Computer Speech and Language (1994) **8**, 297–336

Computational auditory scene analysis

Guy J. Brown and Martin Cooke

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, U.K.

Abstract

Although the ability of human listeners to perceptually segregate concurrent sounds is well documented in the literature, there have been few attempts to exploit this research in the design of computational systems for sound source segregation. In this paper, we present a segregation system that is consistent with psychological and physiological findings. The system is able to segregate speech from a variety of intrusive sounds, including other speech, with some success.

The segregation system consists of four stages. Firstly, the auditory periphery is modelled by a bank of bandpass filters and a simulation of neuromechanical transduction by inner hair cells. In the second stage of the system, periodicities, frequency transitions, onsets and offsets in auditory nerve firing patterns are made explicit by separate auditory representations. The representations, *auditory maps*, are based on the known topographical organization of the higher auditory pathways. Information from the auditory maps is used to construct a symbolic description of the auditory scene. Specifically, the acoustic input is characterized as a collection of time-frequency elements, each of which describes the movement of a spectral peak in time and frequency.

In the final stage of the system, a search strategy is employed which groups elements according to the similarity of their fundamental frequencies, onset times and offset times. Following the search, a waveform can be resynthesized from a group of elements so that segregation performance may be assessed by informal listening tests. The system has been evaluated using a database of voiced speech mixed with a variety of intrusive noises such as music, "office" noise and other speech. A technique for quantitative evaluation of the system is described, in which the signal-to-noise ratio (SNR) is compared before and after the segregation process. After segregation, an increase in SNR is obtained for each noise condition. Additionally, the performance of our system is significantly better than that of the frame-based segregation scheme described by Meddis and Hewitt (1992).

1. Introduction

In 1953, Colin Cherry noted the ability of human listeners to attend selectively to the voice of one speaker in a mixture of many voices, and called this phenomenon the "cocktail party problem". Since then, the perceptual segregation of sound has been the

0885–2308/94/040297+40 $08.00/0

subject of extensive psychological research. Recently, a coherent account of this work has been presented by Bregman (1990). He contends that the mixture of sounds reaching the ears is subjected to an auditory scene analysis (ASA), which occurs in two stages. In the first stage, the acoustic signal is decomposed into a number of sensory components. Subsequently, components that are likely to have arisen from the same source are recombined into a perceptual stream.

Although ASA is documented comprehensively in the literature, there have been few attempts to exploit the known mechanisms of auditory grouping in the design of computational systems for sound source segregation. In this paper, we present a system for the segregation of harmonic sounds which is consistent with psychological and physiological auditory research. The motivation for our system is twofold. Firstly, the performance of automatic speech recognizers in the presence of other interfering sounds is poor compared to that of a human listener. Hence, a segregation system that exploits perceptual grouping principles could provide an improved front-end for automatic speech recognition in noise. Secondly, listeners with sensorineural hearing loss have difficulty in understanding speech in noisy environments (Festen & Plomp, 1983). A segregation system could form the basis for an "intelligent hearing aid", which would amplify a target voice while attenuating interfering noises (such as the voices of competing talkers).

Previous computational systems for source segregation have generally addressed the separation of a known number of sound sources with known characteristics. For example, several perceptual modelling studies have attempted to explain the finding of Scheffers (1983) that the ability of listeners to separately identify concurrent vowels is improved if the vowels have a different fundamental frequency (Scheffers, 1983; Assmann & Summerfield, 1990; Meddis & Hewitt, 1992). Since the average spectral characteristics of vowel sounds are constant over time, these schemes operate on a single auditory excitation pattern. In contrast, nearly all environmental sounds are non-stationary.

Indeed, relatively few models of auditory processing have been described which are able to segregate time-varying sounds. An early attempt is the work of Weintraub (1985), which aims to segregate and reconstruct the voices of two simultaneous speakers. He describes two systems, the most sophisticated of which consists of three main processing stages. Firstly, the pitch period of each voice is determined by analysing the interpeak intervals in the temporal fine structure of each channel of an auditory filterbank. Secondly, the number of active sources and their characteristics are determined by a pair of Markov models. The Markov model for a particular voice can be in one of seven states, corresponding to silence, periodic, non-periodic, onset, offset, increasing periodicity and decreasing periodicity. Finally, the amplitude spectrum of each voice is estimated, given the current state of its Markov model.

Beauvois & Meddis (1991) describe an auditory model in which stream segregation phenomena occur as emergent properties of low-level processing. The model is able to reproduce some simple examples of auditory stream segregation, but does not incorporate a mechanism for grouping components in different spectral regions.

A number of segregation systems have also been proposed that are based on conventional speech processing techniques rather than models of auditory function. Generally, these systems have concentrated on the use of pitch information to segregate simultaneous voices (e.g. Parsons, 1976; Stubbs & Summerfield, 1990) although the work of Denbigh and Zhao (1992) also exploits information about the spatial location of a target voice.

Consideration of previous computational approaches to source segregation suggests that they have suffered from two major limitations. Firstly, in an attempt to simplify the problem, strong assumptions have been made about the number and type of sound sources present. For example, schemes for speech enhancement often assume that the interfering source is another talker with a different average pitch (Weintraub, 1985; Denbigh & Zhao, 1992). These assumptions do not hold in natural acoustic environments, where many sound sources with characteristics that are unknown *a priori* may be active at the same time.

A second limitation of previous approaches arises from the fact that they have been heavily influenced by conventional speech processing techniques. Specifically, they represent the acoustic signal as a series of short-term spectral estimates, so that no information about temporal continuity is taken into account (e.g. Parsons, 1976; Denbigh & Zhao, 1992). Since time and frequency are intrinsically linked in sound, it seems more appropriate that strategies for source segregation should treat time and frequency as equally important dimensions of the acoustic signal.

The segregation system described in this paper addresses these problems by characterizing the auditory scene as a collection of time-frequency symbols. This allows the auditory scene to be searched rapidly, in order to identify symbols with similar properties and combine them into explicit groups. Consequently, our approach does not make strong assumptions about the number or type of sound sources present. A similar philosophy has been adopted in the segregation systems described by Cooke (1993) and Mellinger (1991).

A schematic diagram of the segregation system is shown in Fig. 1. The first stage simulates outer/middle ear filtering, cochlear filtering and neuromechanical transduction in the auditory periphery. In the second stage, information about periodicities, frequency transitions, and onsets and offsets in auditory nerve firing patterns is made explicit by separate auditory representations. The representations, *auditory maps*, are motivated by the known topography of the higher auditory system. In the third stage of the system, information from the map representations is used to construct a symbolic description of the auditory scene, which we call auditory elements. The final stage of the system employs a strategy for searching the *auditory scene*, which identifies elements with common FOs or common onset/offset times and combines them into explicit groups. A waveform may be resynthesized for each group of auditory elements, allowing qualitative and quantitative evaluation of segregation performance.

2. Auditory periphery model

2.1. Outer and middle ear resonances

The outer and middle ears constitute a complex acoustic cavity, which increases and decreases the sound pressure at the tympanic membrane at different frequencies. Since the outer and middle ears are approximately linear for small to moderate sound intensities, their resonances can be modelled by a simple linear filter. Here, a high-pass filter of the form

$$y(t) = x(t) - 0{\cdot}95x(t-1) \tag{1}$$

is used, where $x(t)$ is the input signal at time step t and $y(t)$ is the filtered output signal.

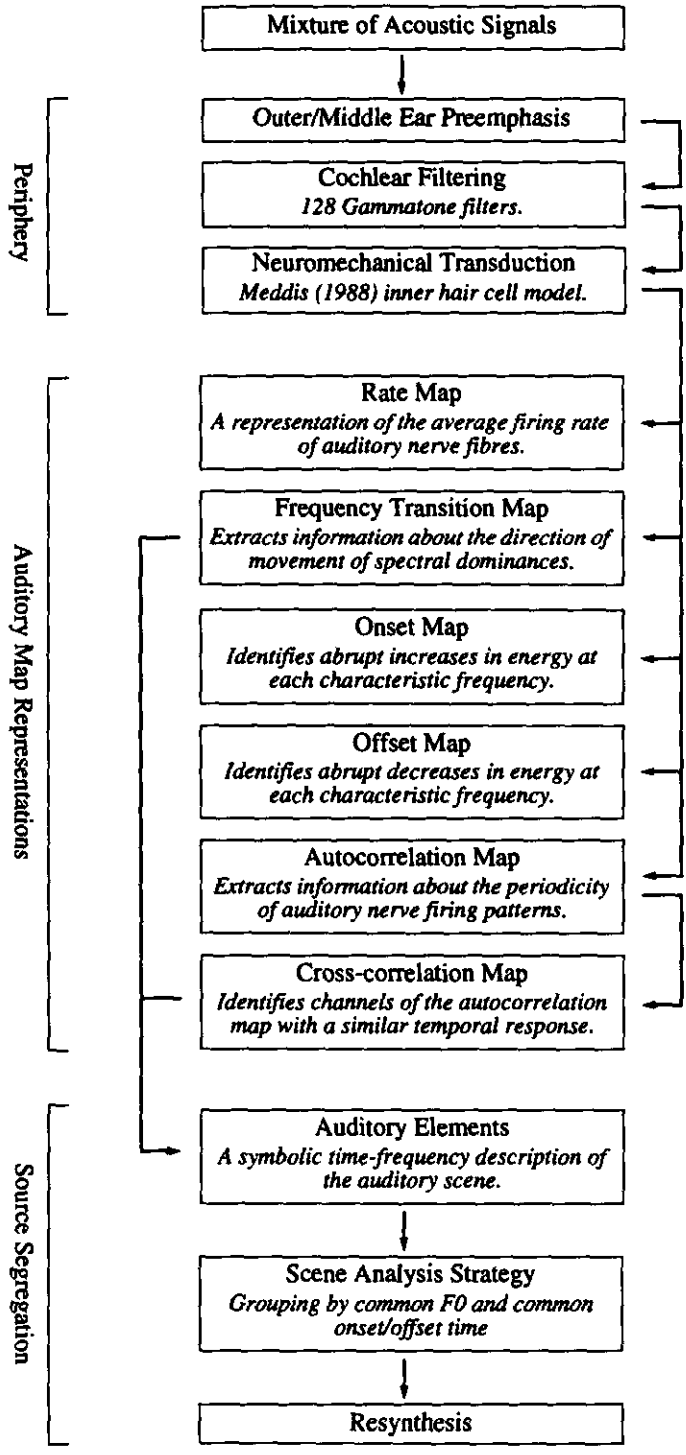

Figure 1. Schematic diagram of the segregation system.

The input signal is sampled at a rate of 16kHz, with 16 bit resolution. Although it is possible to model the transfer function of the outer and middle ears quite closely (e.g. Meddis & Hewitt, 1991), Equation (1) was considered to be an acceptable approximation for the functional approach adopted here.

2.2. Cochlear filtering

The frequency selective properties of the basilar membrane are modelled by a filterbank, in which each filter simulates the frequency response of a particular point along the cochlear partition. Physiological measurements of auditory nerve impulse responses have been made by de Boer and Kuyper (1968) using a "reverse correlation" paradigm. The filterbank used here is based on an analytical approximation of their experimental data, the "gammatone" function proposed by de Boer and de Jongh (1978). The impulse response of the gammatone filter of order n and centre frequency f_0Hz is given by

$$gt(t) = t^{n-1}\exp(-2\pi bt)\cos(2\pi f_0 t + \varphi) \tag{2}$$

where φ is phase and b is related to bandwidth. Here, fourth order filters are used ($n=4$).

In our functional auditory model, it is advantageous to compensate for the phase delays introduced by the filterbank. Specifically, phase is critical in the comparison of onset and offset times in different frequency channels, and the performance of our frequency transition map is improved if the filterbank is phase-compensated. Patterson, Holdsworth, Nimmo-Smith and Rice (1988) describe two types of phase compensation, both of which are employed. Firstly, the peaks of the envelopes of each impulse response can be aligned by introducing a time lead

$$t_c = \frac{n-1}{2\pi b} \tag{3}$$

to the output of the filter. Secondly, a peak in the temporal fine structure can be aligned with a peak in the envelope by the phase correction

$$\varphi_c = -2\pi f_0 t_c. \tag{4}$$

Substituting Equations (3) and (4) into Equation (2), this leads to the phase-compensated gammatone filter

$$gt_c = (t+t_c)^{n-1}\exp(-2\pi b(t+t_c))\cos(2\pi f_0 t) \quad (t \geq -t_c) \tag{5}$$

in which the peak impulse response at time $t=0$ is aligned for each characteristic frequency. Here, a digital approximation of Equation (5) is employed, where an impulse-invariant transform is used to convert from the continuous domain to the digital domain (Cooke, 1993).

Auditory filters are distributed across frequency according to their bandwidths, which increase quasi-logarithmically with the centre frequency of the filter. Here, the gammatone filters are spaced on the equivalent rectangular bandwith (ERB) scale of

Glasberg and Moore (1990). Specifically, 128 overlapping filters were spaced equally in ERB-rate in the range 50–5000 Hz, according to the relation

$$E(f) = 21{\cdot}4 \log_{10}(4{\cdot}37f + 1) \tag{6}$$

where $E(f)$ is the number of ERBs and f is frequency in kHz.

2.3. Neuromechanical transduction

The multiple-reservoir model of inner hair cell transduction described by Meddis (1986) is employed to convert the activity in each filter channel to simulated auditory nerve discharges. Given the output from the gammatone filterbank, the Meddis model computes the probability of a spike occurring in the auditory nerve. Here, the model is configured according to the parameters given in (Meddis, 1988), which simulate an auditory nerve fibre with a high spontaneous firing rate. The Meddis hair cell model is described comprehensively in the literature and is not discussed further here.

2.4. Example representations

Fig. 2 shows a representation of average firing rate in the auditory nerve for three sound sources. Here, the spike probabilities from the Meddis hair cell model have been integrated over a 20 ms Hamming window, and displayed at 10 ms intervals to give a *rate map*. Regions of spectral dominance in speech (harmonics and formants) are clearly represented as dark bands of intense firing activity.

3. Higher auditory representations

3.1. Introduction

A recurring motif in neurophysiology is the *computational map*, a term which describes an array of neurones that are systematically tuned for a particular parameter value (Knudsen, duLac & Esterley, 1982). There is good evidence that computational maps in the higher auditory system have a two-dimensional form, in which characteristic frequency and the value of some other parameter are represented on orthogonal axes. Acoustic parameters that appear to be represented in this way include intensity (Suga & Manabe, 1982), frequency modulation (Shamma, Vranic & Wiser, 1992), amplitude modulation (Schreiner & Langner, 1988) and spatial location (King & Hutchings, 1987).

