TR－H－101

Auditory Signal Processing for The Segregation of Speech from Interfering Sounds: A Computational Investigation of Spatial Location and Periodicity Cues

Guy J. Brown

1994. 9. 27

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2　☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

AUDITORY SIGNAL PROCESSING FOR THE SEGREGATION OF SPEECH FROM INTERFERING SOUNDS: A COMPUTATIONAL INVESTIGATION OF SPATIAL LOCATION AND PERIODICITY CUES

GUY J. BROWN

Department of Computer Science, University of Sheffield, *
Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom
Internet: g.brown@dcs.shef.ac.uk

ABSTRACT

This paper describes two computational schemes for the segregation of speech from interfering sounds, based on auditory signal processing. In the first scheme, spectral components are grouped according to the similarity of their spatial locations. Deficiencies in this system prompted the development of a second system, in which continuity of fundamental frequency and continuity of spatial location are exploited simultaneously within a novel auditory representation, the *pitch-azimuth-time cube*. Results of a quantitative evaluation of the two schemes on a small data set suggest that the second system has a performance advantage over the first. Implications for the design of source segregation systems and auditory scene analysis theory are discussed.

INTRODUCTION

In most listening situations, a mixture of sounds reaches our ears. However, we are able to attend to a single sound source in these situations, such as the voice of a speaker at a cocktail party or a melody played by an instrument in an orchestra. How is this apparently effortless segregation of sounds achieved?

Albert Bregman's recent book, *Auditory Scene Analysis*, presents a coherent account of the perceptual segregation of sound (Bregman, 1990). He contends that the mixture of sounds reaching the ears is subjected to a two-stage parsing process. First, the acoustic signal is decomposed into a number of sensory components. Second, components that are likely to have arisen from the same environmental event are recombined to form perceptual structures (streams) that can be interpreted by higher-level processes.

Essentially, Bregman's book is an investigation of the information processing problems that are posed by hearing. As such, its approach has some similarities with Marr's (1982) computational theory of vision. Indeed, Bregman's account has stimulated research in machine hearing in much the same way that Marr's computational theory has influenced research in machine vision. Recently, a number of workers have proposed sound segregation systems that implement auditory scene analysis principles (Mellinger, 1991; Baumann, 1992; Kashino & Tanaka, 1992; Cooke, 1993; Cooke & Brown, 1993; Brown & Cooke, 1994a,b). These systems provide computational testbeds for studying mechanisms of auditory grouping, and also have potential applications in robust automatic speech recognition, automatic music transcription and advanced hearing prostheses.

In general, previous segregation systems have only addressed the problem of separating sound components that overlap in time, so-called *simultaneous* grouping (Bregman and Pinker, 1978). Undoubtedly, this bias has arisen because simultaneous grouping involves the computation of relatively simple properties of sound components, such as onset time, offset time and harmonic frequency. However, it is clear that mechanisms are also required to group sounds that are widely separated in time, such as a sequence of voiced and unvoiced sounds from a single speaker. Our recent work has focused on the implementation of these *sequential* grouping

* contact addess

principles, which require the analysis of complex attributes such as spatial location, timbre and rhythm (Brown & Cooke, 1994b; Todd & Brown, 1994).

The current study investigates the role of spatial location cues in a computational system for sound segregation. It is well known that listeners tend to assign sounds that originate from the same location in space to the same stream, and sounds that originate from different locations to different streams; Cherry (1953) noted this phenomenon and called it the 'cocktail party' effect. We should expect, therefore, that similarity of spatial location is a powerful sequential grouping principle (Bregman, 1990).

In the remainder of this paper, two segregation systems are described which adopt different strategies for using spatial location information. In the first scheme, it is assumed that the auditory system is able to directly associate a spatial location with the neural activity within individual frequency bands. Spatial location is therefore the primary grouping cue. In the second scheme, a spatial location is associated with a collection of frequency components that have already been grouped because they have a similar pattern of periodicity. Spatial location is therefore seen as a *derived property* of groups that have been formed by other scene analysis principles. The performance of each scheme is evaluated, and the implications of the results for auditory signal processing theory and the design of source segregation systems are discussed.

