Map. An early problem facing perceptual grouping mechanisms is how to relate the auditory representation of an event at a particular time with the representation of the same event at a later time. Here, a map of frequency transition is used to solve this "correspondence problem". Each neuron in the map is tuned to a particular sweep rate, according to the orientation of its receptive field. A similar idea has recently been proposed by Mellinger [9]. When a moving dominance is aligned with a receptive field, a peak of activity occurs at the corresponding frequency and orientation in the map. Therefore, the position and direction of movement of spectral dominances can be identified by locating the maxima in the frequency transition map at each time slice.

Modelling Auditory Scene Analysis

An algorithmic strategy for auditory scene analysis is employed, which emphasizes the time-frequency nature of the grouping process. Initially, the auditory scene is characterized as a collection of *auditory objects*, which are formed by using the frequency transition map to track periodicity groups across time and frequency. Subsequently, a pitch contour is derived for each object using the periodicity map, and auditory objects which have a similar pitch contour are grouped together. Objects can also be grouped if they start or end at the same time. In this case, the onset and offset maps are scanned to ensure that an onset or offset has occurred.



Evaluation

The model has been evaluated using the task of segregating speech from a wide variety of intrusions, such as "cocktail party" noise and other speech. The performance of the model has been assessed in two ways, one qualitative and the other quantitative. Firstly, given the periodicity groups that define a source, it is possible to resynthesize a waveform for the source which can be examined for intelligibility and naturalness in listening tests. The periodicity groups indicate which channels of the auditory filterbank belong to the source at a particular time, so that a resynthesized waveform can be obtained simply by summing these time-frequency regions of the (phase-corrected) gammatone filterbank output. Informally, segregated speech resynthesized in this way is highly intelligible and quite natural. Secondly, performance can be quantified by comparing the signal-to-noise ratio (SNR) before and after segregation (see [3]). On a test set of ten different noise intrusions, an improvement in SNR is obtained in each case.

Conclusions and Future Work

A model of auditory scene analysis has been described which uses information about periodicity, frequency transition, onset and offset to group acoustic components together. The model is able to segregate speech from a wide range of intrusive noise sources. Currently, the model is restricted to segregating periodic sounds (such as voiced speech). Additionally, the approach taken here is entirely data-driven, whereas it is known that learned (schema-driven) principles play an important role in auditory scene analysis [1]. Edelman [5] has suggested that neural maps are central to mechanisms of learning, and we are about to investigate the use of auditory maps in the formation and application of schema-driven grouping principles.

References

[1] A.S. Bregman (1989) *Auditory scene analysis.* MIT Press.

[2] G.J. Brown (1992) *A representational approach to auditory scene analysis.* PhD Thesis, University of Sheffield, in preparation.

[3] G.J. Brown and M.P. Cooke (1992) *A computational model of auditory scene analysis.* Proceedings of ICSLP-92, in press.

[4] M.P. Cooke (1991) *Modelling auditory processing and organization.* PhD Thesis, University of Sheffield.

[5] G.M. Edelman (1989) *Neural Darwinism: The theory of neuronal group selection.* Oxford University Press.

[6] J.C.R. Licklider (1951) *A duplex theory of pitch perception.* Experientia, 7, 128-134.

[7] D. Marr (1982) *Vision.* W.H. Freeman.

[8] R. Meddis (1986) *Simulation of mechanical to neural transduction in the auditory receptor.* Journal of the Acoustical Society of America, 79, 702- 711.

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