Our approach employs functional models of a number of auditory maps, in order to provide primitive information for subsequent scene analysis processing. Effectively, computational maps provide intermediate representations that bridge the gap between the acoustic input and a symbolic auditory description of that input (Brown, 1992). The following sections describe the map representations used in the segregation system, and the motivation for them.

3.2. Autocorrelation map

3.2.1. Motivation

Recently, theories of pitch perception have been proposed which combine features of pattern recognition models (e.g. Goldstein, 1973) and temporal models (e.g. Licklider,

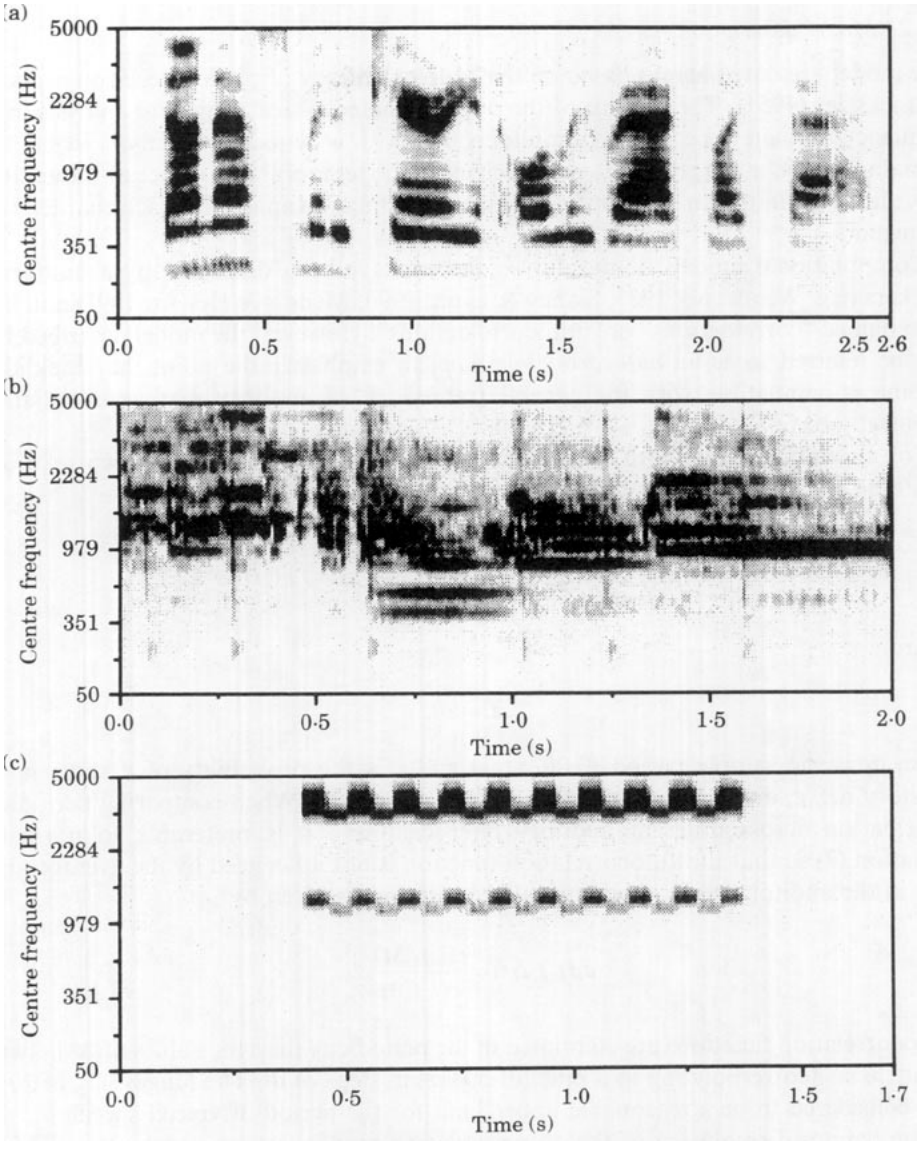

Figure 2. Average auditory nerve firing rate ("rate map") representations of female speech (a), rock music (b) and trill telephone (c). Dark areas indicate regions of intense firing activity. The speech signal is utterance dr1/fdaw0/sa2 ("don't ask me to carry an oily rag like that") from the TIMIT database (Garofolo & Pallet, 1989).

1951) by integrating periodicity information across resolved and unresolved harmonic regions (e.g. Meddis & Hewitt, 1991). Combined models of this type are able to account for many classical psychophysical pitch phenomena. Additionally, it appears that mechanisms similar to those underlying pitch perception can contribute to the perceptual segregation of sounds which have a different fundamental frequency (Scheffers, 1983).

3.2.2. Model description

The model presented here is based on the "duplex" theory of pitch perception proposed by Licklider (1951). The essence of the duplex scheme is that a spectral analysis in the frequency domain is performed simultaneously with a periodicity analysis in the time domain. Licklider suggests that periodicities in the temporal fine structure of auditory nerve firing patterns can be identified by an autocorrelation analysis at each characteristic frequency.

Computational models of the duplex theory have been described by a number of workers (e.g. Weintraub, 1985; Slaney & Lyon, 1990; Meddis & Hewitt, 1991) and have been named "correlograms" or "autocorrelograms". However, the model described here will be referred to as an *autocorrelation map*, to emphasize the point that Licklider's scheme is compatible with the general framework of auditory map representations discussed above.

For an auditory filter with characteristic frequency f, the running autocorrelation c at a time lag Δt is given by

$$c(t,f,\Delta t)=\sum_{i=0}^{\infty} r(t-T,f)r(t-T-\Delta t,f)h(T) \tag{7}$$

where

$$T=idt. \tag{8}$$

Here, dt is the sample period (0·0625 ms) and r is the probability of a spike in the auditory nerve, derived from the Meddis hair cell model. When comparing periodicity information across different auditory filter channels, it is preferable to normalize Equation (7) so that the autocorrelation function is not influenced by the average firing rate in the auditory nerve. The normalized response is given by

$$a_n(t,f,\Delta t)=\frac{c(t,f,\Delta t)}{c(t,f,0)}. \tag{9}$$

Autocorrelation functions are computed in the periodicity map for values of Δt between 0 and 20 ms (corresponding to a pitch of 50 Hz) in steps of dt. The longest lag of 20 ms was considered to be a reasonable upper limit for the period of voiced speech.

The temporal resolution of the autocorrelation map is determined by the width of the window $h(T)$. In his original paper, Licklider (1951) suggests (without any justification) an exponential window with time constant 2–3 ms. However, this window seems too short to give an accurate measurement of the period of sounds with a low pitch, such as the voice of a male speaker. A longer window (about 10 ms) is suggested by Plack and Moore's (1990) study of temporal masking, and their window is an asymmetrical bell-shape rather than the exponential suggested by Licklider. Here, $h(T)$ is a Hamming window of width 10 ms, which gives a reasonable approximation to Plack and Moore's data.

3.2.3. Example representation

An autocorrelation map for the vowel /æ/, excised from the utterance shown in Fig. 2, is shown in the left panel of Fig. 3. Periodicities in each channel are clearly delineated.

Every channel has a peak at the period of the vowel (5·4 ms, corresponding to an F0 of approximately 185 Hz) and its multiples, and these line up across frequency forming "spines" that run vertically through the plot.

3.3. Cross-correlation map

3.3.1. Motivation

It is evident from Fig. 3 that the autocorrelation map contains redundant information. Contiguous sections of the auditory filterbank respond to the same spectral dominance, so that channels with centre frequencies close to the same harmonic or formant have a similar pattern of periodicity. This redundancy provides an early constraint which can be used to group channels of the autocorrelation map that are responding to the same acoustic component. A similar observation has motivated the DOMIN algorithm of Carlson and Granström (1982) and the "pseudospectrum" described by Deng and Geisler (1987).

3.3.2. Model description

Regions of the autocorrelation map that have a similar pattern of periodicity can be identified by cross-correlating the responses of adjacent filter channels. Formally, the similarity at time t of two channels with centre frequencies f_1 and f_2 is given by

$$sim(f_1,f_2,t)=\frac{2\sum_{\Delta t} a_n(t,f_1,\Delta t)a_n(t,f_2,\Delta t)}{\sum_{\Delta t} a_n(t,f_1,\Delta t)^2+\sum_{\Delta t} a_n(t,f_2,\Delta t)^2}. \tag{10}$$

The cross-correlation given in Equation (10) is rate-normalized, so that a difference in the average firing rate of two channels does not affect their similarity score. Consequently, *sim* has a value between zero (no similarity in periodicity) and unity (identical pattern of periodicity).

Given this metric, it is necessary to decide how high the similarity score of adjacent channels in the autocorrelation map must be in order for them to form a group. The approach employed here is to construct a *cross-correlation map*, which indicates the groups that are formed at a series of different similarity scores. A cross-correlation map for the vowel /æ/ is shown in the centre panel of Fig. 3. Like other auditory maps, it is a two-dimensional organization in which characteristic frequency and a tuned parameter (in this case, similarity score) are represented on orthogonal axes. Adjacent channels of the autocorrelation map that have a value of *sim* equal to or greater than the threshold similarity score are allowed to form a group. At the highest similarity threshold (left of the map), no groups occur since adjacent channels are not identical. However, as the threshold is relaxed, channels with a similar pattern of periodicity begin to group together. In the figure, groups of channels that extend across frequency and different threshold of similarity are represented by rectangles, and are referred to as *periodicity groups*.

This technique is motivated by the "dendrogram" method of acoustic-phonetic segmentation described by Glass and Zue (1988). Whereas the dendrogram identifies changes in the spectrum over time, the cross-correlation map identifies changes in periodicity over frequency. However, the principle is similar, since both techniques

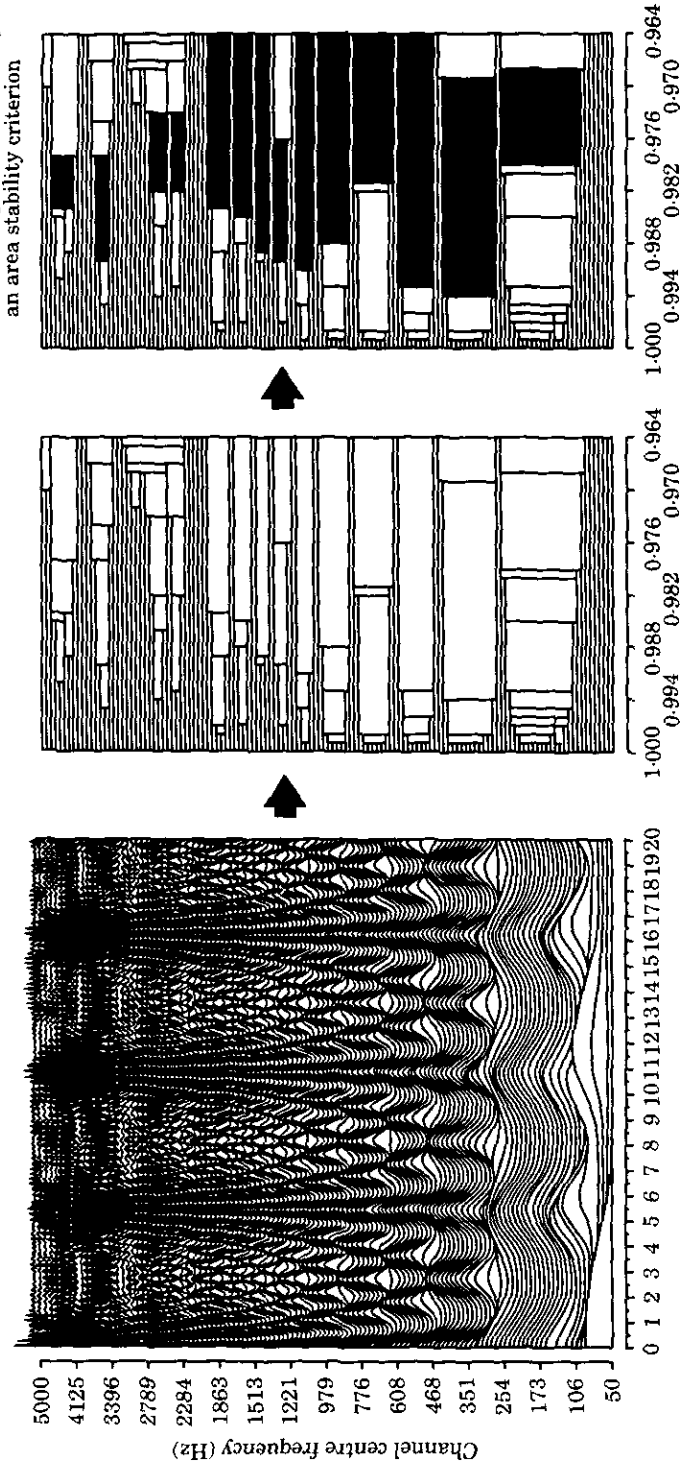

Figure 3. Autocorrelation map (left), cross-correlation map (centre) and selected periodicity groups (right) *for the vowel /æ/,* excised from the utterance shown in Fig. 2. The periodicity groups delineate channels *in the* autocorrelation map that have a similar temporal response.

attempt to find features in a representation that are stable across different scales of comparison. Clearly, the cross-correlation map in Fig. 3 contains many alternative groupings at different thresholds of similarity. Here, "good" groups are taken to be those that are stable across frequency and similarity threshold. Specifically, an "area stability criterion" is used, in which periodicity groups are selected if they have no descendents with a greater area in frequency-similarity space (the descendents of a group lie to its left in the map). The selected groups are shown as grey rectangles in the right panel of Fig. 3. As required, the groups delineate areas of similar periodicity in the vicinity of harmonics and formants.

3.3.3. Example representations

Periodicity groups for three sound sources, selected from each frame of the cross-correlation map by an area stability criterion, are shown in Fig. 4. The groups (black blocks) indicate regions of the auditory filterbank that have a similar response across characteristic frequency. Comparison of Fig. 4 with the rate maps in Fig. 2 indicates that areas of spectral dominance (e.g. harmonics and formants of speech) are clearly delineated by the periodicity groups.

3.4. Frequency transition map

3.4.1. Motivation

An early problem facing perceptual grouping mechanisms is how to match the auditory representation of an acoustic event at a particular time with the representation of the same event at a later time. This task is the auditory analogue of the *correspondence problem* which arises in the perception of visual motion (Ullman, 1979). It is likely that the auditory system uses two cues, frequency proximity and alignment on a common time-frequency trajectory, to solve the correspondence problem (Tougas & Bregman, 1985). Since many natural sounds (such as speech) consist of glides in frequency, it might be supposed that trajectory is an important grouping cue. Indeed, there is some evidence that the auditory system measures frequency transitions, and that it uses this information to group frequency components across time according to their trajectories (Ciocca & Bregman, 1987).

3.4.2. Model description

A schematic of the model frequency transition map is shown in Fig. 5. Cells in the map (spheres) are arranged in a two-dimensional framework, with characteristic frequency represented on one axis and frequency transition represented on the other. Each neurone is tuned to a particular rate and direction of frequency sweep, depending on the orientation of its receptive field. Similar schemes have been proposed by Mellinger (1991) and, in a non-auditory context, by Riley (1989).

The firing rate of each neurone in the map is determined by convolving its receptive field with the simulated auditory nerve response from the Meddis hair cell model. Hence, for a cell with characteristic frequency f and receptive field orientation θ, the firing rate $s(t,f,\theta)$ at time t is given by

$$s(t,f,\theta)=\sum_{i=-N}^{N}\sum_{j=-M}^{M} r(t+i,f+j)g_{\theta}(i,j) \tag{11}$$

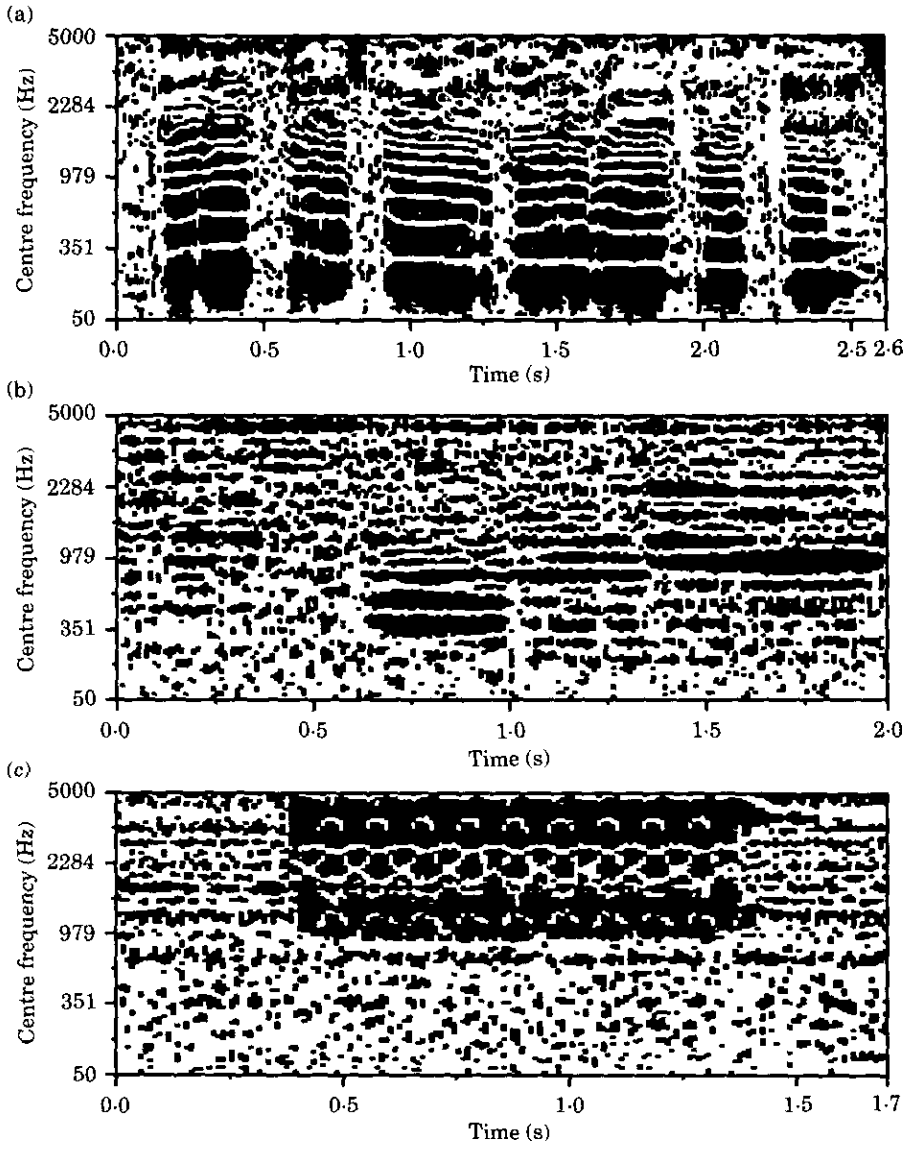

Figure 4. Periodicity groups for the speech (a), music (b) and telephone sound (c) sources. Black blocks indicate regions of similar temporal response across channels of the auditory filterbank.

where $2N+1$ and $2M+1$ define the width in time and frequency of the receptive field g_θ. As before, r is the probability of a spike in the auditory nerve.

Each neurone in the map is required to be tuned to a particular rate of frequency transition. This implies that the receptive field of a cell must elicit a maximal response when it is aligned with a spectral peak which is moving at the cell's preferred rate. The function

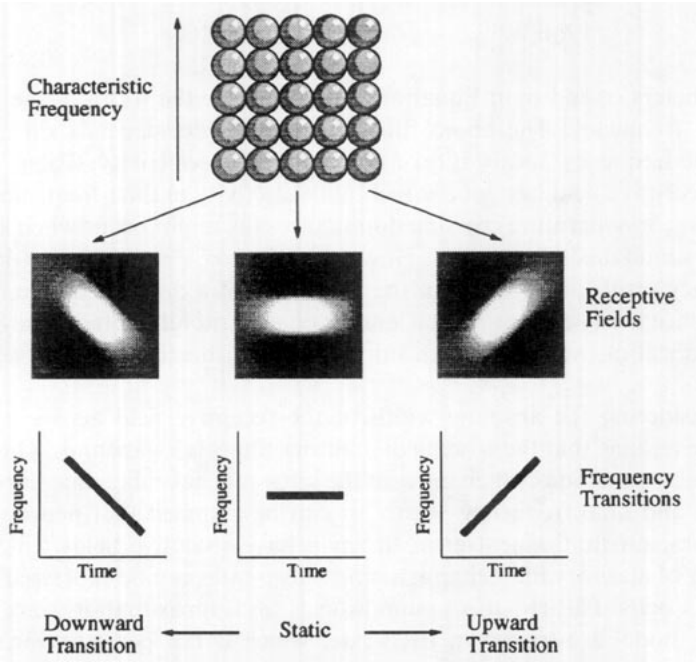

Figure 5. Schematic of the frequency transition map. Cells in the map (spheres) are tuned to different rates of frequency transition, depending on the orientation of their receptive fields.

$$g(t,f)=\frac{\partial^2}{\partial f^2}G(t,f) \tag{12}$$

suggested by Riley (1989) is used to satisfy this condition, where $G(t,f)$ is a two-dimensional Gaussian

$$G(t,f)=\exp\left(-\frac{t^2}{2\pi\sigma_t^2}-\frac{f^2}{2\pi\sigma_f^2}\right). \tag{13}$$

A plot of the receptive field $g(t,f)$ is shown in the middle of Fig. 5. It consists of a central excitatory (positive) region and two flanking inhibitory (negative) regions which confer directional selectivity. In the form given in Equation (12), $g(t,f)$ responds maximally when it is centred on a dominance that is static in frequency, such as a pure tone. Receptive fields tuned to particular rates and directions of frequency transition are obtained by rotating $g(t,f)$ in the time-frequency plane. The operator

$$R_\theta(t,f)=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}t\\ f\end{pmatrix} \tag{14}$$

rotates a point by θ radians in time and frequency, and thus the receptive field at a particular orientation may be written as

$$g_\theta(t,f) = gR_\theta(t,f). \tag{15}$$

The parameters σ_t and σ_f in Equation (13) determine the width of the receptive field in time and frequency. The choice of these parameters depends on a compromise between time-frequency localization and directional selectivity (Riley, 1989). When $\sigma_t \neq \sigma_f$, the receptive field has good directional selectivity in time-frequency, which gives it an advantage in separating crossing dominances (as might occur when several sounds sources are simultaneously active). However, when $\sigma_t \neq \sigma_f$, optimum localization in time-frequency results, and bends in the trajectory of a dominance are resolved more effectively. Since many environmental sounds change rapidly in frequency (e.g. speech), accurate localization was considered important and therefore σ_t was set to the same value as σ_f.

When considering the absolute width of the receptive field across frequency, it is convenient to assume that the spacing of auditory filters is logarithmic. On a logarithmic scale, frequency transitions which move at the same rate have the same slope, irrespective of the initial and final frequency. Hence, it can be assumed that neurones centred on different characteristic frequencies in the map have receptive fields which occupy the same number of auditory filter channels, and sweep rates can be expressed in convenient units such as oct/s. Clearly, this assumption is an approximation since the auditory filters in the model are spaced in ERB-rate, which is not perfectly logarithmic. This discrepancy would present a problem if the frequency transition map formed a basis for grouping components with common rates of frequency modulation, since dominances moving at the same rate in different frequency regions would not have the same slope. However, the map is used here only to track spectral dominances across time, so the error between ERB-rate and logarithmic spacing was considered acceptable.

The frequency width of the receptive field was determined by practical considerations. Receptive fields that are wide in frequency do not localize spectral peaks as accurately as receptive fields which are narrow in frequency. Conversely, the narrowness of the receptive field in frequency is limited by the number of auditory filters used in the model. For a filterbank with 128 channels in the range 50 Hz–5 kHz, a frequency spread of seven channels (approximately 1·4 ERB) was found to be a good compromise.

The absolute width in time of the receptive field should be at least as wide as the longest fundamental period expected for a periodic source, otherwise the map will be integrating auditory nerve activity over an uneven temporal window. Since the lowest fundamental frequency expected is 50 Hz, this suggests a lower limit of 20 ms for the time width. Additionally, Nabelek and Hirsch (1969) have found that the ability of listeners to discriminate between different rates of frequency transition is optimal for sweep durations of 30 ms. Consequently, the time width of the receptive field was set to 30 ms in the model. Since the frequency spread was seven auditory filter channels, 30 ms of auditory nerve firings were collapsed into seven bins in order to give the receptive field an equal width in time and frequency.

Since the map is intended to detect frequency transitions in many types of environmental sounds, it is necessary to know the maximum rate at which sweeps in frequency are likely to occur. Some of the most rapid changes in frequency are observed in formant transitions of speech, the majority of which occur at rates of less than 20 oct/s (Lehiste & Peterson, 1961). Accordingly, neurones were tuned to a maximum upward transition rate of 20 oct/s, and a maximum downward transition rate of -20 oct/s.

A final consideration is the distribution of receptive fields across frequency. Here,

the spacing of receptive fields is determined by deriving a tuning curve for the receptive field $g(t,f)$, which quantifies its selectivity to different rates of frequency sweep. Riley (1989) has shown that the tuning curve for $g(t,f)$ is given by

$$\Gamma(\varphi, \zeta) \propto \frac{\cos^2\varphi}{\sqrt{1+(\zeta^2-1)\sin^2\varphi}} \tag{16}$$

where

$$\xi = \frac{\sigma_t}{\sigma_f} \tag{17}$$

and φ is the slope of a pure tone rising linearly in log frequency. Since σ_t and σ_f are equal in the model, ξ is unity and Equation (16) reduces to

$$\Gamma(\varphi) = \cos^2\varphi. \tag{18}$$

Here, receptive fields are spaced so that their tuning curves overlap at their 3 dB points. This corresponds to a spacing of 1·82 oct/s between the preferred sweep rate of neurones in the map.

3.5. Example representations

Although the map is intended to track peaks in the auditory nerve response, it actually measures the rate of frequency transition at every characteristic frequency. Spectral peaks can be located in the map by looking for maxima in response along the frequency axis when $\theta = 0$. Formally, spectral peaks in the map occur at characteristic frequencies which satisfy the condition

$$\frac{\partial}{\partial f} s(t,f,0) = 0. \tag{19}$$

This technique for identifying spectral peaks is generally very reliable, since $g(t,f)$ is sufficiently wide to ensure that a moving dominance generates activity in the map along the line where $\theta = 0$. The direction in which a dominance is moving is determined by locating the maximum along the sweep rate axis of the map, at the characteristic frequency of the spectral peak. Hence, the condition

$$\frac{\partial}{\partial \theta} s(t,f,\theta) = 0 \tag{20}$$

identifies the rate of frequency transition at a particular characteristic frequency.

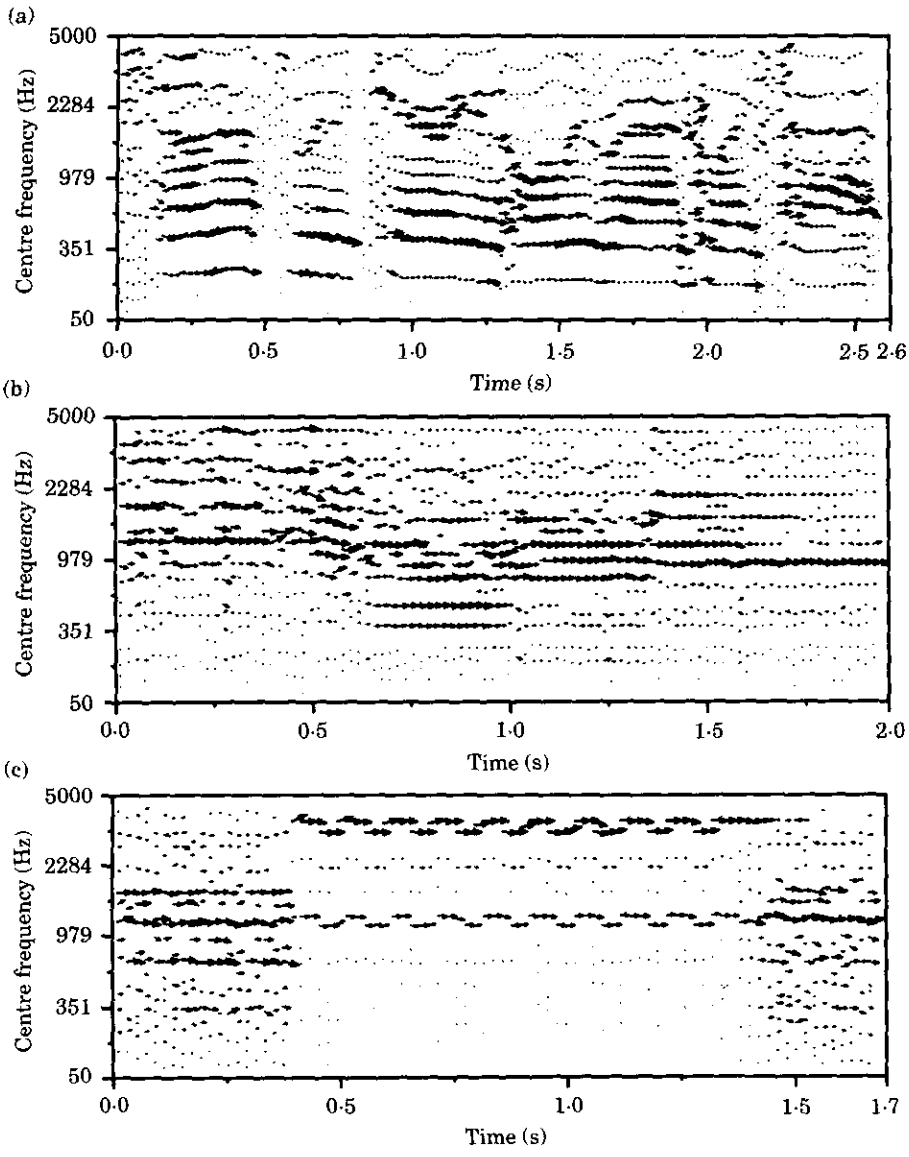

Figure 6. Frequency transition map representations of female speech (a), rock music (b) and trill telephone (c). Vector size is related to amplitude, and direction is related to sweep rate.

Fig. 6 shows the position and orientation of spectral dominances in three sound sources, identified by applying the conditions given in Equations (19) and (20) to the frequency transition map at 10 ms intervals. Spectral peaks are represented by vectors, which have a size related to the amplitude of the peak in the map and a direction related to the rate of frequency transition. Clearly, the map provides primitive information that could be used to track spectral peaks across time using a trajectory principle.

3.6. Onset and offset maps

3.6.1. Motivation

In normal listening situations, it is unlikely that independent sound sources will start and end at the same time. There is good evidence that the auditory system exploits this fact by grouping together acoustic components which have the same onset and offset times. For example, it has been demonstrated that a harmonic which starts before or ends after the other components of a synthetic vowel contributes less to the vowel percept than a synchronous harmonic (Darwin, 1984).

3.6.2. Model description

Cells which respond with a brief burst of activity at the onset or offset of a tonal stimulus are found throughout the higher auditory nuclei. One possible mechanism of these cells would be an excitatory input to the cell at the start of the stimulus, followed by a strong inhibitory input which prevents activity throughout the remaining stimulation (Shofner & Young, 1985). This mechanism can be approximated by writing the membrane potential $p_{on}(t)$ as a leaky sum of the excitatory and inhibitory inputs to the cell,

$$p_{on}(t) = p_{on}(t-1)c_d + E_{psp}r(t) - I_{psp}r(t-\Delta t_I) \tag{21}$$

where E_{psp} and I_{psp} are the excitatory and inhibitory inputs respectively, Δt_I determines the time before inhibition, and the decay constant c_d is given by

$$c_d = \exp\left(-\frac{dt}{\tau_d}\right). \tag{22}$$

The firing rate of the onset cell, $s_{on}(t)$, is determined by the value of the membrane potential when it exceeds a threshold Th,

$$s_{on}(t) = \begin{cases} p_{on}(t) & \text{if } p_{on}(t) > Th \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

The input r to the cell is the envelope of the auditory nerve response at the characteristic frequency of the onset cell, obtained by integrating the output of the Meddis hair cell model over 20 ms with a Hamming window. A wide smoothing window is used in order to remove amplitude fluctuations accompanying the glottal pulses of speech stimuli. In order to model strong inhibition following excitation, the value of the delayed inhibitory input I_{psp} (1·01) is set larger than the value of the excitatory input E_{psp} (1·00).

The parameter τ_d determines the time taken for the membrane potential to decay to 1/e of its maximum deviation from the resting level. Onset cells are able to fire on every click in a pulse train at rates of up to 400–700 clicks/s, which suggests that the membrane can reset within a few ms of firing. Thus, τ_d was set to a short value (1·5 ms) in the model.

The firing threshold Th is set depending on the inhibitory delay of the onset cell

model. Cells with short inhibitory delays (1–5 ms) only produce positive output at abrupt onsets, so in these cases *Th* can be set to zero. For cells with longer delays, *Th* can be increased to remove activity caused by small fluctuations in amplitude over large time intervals. In the examples shown here, *Th* is set to zero.

The time delay before inhibition Δt_I determines the rate of amplitude change that the cell is sensitive to. When the inhibitory delay is short, the model will be sensitive to rapid increases in amplitude but will respond less vigorously to a stimulus with a slow rise time. Here, Δt_I is set to 5 ms in order to detect abrupt amplitude changes.

Intuitively, detecting an offset is rather like the "reverse" of detecting an onset, which suggests that offset cells may receive their excitatory and inhibitory inputs in the opposite order to onset cells. Hence, the membrane potential $p_{off}(t)$ for an offset cell can be written as

$$p_{off}(t) = p_{off}(t-1)c_d + E_{psp}r(t-\Delta t_E) - I_{psp}r(t) \tag{24}$$

where excitation is now delayed relative to inhibition. Similarly, the firing rate of the offset cell is given by

$$s_{off}(t) = \begin{cases} p_{off}(t) & \text{for } p_{off}(t) > Th \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

where the delay before excitation Δt_E is 5 ms. The remaining parameters have the same values as those in Equations (21)–(23).

3.6.3. Example representations

Fig. 7 shows onset map representations for three sound sources. The positions of onsets are identified by bands of brief firing activity extending across frequency. Representations derived from the offset map identify the cessation of a sound source in a similar manner.

4. Auditory elements

4.1. Introduction

Thus far, the auditory representation of an acoustic source has been expressed in terms of the activity in separate neural maps over time. Clearly, a representation of auditory activity is required which combines the information from the different maps, and is amenable to the application of grouping principles in a scene analysis strategy.

An important issue to be considered here is the representation of time. The majority of auditory models described in the literature employ a frame-based representation of time, in which the neural activity across characteristic frequencies is coded as a one-dimensional vector of coefficients at regular time intervals. In many cases, this strategy is adopted because the output of the auditory model is required in a form that is compatible with frame-based automatic speech recognition system (e.g. Beet, 1990). Similarly, frame-based representations of time have been used in the majority of systems which attempt to segregate simultaneous sounds, principally because of the influence of speech processing techniques (Parsons, 1976; Scheffers, 1983; Stubbs & Summerfield, 1990; Varga & Moore, 1990).

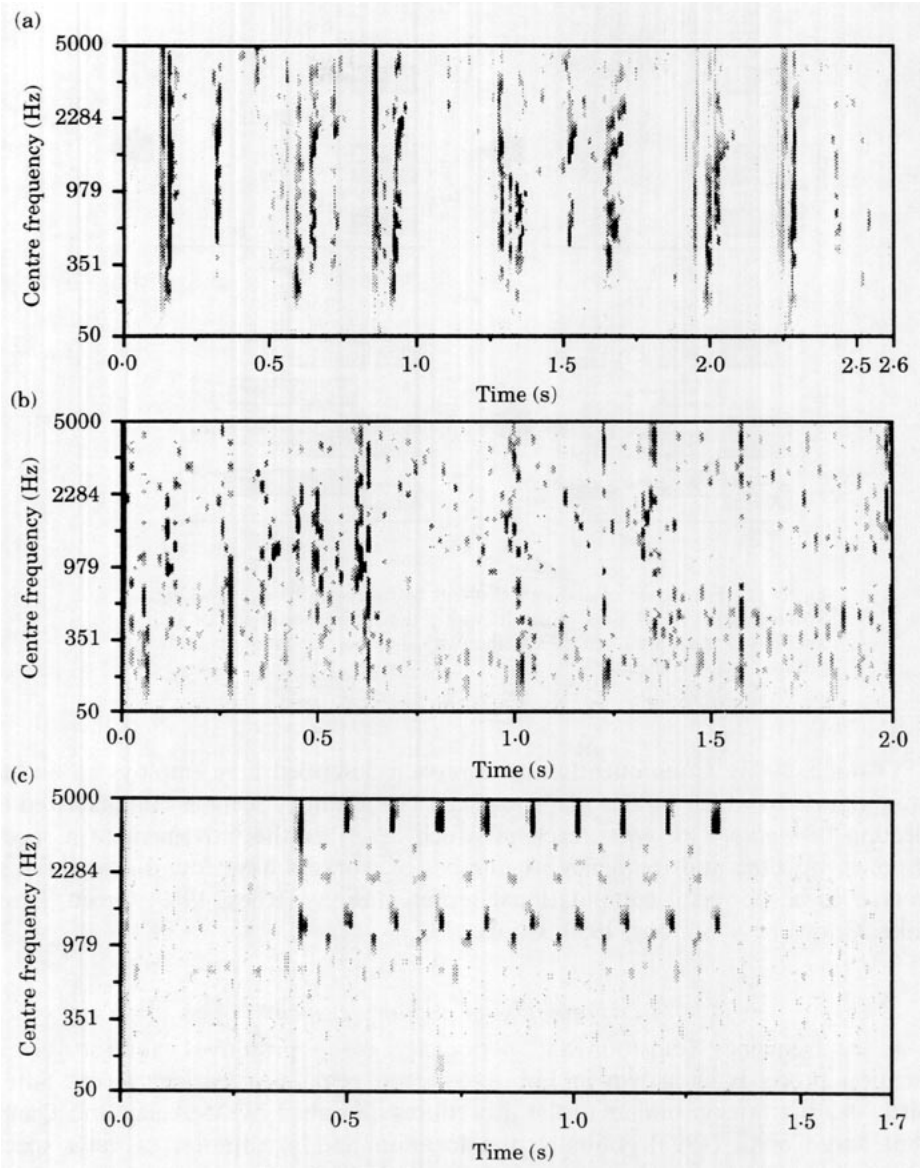

Figure 7. Onset map representations of female speech (a), music (b) and trill telephone (c). Dark areas of the image indicate regions of intense firing activity.

Although frame-based auditory representations provide a good *visual* description of acoustic events, they are inadequate as a basis for ASA algorithms. Specifically, they do not contain explicit information about the way in which the acoustic components vary across time. Examination of Fig. 2 suggests that time is an intrinsic dimension of the auditory rate map representation—visually, we see a two-dimensional time-frequency surface rather than a series of one-dimensional spectra. This observation has been made by a number of workers (e.g. Riley, 1989; McAulay & Quatieri, 1986; Heinbach,

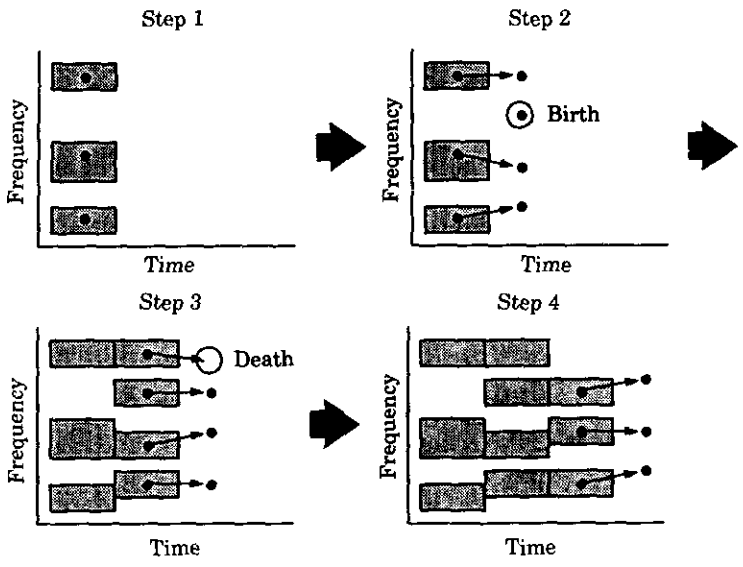

Figure 8. Formation of auditory elements by birth–death peak tracking. Spectral peaks (black dots) which lie within the acceptance region of an auditory element are recruited together with their corresponding periodicity groups (grey rectangles).

1988; Cooke, 1993). Consequently, the approach described here employs an auditory representation in which time is made explicit. The auditory scene is characterized as a collection of *auditory elements*, each of which describes the movement of a spectral component in time and frequency. A number of workers have found descriptions of this type to be powerful computational representations (Riley, 1989; Green, Brown, Cooke, Crawford & Simons, 1990; Cooke, 1993).

4.2. Formation of auditory elements

Given the frequency transition and periodicity group primitives, auditory element formation proceeds as shown in Fig. 8. Spectral peaks are tracked across time by a birth–death strategy, similar to the procedures described by McAulay and Quatieri (1986) and Cooke (1993). Initially, the location and orientation of each spectral peak in a particular time frame are derived by finding the maxima in the frequency transition map, as described in Section 3.5. Subsequently, the movement of a peak at time frame t to a new frequency channel f at the next time frame $t+1$ is predicted by a simple linear extrapolation of the peak's orientation. In practice, it is desirable to allow some tolerance in the predicted position of the peak, so an acceptance region $\omega(f)$ is computed which is centred on f. Formation of an auditory element then proceeds according to the following three rules:

Rule 1. For an existing element at time frame t, a peak that lies within the acceptance region $\omega(t)$ at time frame $t+1$ is recruited to the element. If the recruited peak falls within the boundaries of a periodicity group for time frame $t+1$, then the frequency spread of the element at that time is taken as the width of the periodicity group.

Otherwise, the element is assumed to occupy a single channel of the filterbank at time $t+1$ (steps 2, 3 and 4 in the figure).

Rule 2. If an existing element at time t is unable to recruit a new peak at time $t+1$, the element "dies" (step 3 in the figure).

Rule 3. Peaks at time $t+1$ which do not fall within the acceptance region of an existing element are "born" as new elements. Periodicity groups are matched to the new element as described in the first rule (step 2 of the figure).

The use of an acceptance region around the predicted location of a peak is consistent with the findings of Ciocca and Bregman (1987). They found that when listeners were asked to judge the continuity of a glide through a band of noise, listeners tolerated a disparity in the starting frequency of the post-noise glide. Unfortunately, Ciocca and Bregman did not quantify the width of the acceptance region for different glide slopes, so their data cannot be used to calibrate our system. Rather, the width of $\omega(f)$ was derived empirically. A tolerance of one channel either side of the predicted peak position (corresponding to a $\omega(f)$ of 0·6 ERB) was found to be suitable. Wider acceptance regions tended to produce longer elements, but increased the number of tracking errors (e.g. joining two components with different harmonic numbers).

Not all of the auditory elements are retained for further processing. Specifically, elements that span fewer than two time frames are eliminated. Very short auditory elements are unlikely to have a significant acoustic correlate, and removing them eases the computational burden on the subsequent scene analysis strategy.

4.3. Auditory element representations

The auditory element representations of three sound sources are shown in Fig. 9. Each grey shape is a symbol which traces the path of a spectral dominance through time and frequency. Individual harmonics and formants of speech are generally represented as a single element.

5. Grouping auditory elements

Now that the auditory scene is represented as a collection of symbolic auditory elements, the scene analysis process can be phrased as the problem of finding elements that are likely to have originated from the same acoustic source. Such elements can be identified by exploiting the Gestalt principle of *common fate*. The Gestalt psychologists (e.g., Koffka, 1936) described many principles of perceptual organization which, although generally described first in relation to vision, are also applicable to audition. The term "common fate" describes the tendency to group sensory elements which change in the same way at the same time. In our segregation system, auditory elements are grouped if they have a common F0 or a common onset or offset time. These perceptual grouping cues have been documented extensively in the psychophysical literature (see Sections 3.2.1 and 3.6.1).

6. Grouping by common fundamental frequency

A strategy for segregating concurrent periodic sounds is now described, which partitions the channels of the autocorrelation map into groups that are likely to have the same

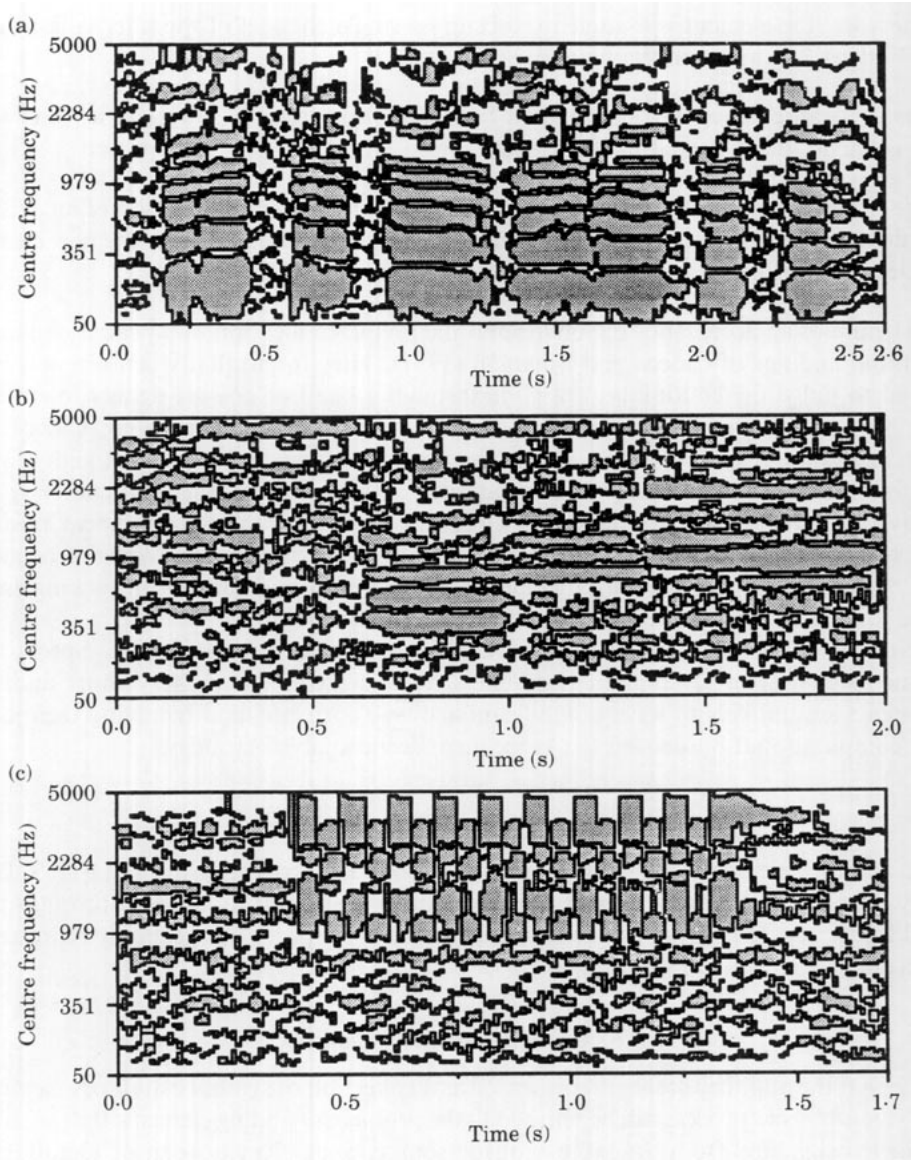

Figure 9. Auditory element representations of speech (a), music (b) and trill telephone (c).

F0. Before proceeding, some previous autocorrelation-based approaches to source segregation are considered.

6.1. Frame-based segregation schemes

Assmann and Summerfield (A&S) (1990) and Meddis and Hewitt (M&H) (1992) have proposed autocorrelation-based segregation strategies, which attempt to model the perceptual processes underlying the ability of human listeners to identify concurrent

vowels with different F0s. As such, they are limited to processing static sounds, and operate on a single frame of an autocorrelation map. However, the strategies could equally be applied to successive frames of a map in order to segregate time-varying stimuli. Weintraub (1985) describes a segregation system based on this principle.

A&S propose several schemes for segregating double vowels. Their "non-linear place-time" model is considered here, which employs an autocorrelation map similar to that described in Equations (7) and (8). Initially, a *summary autocorrelation function* is formed by averaging the channel autocorrelation functions across frequency,

$$s(t, \Delta t) = \frac{1}{M} \sum_{f=1}^{M} a_n(t, f, \Delta t) \tag{26}$$

where M is the number of auditory filter channels. The two largest peaks in the summary are identified, and the delays at which these peaks occur are assumed to correspond to the fundamental periods of the two vowels. Subsequently, the spectrum of each vowel is estimated by sampling the channels of the autocorrelation map at the delay corresponding to the vowel's period. Hence, two "synchrony spectra" are obtained, which indicate the degree of synchronization to each vowel in the auditory nerve.

An alternative strategy has been proposed by M&H. Given that there are two vowels present with different F0s, the M&H scheme partitions the autocorrelation map into two mutually exclusive sets of channels. Initially, the largest peak in the summary autocorrelation is identified, and this is taken to be the period of the dominant vowel. Channels with a peak in their autocorrelation functions at this delay are removed from the map, and the remaining channels are assumed to belong to the second vowel.

Although both of these schemes provide a good match to listeners' responses for vowel segregation tasks, they suffer from potential problems as algorithms for the segregation of arbitrary concurrent sounds. In particular, both the A&S and M&H strategies require *a priori* knowledge of the number of sound sources that are present. Additionally, the A&S and M&H segregation strategies assume that the F0 of a source is identified first, and then this F0 is used to group the components of the source together. However, the work of Darwin and Ciocca (1992) suggests that mechanisms of pitch perception must take into account the temporal history of the components of a harmonic complex, in order to exclude those that differ in onset time. Hence, it appears that perceptual grouping determines perceived pitch, rather than *vice versa*.

6.2. Principles of the new strategy

A new autocorrelation-based segregation strategy is now presented which avoids many of the limitations of the A&S and M&H schemes. In particular, it exploits the fact that temporal continuity has been made explicit in the auditory element representation [see Brown and Cooke (1992) for further details].

Our strategy differs from the other segregation schemes in that it identifies a "local" F0 contour for each element in the auditory scene. Subsequently, elements are grouped if their F0 contours are similar. In contrast, the A&S and M&H schemes attempt to partition the energy in the autocorrelation map using "global" pitch information derived from the summary autocorrelation function.

The summary autocorrelation of a periodic sound has peaks at integer multiples of

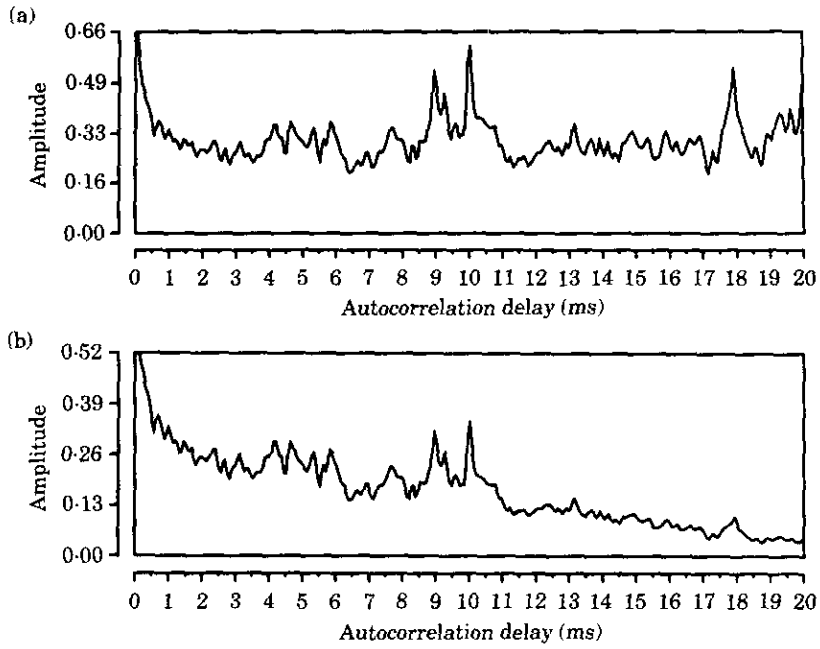

Figure 10. Summary autocorrelation functions for a mixture of the synthetic vowels /a/ (FO 112 Hz) and /ɛ/ (FO 100 Hz), before weighting (a) and after weighting (b).

the fundamental period, as well as a peak at the fundamental period itself. In order to reduce the influence of these "false" peaks on the segregation strategy described here, a weighting is applied to the summary autocorrelation which attenuates peaks at longer delay times. Specifically, a modified summary autocorrelation

$$s_w(t,\Delta t)=\frac{w(\Delta t)}{M}\sum_{f=1}^{M}a_n(t,f,\Delta t) \tag{27}$$

is computed, where the weighting function $w(\Delta t)$ is defined by

$$w(\Delta t)=1{\cdot}0-0{\cdot}9\frac{\Delta t}{\Delta t_{max}} \tag{28}$$

as suggested by Weintraub (1985). Here, Δt_{max} is the longest autocorrelation delay, and the other parameters are defined in Section 3.2.2. The function $w(\Delta t)$ imposes a linear weighting on the summary autocorrelation, which varies from 1·0 at zero delay to 0·1 at the longest delay. This ensures that the peak at the period of a source is larger than the peaks at integer multiples of the period. For example, Fig. 10 shows summary autocorrelation functions for a mixture of the synthetic vowels /a/ (F0 112 Hz) and /ɛ/ (F0 100 Hz). A peak occurs in the summary autocorrelation at the period of each vowel (8·93 and 10·0 ms), and also at twice the period of the /a/ (17·86 ms). After weighting, this spurious peak has been attenuated.

The next stage in the algorithm exploits the fact that auditory elements generally occupy more than one channel of the autocorrelation map at each time frame, since they have been derived by tracking periodicity groups across time. Specifically, a local summary autocorrelation is computed, which averages the channel autocorrelation functions over the frequency spread of the element. For an auditory element which occupies channels f_1 to f_2 of the autocorrelation map at time t, the local summary autocorrelation l is given by

$$l(t,f_1,f_2,\Delta t)=\frac{1}{f_2-f_1+1}\sum_{f=f_1}^{f_2}a_n(t,f,\Delta t). \tag{29}$$

Note that the effect of this averaging will be small, since the channels occupied by an element at a particular time frame are, by definition, very similar.

The weighted summary autocorrelation $s_w(t,\Delta t)$ is an average measure of the periodicities present in the autocorrelation map. As such, it indicates the likelihood of a period Δt occurring in the map at time t. Similarly, the local summary autocorrelation functions $l(t,f_1,f_2,\Delta t)$ indicate the likelihood of a particular period occurring in a channel of the map. Therefore, the product of these two quantities gives an estimate of the probability[1] that the response of a channel f is dominated by a source with period Δt at time t,

$$Pr(t,f_1,f_2,\Delta t)=l(t,f_1,f_2,\Delta t)s_w(t,\Delta t). \tag{30}$$

From Equation (30), it is possible to predict the period of the source that a channel is most likely to be dominated by. Specifically, the predicted period is given by the autocorrelation delay at which $Pr(t,f_1,f_2,\Delta t)$ is highest. Although Pr could be computed in a frame-by-frame manner, such an approach would not take advantage of temporal continuity. Rather, Pr is computed at every time frame occupied by the auditory element, and the best path through this series of functions is found by a dynamic programming algorithm (Cooper & Cooper, 1981). Since the optimum F0 contour passes through peaks in Pr, the dynamic programming algorithm actually finds the best path through the series of functions

$$m(t,\Delta t)=\begin{cases}Pr(t,f_1,f_2,\Delta t) & \text{if } \dfrac{\partial}{\partial\Delta t}Pr(t,f_1,f_2,\Delta t)=0.\\ 0 & \text{otherwise}\end{cases} \tag{31}$$

Here, $m(t,\Delta t)$ is zero except at delays where a local maximum occurs. Equation (31) is computed by using a finite difference approximation to the differential, checking the size of zero crossings to ensure that a maximum has been found rather than a minimum.

The dynamic programming algorithm proceeds as follows. The dynamic programming score $ds(t,\Delta t)$ for a period Δt at time frame t is defined as the dynamic programming score at the previous frame, plus the transition score gained by moving to the current period. Formally, the recursive relation

[1] Note that the term "probability" is used loosely. $Pr(t,f,\Delta t)$ is not a true probability since, in general, $\sum_{\Delta t} \Pr(t,f,\Delta t)\neq 1$.

$$ds(t,\Delta t)=\begin{cases} ds(t-1,\Delta t_p)+\max\limits_{\Delta t} ts(\Delta t_p,\Delta t,t) & \text{if } t_s<t\leqslant t_e \\ 0 & \text{if } t=t_s \end{cases} \tag{32}$$

is calculated for each time frame t between the start time t_s and end time t_e of the auditory element, for values of Δt between 2 and 20 ms (corresponding to pitches in the range 50–500 Hz). The transition score $ts(\Delta t_p,\Delta t,t)$ quantifies the cost of moving from a period Δt_p in the previous frame to a period Δt in the current frame, and is given by

$$ts(\Delta t_p,\Delta t,t)=m(t,\Delta t)\exp\left(-\frac{(\Delta t-\Delta t_p)^2}{2\delta_t^2}\right). \tag{33}$$

Hence, the transition score for a new period depends upon its probability, and its distance from the previous period. The exponential term in Equation (33) applies a Gaussian weighting to the difference in period, so that smaller changes in period give a higher transition score. In the absence of any experimental data, the standard deviation δ_t of the Gaussian was derived empirically. A value of 0·6 ms was found to give good results.

A dynamic programming score $ds(t,\Delta t)$ is computed for each initial period Δt at time t_s, and the period with the highest score is taken to be the start of the best path. Subsequently, the best path is retraced through the series of functions $m(t,\Delta t)$ in order to determine the F0 contour. This process is repeated for each element in the auditory scene, as shown in the left panel of Fig. 11. Here, the F0 contours have been derived for each element in a mixture of speech and a synthetic siren. The contours cluster into two distinct groups, corresponding to the F0s of the two sources. Additionally, a small number of contours occur at twice the fundamental period of the speech, which are due to suboctave errors in the tracking procedure.

Given a predicted F0 contour for each element in the auditory scene, segregation can now be achieved by application of the following grouping principle:

> Auditory elements which overlap in time are grouped together if their predicted F0 contours are sufficiently similar.

For two elements that overlap in time, the similarity of their F0 contours $p_1(t)$ and $p_2(t)$ can be quantified by the metric

$$sim(p_1,p_2)=\frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2}\exp\left(\frac{-[p_1(t)+p_2(t)]^2}{2\delta_p^2}\right). \tag{34}$$

Here, t_1 and t_2 define the first and last time frames at which the two elements overlap. This similarity metric computes the average Gaussian-weighted difference between the two F0 contours. As such, $sim(p_1,p_2)$ varies between unity (identical F0 contours) and zero (very different F0 contours). The standard deviation δ_p of the Gaussian determines the amount of tolerance in the comparison. Here, δ_p was set to 0·3 ms by inspection.

Finally, two elements are allowed to form a group if their $sim(p_1,p_2)$ score exceeds a threshold value. In practice, the F0 contours of elements that belong to the same source tend to be very similar, so the threshold can be set quite high. A value of 0·9 is used here. Clearly, this process groups auditory elements in a pairwise manner. Section 8.2

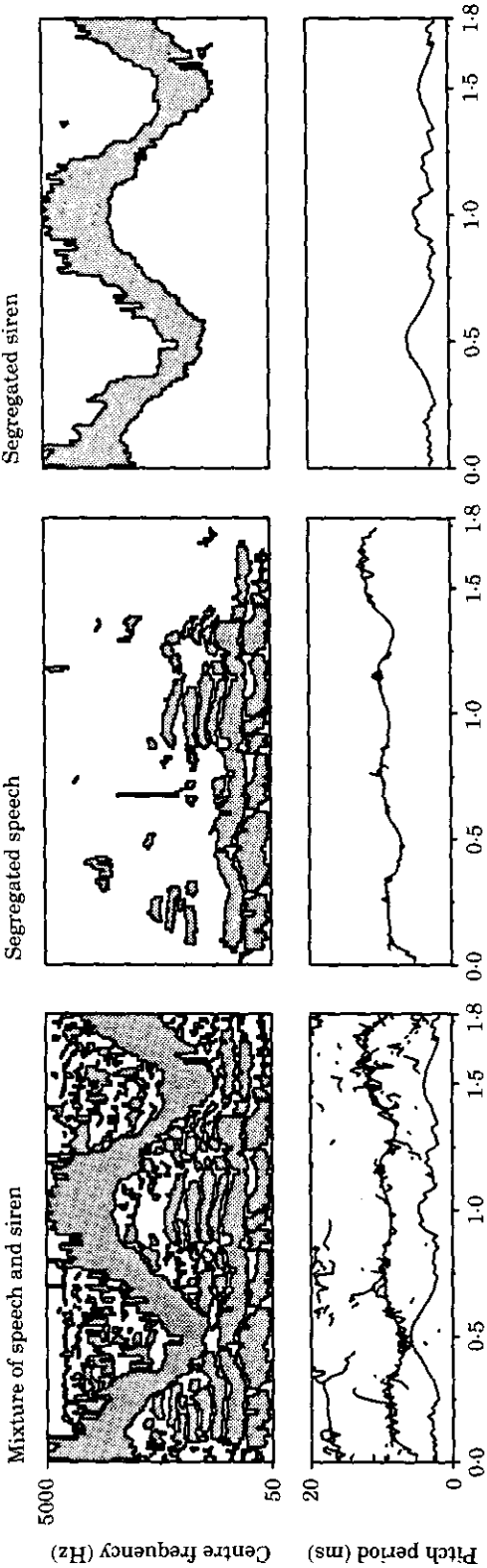

Figure 11. Segregation of speech from a synthetic siren by the system. The original mixture is shown in the left panel. The two groups found by the system are shown in the centre and right panels, which correspond to the speech and siren, respectively. Each panel shows the auditory elements (top) and their F0 contours (bottom).

describes a strategy for searching the auditory scene which forms larger groups from these pairwise comparisons.

This autocorrelation-based approach has a couple of advantages when compared with the A&S and M&H segregation strategies. Firstly, no prior knowledge of the number of sources present in the stimulus is required. Rather, the number of groups that are formed is determined by the number of different predicted fundamental periods. Secondly, the new strategy does not attempt to identify a *global* pitch for each source. Rather, it predicts a *local* pitch for every channel in the map, and groups channels with the same local pitch. This approach is consistent with the view that grouping determines the perceived pitch of a source, rather than *vice versa*.

7. Grouping by common onset and offset

A simple way of grouping by common onset and offset would be to group auditory elements which start and end at the same time. However, auditory elements are formed by a tracking strategy that prefers to break an element rather than make a tracking error. Therefore, the start and end times of an auditory element do not necessarily correspond to the appearance and disappearance of an acoustic event.

The onset and offset maps provide a solution to this problem, since the presence of activity in the maps indicates that an onset or offset of an acoustic event has occurred. Therefore, the following principle can be applied to group elements with a common onset or offset time:

> Auditory elements which start or end synchronously are more likely to form a group, providing that there is sufficient activity in the onset or offset map at the appropriate time.

In practice, elements tend not to be exactly synchronous, so it is desirable to allow a tolerance in the comparison of onset and offset times. Darwin (1984) finds onset and offset segregation effects at disparities of 30 ms, so the tolerance should clearly be less than this. Here, elements are judged to be synchronous if the difference between their start or end times is not more than two time frames, corresponding to a tolerance of 20 ms.

Given the start or end time of an element, the onset or offset map is checked to ensure that an acoustic event has actually started or stopped. Again, it is desirable to allow a tolerance when comparing the start/end time of an element with the time of activity in the onset/offset map. This is because auditory filters tend to ring at their centre frequencies for a few milliseconds after an abrupt onset, which delays the formation of periodicity groups. Similarly, periodicity groups may extend for a few milliseconds after a sudden offset, because the filters continue to ring at the frequency of the stimulus. Therefore, the activity $act(t)$ in the onset or offset map $o(t,f)$ at the start/end time t of an auditory element is quantified by

$$act(t)=\sum_{f=f_1}^{f_2}\sum_{\tau=-2}^{2} o(t+\tau,f). \tag{35}$$

Here, f_1 and f_2 define the range of channels in the filterbank occupied by the element during its first (in the case of onset) or last (in the case of offset) time frame. As before, a two frame (20 ms) tolerance is allowed either side of the start/end time t. An onset or offset is indicated when the activity in the map $act(t)$ exceeds zero.

When two auditory elements start or end at the same time, and an onset or offset is indicated by the maps, the tendency of two elements to group is increased by adding a constant weighting to the similarity score $sim(p_1,p_2)$ defined in Equation (34). The weightings for common onset and common offset are both set to 0·5. Recall that elements are allowed to fuse if their $sim(p_1,p_2)$ score exceeds a threshold value of 0·9. Therefore, elements which have a common onset *and* a common offset will form a group regardless of their F0 contour similarity, since their onset and offset score (1·0) exceeds the threshold. However, elements with a common onset *or* a common offset must also have an F0 contour similarity of at least 0·4, in order to exceed the threshold and form a group. This requirement is consistent with the suggestion of Darwin and Sutherland (1984) that common onset and common offset are neither necessary nor sufficient conditions for grouping the components of speech. In natural speech, formants move rapidly in frequency so that nearby harmonics are amplified and attenuated at different times. Hence, it would be inappropriate to group only those harmonics which are exactly synchronous.

8. Searching the auditory scene

An algorithmic search strategy is now described, which aims to partition the auditory scene into groups of elements that are likely to have arisen from the same environmental event. Similar schemes have been proposed by Cooke (1993) and Mellinger (1991). Currently, the strategy groups elements according to their F0s, onset times and offset times. As such, the algorithm is restricted to *primitive* grouping, and does not attempt to use learned (schema-driven) grouping principles (Bregman, 1990). Additionally, the strategy is limited to searching for simultaneous organization in the auditory scene, and is therefore unable to group elements that are widely separated in time. However, the time-frequency nature of the auditory element representation does allow the sequential propagation of groups in situations where elements overlap.

8.1. Motivation

The issues that arise in formulating a strategy for searching the auditory scene have been comprehensively discussed by Cooke (1993). Here, a new strategy is proposed that is motivated by several of Cooke's observations.

Firstly, the strategy employed here assumes that every element in the auditory scene must be allocated to a group. Hence, the search terminates when all of the elements in the scene have been accounted for. In some cases, a group may consist of a single element.

A second point concerns the allocation of auditory elements between groups. Since the segregation strategy described in Section 6.2 allocates channels of the autocorrelation map exclusively to a single source, an auditory element cannot belong to more than one group. Hence, once an element has been assigned to a group by the search strategy, it is effectively "removed" from the auditory scene. As such, our system applies a "principle of exclusive allocation" (Bregman, 1990).

One potential problem in rigidly applying a principle of exclusive allocation is that the search strategy may find different organizations in the auditory scene if it starts from different elements. For example, a frequency component which could belong to two harmonic series might be assigned arbitrarily to the harmonic series that was identified first by the search strategy. However, the algorithm proposed here does not suffer from this problem, for two reasons. Firstly, elements are grouped according to the similarity of their predicted F0 contours, rather than by harmonicity *per se*. It is very unlikely that the F0 contours of two groups will be so similar that they will compete for the same elements. Secondly, exclusive allocation is not imposed at the level of the search strategy. Rather, it emerges as a consequence of the fact that elements are assigned to a single predicted F0 contour.

In practice, the search time can be reduced by starting from "dominant" elements in the auditory scene. Here, the length of an element is taken as an indication of its dominance, although other properties (such as time-frequency *area*) could also be used. Long elements generally give rise to large groups, and are likely to have a significant acoustic correlate. Therefore, the search for a new group starts with the longest element in the auditory scene, and long elements are recruited to groups before shorter elements.

8.2. *The search strategy*

The algorithm used to search the auditory scene proceeds as follows. Initially, the longest element in the auditory scene is selected as the start of a new group. Then, every element remaining in the scene is considered as a possible match ("focus") to the group. A similarity score $sim(p_1,p_2)$ is calculated between the F0 contour of the focus element and every element in the group that it overlaps in time, as described in Section 6.1. Subsequently, the score is adjusted if the elements being compared have a common onset or a common offset (see Section 7). If the focus element has a $sim(p_1,p_2)$ score greater than 0·9 for every element in the group that it overlaps, it is added to the group. This process iterates until the group cannot recruit any more elements. Then, a new group is started if there are any elements remaining in the auditory scene.

Note that elements are recruited to groups under very tight constraints. Specifically, a focus element can only be recruited to a group if it is sufficiently similar to *all* the members of the group that it overlaps in time. This constraint is imposed to prevent small elements in a group from acting as a "bridge" to dissimilar elements. For example, a focus element which generally has a different F0 contour to a group, but is similar for a short time, could be recruited by a short element in the group that spans the period when the pitch tracks are similar. Checking that the focus element is consistent with every member of the group alleviates this problem.

An example of grouping by the system is shown in Fig. 11. The left panel shows the auditory elements for a mixture of speech and a synthetic siren, together with the F0 contour for each element. After application of the search strategy, two groups of elements have been identified which correspond to the speech (centre panel) and siren (right panel). Examination of the F0 contours for these groups suggests that the two sources have been segregated very effectively.

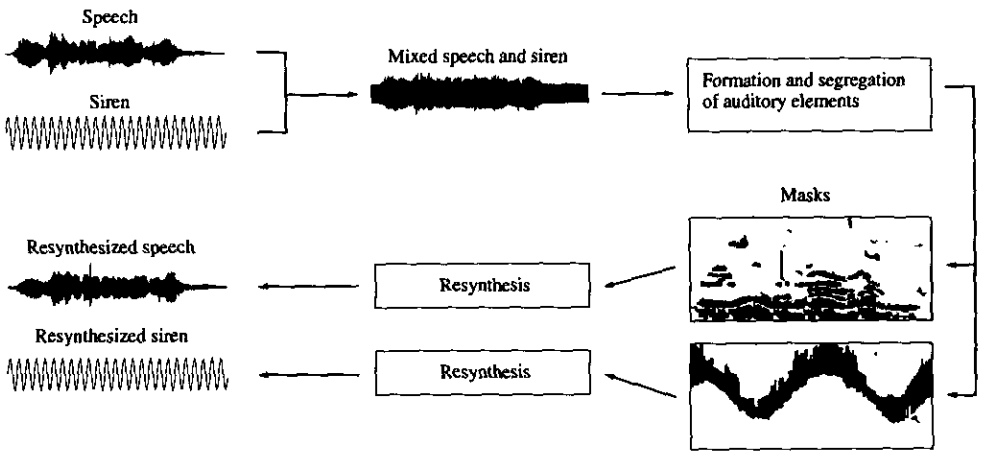

Figure 12. Schematic diagram of the resynthesis process, showing the segregation of speech from a synthetic siren and the resynthesis of each source.

9. Resynthesis

9.1. Motivation

A number of workers have used resynthesis to determine whether an auditory representation preserves perceptually important features of the acoustic input (e.g. Heinbach, 1988; Hukin & Damper, 1989). Resynthesis also provides a convenient means of assessing the performance of systems that attempt to segregate concurrent sounds (Parsons, 1976; Weintraub, 1985; Denbigh & Zhao, 1992; Cooke, 1993). By listening to the segregated output, it is possible to assess how much of the signal has been retained, and how much of the noise intrusion has been rejected.

9.2. Resynthesis from auditory elements

The resynthesis technique employed here is similar to the scheme described by Weintraub (1985). Fig. 12 illustrates the process for a mixture of speech and a siren intrusion, although the technique can be applied to any arbitrary input.

Segregation by the system produces a number of groups of auditory elements. The first stage in resynthesizing a waveform for a group is to form a *mask*. If a channel of the auditory filterbank is occupied by an element in the group at a particular time frame, the value of the mask at that time and channel is unity. Otherwise, the value of the mask is zero. Hence, the mask consists of a matrix of binary weights, that indicate which frequency channels of the filterbank belong to the group at each time frame.

Subsequently, a resynthesized waveform is constructed from the gammatone filter output. In order to remove any across-channel phase differences, the output of each filter is time-reversed, filtered a second time, and time-reversed again. Then, each time-frequency region of the phase-corrected filter output is multiplied by the corresponding weight in the mask. The weights are applied to 20 ms segments of the filter output, which overlap by 10 ms and are windowed with a raised cosine. Finally, the resynthesized waveform is obtained by summing the weighted filter outputs across all channels of the filterbank.

The validity of this resynthesis technique has been confirmed by resynthesizing a signal when every element in the mask is unity, so that all of the time-frequency regions of the filterbank output are included. Speech resynthesized in this way is of very high quality. Additionally, segregated speech obtained after grouping by the system has been resynthesized from each of the 100 mixtures described in Section 10.3. Generally, the resynthesized speech is highly intelligible and quite natural. The best exemplars occur when the noise intrusion is narrowband (1 kHz tone, siren), and the worst occur when the noise is random and wideband (laboratory noise, random noise).

10. Quantitative evaluation

10.1. Motivation

If a resynthesis path is available from a source segregation system, performance can be quantified by assessing the intelligibility of the segregated output in formal listening tests (Hanson & Wong, 1984; Stubbs & Summerfield, 1990). However, this approach may be time-consuming, and subjects require training in order to perform the task. Alternatively, listeners can be replaced in intelligibility tests by an automatic recognizer (e.g. Weintraub, 1985). Unfortunately, interpretation of results may be difficult if an auditory representation is used as an input to the recognizer. For example, Beet (1990) has shown that the output of an auditory model can be an unsuitable input for a conventional speech recognition system.

In the following section, a new evaluation technique is described which allows an SNR to be computed before and after segregation by the system. This evaluation methodology is fast, simple to implement and leads to an easily interpreted metric. Additionally, quoting the performance of the system in terms of an improvement in SNR allows our results to be compared with those of other workers.

10.2. Comparison of SNRs

Generally, the sounds in the test set of mixtures used here are non-stationary. Therefore, a running short-term SNR is computed, which takes the form

$$snr(t) = \frac{2}{\pi}\text{atan}\left(\frac{\sum_{i=0}^{w-1} s^2(t+i)}{\sum_{i=0}^{w-1} n^2(t+i)}\right) \tag{36}$$

where s and n are the speech and noise waveforms, respectively. Here, a 10-ms non-overlapping window of size w samples is used, and results are expressed as the mean $snr(t)$ over every time frame t in the mixture.

Following segregation by the system, all of the noise intrusion n may have been removed within a particular time window. Clearly, this is an ideal result, but it gives rise to an infinite SNR. Hence, an arctangent compression is applied in Equation (36), which ensures that $snr(t)$ is always finite. In practice, this leads to a highly intuitive metric. When there is no signal in the mixture, $snr(t)$ is zero. Similarly, when there is no noise in the mixture, $snr(t)$ is unity. An $snr(t)$ of 0·5 indicates that the levels of signal and noise are equal.

In order to express the performance of the system as an improvement in SNR, it must be possible to obtain separate signal and noise waveforms *after* segregation. This is possible because the resynthesis process is linear, since the gammatone is a linear filter and resynthesis essentially consists of two passes of gammatone filtering. A linear system R satisfies the property of superposition, namely,

$$R(s+n)=R(s)+R(n). \tag{37}$$

Consider the case where the system R represents the resynthesis of a waveform from a mask, and s and n represent the signal and noise, respectively. Equation (37) implies that the proportion of signal in a segregated mixture can be obtained by resynthesizing the signal waveform from the mask, and that the proportion of noise can be obtained by resynthesizing the noise waveform from the mask. Hence, separate signal and noise waveforms can be obtained from a segregated mixture. Furthermore, this technique can be applied to any representation from which a linear resynthesis path is available.

This approach has a number of useful properties. Firstly, it is possible to compute $snr(t)$ after segregation. Secondly, visual examination of the resynthesized speech and noise waveforms indicates how much of the signal has been retained, and how much of the noise has been removed. Finally, it is possible to listen separately to the proportion of signal and proportion of noise in the segregated output. Hence, the degradation of the signal and noise waveforms after segregation can be assessed in informal listening tests.

10.3. Mixture test set

The database of speech and noise mixtures employed by Cooke (1993) has been used as a test set for quantifying the performance of the system. Although the majority of segregation systems have been evaluated using the task of separating speech from other interfering speech (e.g. Parsons, 1976; Hanson & Wong, 1984; Weintraub, 1985), it is clear that a wide variety of noise intrusions occur in natural listening environments. Hence, Cooke's test set contains a range of 10 different noise sources, which include synthetic stimuli (1 kHz tone, random noise) and environmental sounds (music and "office" noise).

Here, our system is evaluated on a set of 100 mixtures, obtained by adding the waveforms of each of the 10 intrusions to each of 10 voiced utterances (five sentences spoken by two male speakers). Fully voiced utterances have been used since our system is not able to sequentially group a stream of voiced–unvoiced speech sounds.

10.4. Results

Each of the 100 mixtures of speech and noise in the test set were processed by the system. The groups corresponding to the speech were identified visually from the auditory element representation, or by listening to the resynthesized waveforms of each group.

Two evaluation metrics are used here. Firstly, the proportions of speech and noise in each group have been derived, allowing the mean $snr(t)$ to be computed after segregation using Equation (36). Similarly, the mean $snr(t)$ has been computed for the original mixture, so that performance can be quantified as an improvement in SNR.

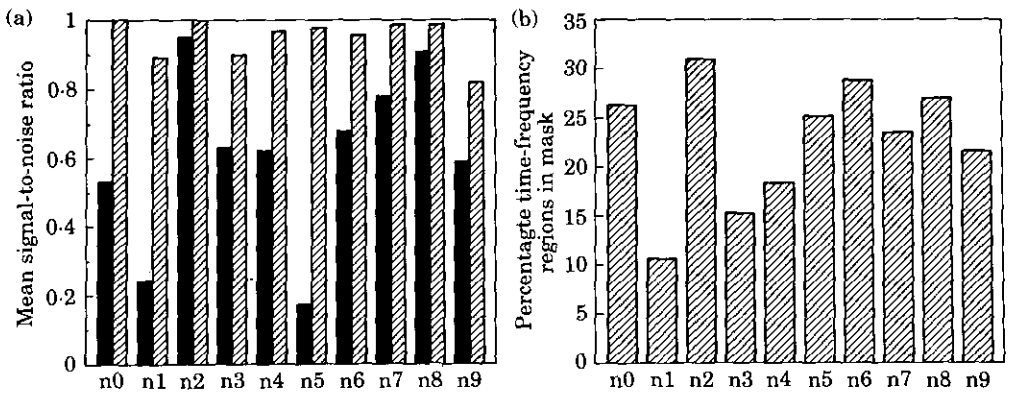

Figure 13. Comparison of signal-to-noise ratio before and after segregation by the system (a). Percentage time-frequency regions allocated to the speech source (b). The intrusions are n0 = 1 kHz tone; n1 = random noise; n2 = noise bursts; n3 = "cocktail party" noise; n4 = rock music; n5 = siren, n6 = trill telephone; n7 = female speech; n8 = male speech; n9 = female speech. (■), Original mixture; (▨), after segregation.

Secondly, the number of non-zero time-frequency regions (TFRs) in the mask is determined for each group. This gives an estimate of how much of the auditory scene has been recovered by the grouping process.

In each of the noise conditions the voiced utterances gave similar results using the SNR and TFR metrics. Hence, the results for each intrusion have been averaged over the 10 utterances.

10.4.1. Segregation by the system using common F0 contour, onset and offset cues

The mean *snr*(*t*) for each noise condition is shown in Fig. 13(a), for the original mixture and for segregated speech obtained after processing by our system using common F0 contour, onset and offset grouping cues. It is apparent that segregation has improved the mean *snr*(*t*) in each case. For some intrusions (n0, n1 and n5) the improvement is very significant. Fig. 13(b) shows the mean number of TFRs allocated to the speech component of the mixture after grouping by our system. About 10–30% of the TFRs in the mask are allocated to the speech, depending on the noise condition. Clearly, more of the TFRs will be allocated to the speech when the intrusion is narrowband (e.g. n0) than when it is broadband (e.g. n1).

10.4.2. Random grouping

The significance of the previous results can be assessed by determining how well a segregation system would perform if it grouped frequency channels randomly at each time frame. Fig. 14 shows the mean values of *snr*(*t*) before and after random grouping. As might be expected, the proportions of signal and noise in a random group are approximately the same as they are in the original mixture. However, small increases in *snr*(*t*) occur with some intrusions (n0, n1, n5 and n9). Comparison with Fig. 13(a) confirms that the performance of our system is better than random grouping for every noise condition.

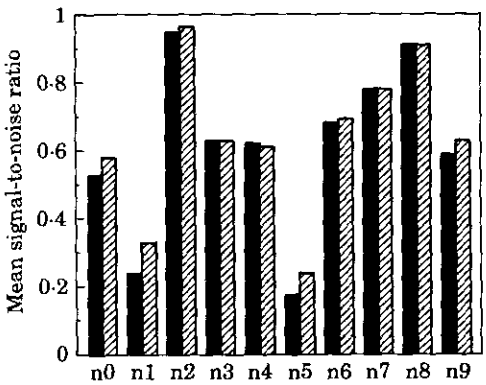

Figure 14. Comparison of signal-to-noise ratio before and after random grouping. The intrusions are n0 = 1 kHz tone; n1 = random noise; n2 = noise bursts; n3 = "cocktail party" noise; n4 = rock music; n5 = siren; n6 = trill telephone; n7 = female speech; n8 = male speech; n9 = female speech. (■), Original mixture; (▨), after random grouping.

10.4.3. Comparison with a frame-based segregation strategy

It is instructive to compare the performance of our system with that of a conventional frame-based autocorrelation segregation strategy. Here, a frame-based strategy similar to the one proposed by Meddis and Hewitt (1992) has been used. Initially, F0 contours were derived for each of the 10 voiced utterances. This was achieved by computing a summary autocorrelation representation for the clean speech, and identifying the location of the largest peak in each time frame. Where necessary, octave errors were manually corrected. Subsequently, these F0 contours were used to guide the segregation of speech from mixtures of speech and noise. Specifically, an autocorrelation map of the mixture was computed at each time frame, and channels of the map which had a peak at the given fundamental period were allocated to the speech source.

Clearly, this approach gives the frame-based strategy an unfair advantage in the comparison, since it has *a priori* knowledge of the fundamental period of the speech at each time frame. Normally, the periods of the two sources would have to be estimated from the summary autocorrelation function of the mixture. The results here assume that this difficult task has been performed without any errors. As such, the results represent the *optimum* performance of a frame-based autocorrelation segregation strategy on the test set.

Fig. 15(a) shows the mean value of $snr(t)$ after segregation, for our system and the frame-based strategy. The performance of our system is better for every intrusion except n9, for which it is the same. In the majority of conditions, our system also recruits more TFRs than the frame-based strategy [Fig. 15(b)]. Generally, therefore, our system is able to recover more of the speech source from the auditory scene. Undoubtedly, the poorer performance of the frame-based autocorrelation strategy arises from the fact that it does not exploit temporal continuity.

11. Summary and discussion

This paper has described a source segregation system which is able to group acoustic components on the basis of their F0 contours and onset/offset times. The novel contributions and limitations of the system are now discussed.

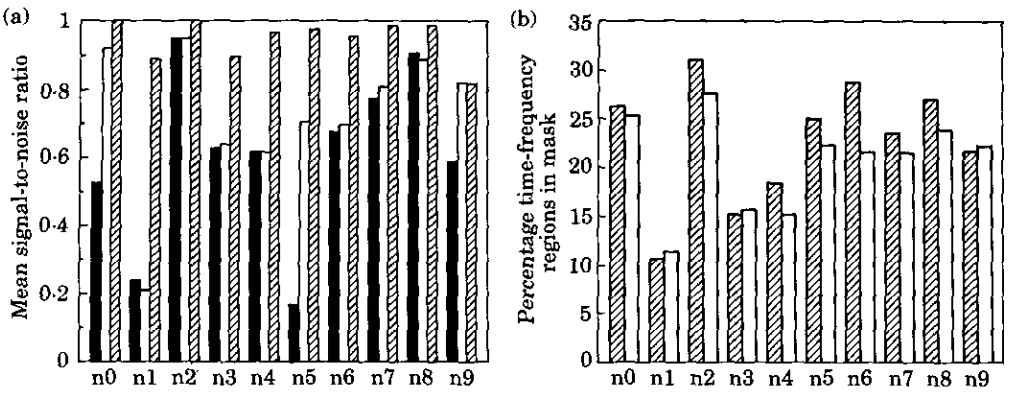

Figure 15. Comparison of signal-to-noise ratio before and after segregration by the system and by a frame-based segregation scheme (a). Comparison of time-frequency regions allocated to the speech source by the system and a frame-based segregation scheme (b). The intrusions are *n0* = 1 kHz tone; *n1* = random noise; n2 = noise bursts; n3 = "cocktail party" noise; n4 = rock music; n5 = siren; n6 = trill telephone; n7 = female speech; n8 = male speech; n9 = female speech. (■), Original mixture; (▨), after segregation; (□), after segregation by a frame-based autocorrelation scheme.

11.1. Signals and symbols in hearing

Perhaps the most significant characteristic of the system is its physiologically-principled, multi-representational view of auditory function. Here, computational models of auditory maps have been employed to provide a rich representational description of the auditory scene. Specifically, the maps extract information about onsets, offsets, frequency transitions and periodicities in different spectral regions. A similar approach has previously been advocated by Darwin (1984).

This approach is similar in concept to the computational approach to vision described by Marr (1982). Marr suggested that the first stage in the description of a visual image should be a rich representation of intensity-level changes, which he called the *primal sketch*. In subsequent stages, a number of processes operate on the primal sketch to identify more abstract levels of structure. Similarly, the auditory maps employed here provide a primitive, but rich, representation of the auditory scene. These primitives form the basis for deriving abstract time-frequency elements, which can be manipulated rapidly and effectively. Hence, auditory maps play a central role in bridging the gap between an acoustic signal and its description as a collection of symbolic auditory elements.

It should be stressed that we are not claiming that auditory elements are actually computed and manipulated at high levels of the auditory pathway. Rather, auditory elements are an abstraction which has been introduced to support a functional description of ASA. The principle advantage of the auditory element representation is the ease with which it allows subsequent grouping algorithms to proceed. For example, there is good evidence that the auditory system is able to correlate patterns of periodicity in widely separated frequency regions in order to identify spectral components that are excited by a common fundamental (Carlyon, Demany & Semal, 1992). Such cross-correlation of filter channels is very expensive in computational terms. However,

auditory elements can be correlated very rapidly, since channels with a similar temporal response have been grouped early in the processing sequence.

11.2. Limitations of the system

A possible limitation of the system arises from the fact that auditory elements are exclusively allocated to one group. When an intrusion is removed from the auditory scene, it leaves a "gap" in the spectrum which can often be heard in the resynthesized waveform. In Cooke and Brown (1993), we describe some preliminary work which addresses this problem by using principles of perceived continuity to extrapolate missing parts of the spectrum.

The system is also limited by the small number of grouping principles that are currently implemented. In particular, the search strategy described in Section 8.2 is unable to group components which are widely separated in time, such as a sequence of speech sounds from a single speaker. Such sequential grouping (Bregman, 1990) is influenced by the timbre, spatial location, temporal proximity, F0 and intensity of successive sounds. Incorporating these cues into the system is a challenging issue for future research.

Currently, only primitive (data-driven) grouping principles are employed in the system. However it is known that listeners are also able to use learned (schema-driven) principles to segregate concurrent sounds. Incorporating schema-driven processing into the system will require a flexible computational framework in which top-down and bottom-up information can influence the groups that are formed. Some preliminary work on this problem, which employs a "blackboard" expert system architecture (Erman & Lesser, 1975), is reported in (Crawford, Cooke & Brown, 1993).

Once formed, auditory elements are not subjected to any further modification in the system. For example, an element cannot be split across time or frequency. A possible limitation of this approach is suggested by an experiment by Darwin and Sutherland (1984). They measured the changes in vowel percept that were caused by adding a tone to the first formant region of a vowel. In one condition, the tone started 30 ms before the vowel. When a harmonic of the leading tone was added which stopped as the vowel started, listeners were more likely to hear a change in the vowel colour. This suggests that the two leading tones formed a separate perceptual group, which ended at the start of the vowel. Currently, the system cannot reproduce this result, since it requires the element representing the leading tone to be broken at the point where the vowel starts. Clearly, this limitation questions the validity of the time-frequency auditory element representation used here. Further research is required to address this issue.

Another limitation of the auditory element representation is that impulsive sounds, such as plosives, are poorly represented (see Fig. 9). Finally, the high computational load of our segregation system currently precludes its use for real-time speech processing tasks.[2]

11.3. Role of common onset and common offset

Grouping by common onset and common offset is subject to tight constraints in the system. Specifically, auditory elements must have some similarity in their F0 contours

[2] The segregation system operates at approximately 4000 times real time, running under UNIX on a SUN SPARCstation 1.

in order for grouping by common onset or common offset to become effective. As a result, common onset and offset cues only contribute to grouping when an auditory element has been excluded from a group because of small irrregularities in its F0 contour (Brown, 1992). In order to reproduce the results of Darwin (1984), speech-specific constraints would have to be included in the system which would allow components of the same voice to be grouped even if they had different onset or offset times.

11.4. Default grouping condition

In the system, it is assumed that elements in the auditory scene are segregated unless there is evidence to group them together. However, segregation may not be the default condition of organization. Rather, the auditory system may prefer to *fuse* all the components in the auditory scene, so that elements are only segregated when there is evidence for doing so (Bregman, 1990). Fusion could be made the default condition in the system by rejecting an element from a group if its $sim(p_1,p_2)$ score with the members of the group was sufficiently low. Whether this approach would have any advantages over the strategy presented here is an issue for further investigation.

11.5. Retroactive effects in grouping

The search algorithm allows auditory elements at a particular time to be recruited to a group that starts at a later time. In fact, there is good evidence that perceptual grouping mechanisms are able to operate retroactively (e.g. Darwin, 1984). However, in our system there is no limit on how far the scene analysis strategy can search back through time. Perceptual grouping mechanisms may actually operate over a temporal window of a few hundred milliseconds.

12. Conclusions

This article has presented a source segregation system which is motivated by the known mechanisms of ASA. The system constructs a symbolic description of the auditory scene, which is searched for acoustic components with common F0 contours, common onset times and common offset times. Components with similar properties are combined into explicit groups, from which a waveform can be obtained by a resynthesis path. The system has been tested on a database of speech mixed with various noise intrusions, with encouraging results. In particular, our results suggest that the use of temporal continuity constraints in the system gives it an advantage in performance over frame-based segregation strategies.

Thanks to Malcolm Crawford for software support, and to Malcolm Slaney and two anonymous reviewers for their incisive comments on an earlier draft of this paper. This work was supported by grant GR/H53174 from the Science and Engineering Research Council Image Interpretation Initiative. GJB thanks the Nuffield Foundation for an equipment grant.

References

Assmann, P. F. & Summerfield, Q. (1990). Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America* **88**, 680–697.

Beauvois, M. W. & Meddis, R. (1991). A computer model of auditory stream segregation. *Quarterly Journal of Experimental Psychology* **43A**, 517–541.

Beet, S. W. (1990). Automatic speech recognition using a reduced auditory representation and position-tolerant discrimination. *Computer Speech and Language* **4**, 17–33.

Boer, E. de & Jongh, H. D. de (1978). On cochlear encoding: potentialities and limitations of the reverse-correlation technique. *Journal of the Acoustical Society of America* **63**, 115–135.

Boer, E. de & Kuyper, P. (1968). Triggered correlation. *IEEE Transactions on Biomedical Engineering* **15**, 169–179.

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

Brown, G. J. (1992). Computational auditory scene analysis: A representational approach. PhD Thesis, University of Sheffield.

Brown, G. J. & Cooke, M. P. (1992). Grouping sound sources using common pitch contours. *Proceedings of the Institute of Acoustics* **14**, 439–446.

Carlson, R. & Granström, B. (1982). Towards an auditory spectrograph. In *The Representation of Speech in the Peripheral Auditory System* (R. Carlson & B. Granström, eds). Elsevier, Amsterdam, Holland.

Carlyon, R. P., Demany, L. & Semal, C. (1992). Detection of across-frequency difference in fundamental frequency. *Journal of the Acoustical Society of America* **91**, 279–292.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America* **25**, 975–979.

Ciocca, V. & Bregman, A. S. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception and Psychophysics* **42**, 476–484.

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

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

Cooper, L. L. & Cooper, M. W. (1981). *Introduction to Dynamic Programming.* Pergamon Press, Elmsford, New York.

Crawford, M. D., Cooke, M. P. & Brown, G. J. (1993). Interactive computational auditory scene analysis: An environment for exploring auditory representations and groups. *Journal of the Acoustical Society of America* **93**, 2308.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America* **76**, 1636–1647.

Darwin, C. J. & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *Journal of the Acoustical Society of America* **91**, 3381–3390.

Darwin, C. J. & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic?. *Quarterly Journal of Experimental Psychology* **36A**, 193–208.

Denbigh, P. N. & Zhao, J. (1992). Pitch extraction and separation of overlapping speech. *Speech Communication* **11**, 119–125.

Deng, L. & Geisler, C. D. (1987). A composite auditory model for processing speech sounds. *Journal of the Acoustical Society of America* **82**, 2001–2012.

Erman, L. D. & Lesser, V. R. (1975). A multi-level organisation for problem solving using many diverse cooperating sources of knowledge. *Proceedings of the International Joint Conference on Artificial Intelligence* **2**, 483–490.

Festen, J. M. & Plomp, R. (1983). Relations between auditory functions in impaired hearing. *Journal of the Acoustical Society of America* **73**, 652–662.

Garofolo, J. S. & Pallet, D. S. (1989). Use of CD-ROM for speech database storage and exchange. *Proceedings of the European Conference on Speech Communication and Technology*, 309–315.

Glasberg, B. R. & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research* **47**, 103–138.

Glass, J. R. & Zue, V. W. (1988). Multi-level acoustic segmentation of continuous speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 429–432.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America* **54**, 1496–1516.

Green, P. D., Brown, G. J., Cooke, M. P., Crawford, M. D. & Simons, A. J. H. (1990). Bridging the gap between signals and symbols in speech recognition. In *Advances in Speech, Hearing and Language Processing Volume 1* (W. Ainsworth, ed.), pp. 149–192. Academic Press, London.

Hanson, B. A. & Wong, D. Y. (1984). The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* **18A**, 5.1–5.4.

Heinbach, W. (1988). Aurally adequate signal representation: The part-tone-time-pattern. *Acustica* **67**, 113–121.

Hukin, R. W. & Damper, R. I. (1989). Testing an auditory model by resynthesis. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)* **1**, 243–246.

King, A. J. & Hutchings, M. E. (1987). Spatial response properties of acoustically responsive neurons in

the superior colliculus of the ferret: A map of auditory space. *Journal of Neurophysiology* **57**, 596–624.
Knudsen, E. I., duLac, S. & Esterley, S. D. (1982). Computational maps in the brain. *Annual Review of Neuroscience* **10**, 41–65.
Koffka, K. (1936). *Principles of Gestalt Psychology*. Harcourt Brace, New York.
Lehiste, I. & Peterson, G. E. (1961). Transitions, glides and diphthongs. *Journal of the Acoustical Society of America* **33**, 268–277.
Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experentia* **7**, 128–134.
Marr, D. (1982). *Vision*. Freeman, San Francisco.
McAulay, R. J. & Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**, 744–754.
Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America* **79**, 702–711.
Meddis, R. (1988). Simulation of auditory–neural transduction: Further studies. *Journal of the Acoustical Society of America* **83**, 1056–1063.
Meddis, R. & Hewitt, M. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. Pitch identification. *Journal of the Acoustical Society of America* **89**, 2866–2882.
Meddis, R. & Hewitt, M. (1992). Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America* **91**, 233–245.
Mellinger, D. K. (1991). Event formation and separation in musical sound. PhD Thesis, Stanford University.
Nabelek, I. & Hirsh, I. J. (1969). On the discrimination of frequency transitions. *Journal of the Acoustical Society of America* **45**, 1510–1519.
Parsons, T. W. (1976). Separation of speech from interfering noise by means of harmonic selection. *Journal of the Acoustical Society of America* **60**, 911–918.
Patterson, R. D., Holdsworth, J., Nimmo-Smith, I. & Rice, P. (1988). SVOS final report, Part B: Implementing a gammatone filter bank. *APU report 2341*.
Plack, C. J. & Moore, B. C. J. (1990). Temporal window shape as a function of frequency and level. *Journal of the Acoustical Society of America* **87**, 2178–2187.
Riley, M. D. (1989). *Speech Time-Frequency Representations*. Kluwer Academic Publishers, Boston.
Scheffers, M. T. M. (1983). Sifting vowels: Auditory pitch analysis and sound segregation. PhD Thesis, University of Gröningen.
Schreiner, C. E. & Langner, G. (1988). Periodicity coding in the inferior colliculus of the cat. II. Topographical organization. *Journal of Neurophysiology* **60**, 1823–1840.
Shamma, S. A., Vranic, S. & Wiser, P. (1992). Spectral gradient columns in primary auditory cortex: Physiological and psychoacoustical correlates. In *Advances in the Biosciences Volume 83*, (Y. Cazals, L. Demany & K. Homer, eds). Pergamon Press, Oxford.
Shofner, W. P. & Young, E. D. (1985). Excitatory/inhibitory response types in the cochlear nucleus: Relationships to discharge patterns and responses to electrical stimulation of the auditory nerve. *Journal of Neurophysiology* **54**, 917–939.
Slaney, M. & Lyon, R. F. (1990). A perceptual pitch detector. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 357–360.
Stubbs, R. J. & Summerfield, Q. (1990). Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America* **87**, 359–372.
Suga, N. & Manabe, T. (1982). Neural basis of amplitude-spectrum representation in auditory cortex of the mustached bat. *Journal of Neurophysiology* **47**, 225–255.
Tougas, Y. & Bregman, A. S. (1985). Crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance* **11**, 788–798.
Ullman, S. (1979). *The Interpretation of Visual Motion*. MIT press, London.
Varga, A. P. & Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 845–848.
Weintraub, M. (1985). A theory and computational model of monaural auditory sounds separation. PhD Thesis, Stanford University.

(*Received 19 November 1993 and accepted for publication 8 March 1994*)