EXPERIMENT ONE: LOCATION CUES ONLY

A number of psychophysical studies, notably those investigating the phenomenon of binaural masking level difference, have suggested that the auditory system is able to compute spatial location on a frequency-by-frequency basis (e.g. Webster, 1951). Such frequency-specific comparisons could underly the ability of listeners to segregate the components of concurrent sounds that originate from different locations in space. Bregman (1990) suggests a scene analysis interpretation of this mechanism, which is compatible with a Gestalt principle of grouping by similarity:

> *"...if two parts of the spectrum come from the same source, then they should be coming from the same spatial location and be subject to the same echoes; therefore, the delay between time of arrival in the two ears should be the same for both parts. Working this reasoning backward, if the auditory system receives two spectral regions with the same interaural phase difference, it should fuse them; however, when the two regions show different interaural phase differences, the regions should be segregated..." (page 324)*

This principle has been used in a computational model of binaural segregation described by Lyon (1988). In Lyon's model, the auditory nerve responses of corresponding frequency channels from the two ears are cross-correlated, forming a 'cross-correlogram'. This mechanism is equivalent to the coincidence scheme proposed by Jeffress (1948) for computing interaural time differences. Channels of the cross-correlogram that are dominated by frequency components with the same interaural time difference exhibit a peak at the same correlation delay. Hence, Lyon's model identifies channels with similar peaks and assigns the energy in those channels to the same source.

A Segregation System Using Location Cues Only

A block diagram of a segregation system which uses location cues is shown in Figure 1. The system consists of four processing stages. First, signals recorded from the left and right ears of a head and torso simulator are processed by a model of the auditory periphery. The output of the peripheral auditory model is a probabilistic representation of auditory nerve firing activity for each of 100 filter channels. In the second stage, simulated auditory nerve firings for the left and right ears of the same filter channel are cross-correlated, as in Lyon's (1988) model. Thirdly, the locations of auditory events are estimated from the cross-correlogram, and these locations are tracked across time. Finally, auditory events with similar location-time tracks are allocated to the same source, and a time-frequency mask is constructed that allows the spectrum of the source to be recovered. The following sections describe each stage of the model in detail.

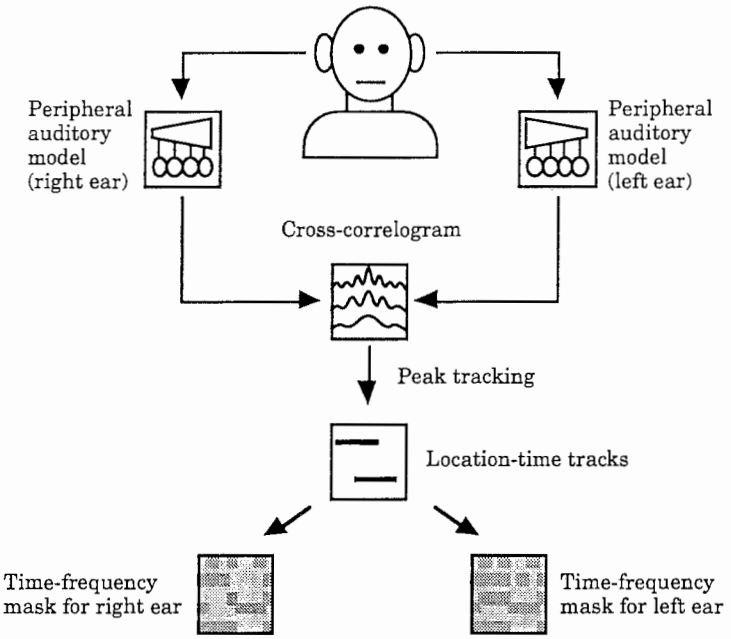

Figure 1: A segregation scheme using location cues only. The location of each auditory event is computed by tracking peaks in a cross-correlogram. Events that have a similar location-time track are allocated to the same source.

Auditory periphery

Peripheral auditory filtering is simulated by a bank of bandpass 'gammatone' filters (Patterson *et al*, 1988), each of which models the response of one point along the basilar membrane. The impulse response of a gammatone filter with order n and centre frequency f_c Hz is

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

where t is time, ϕ is phase and b is related to bandwidth. Here, fourth order filters are used, with centre frequencies distributed according to the equivalent rectangular bandwidth (ERB) scale of Glasberg & Moore (1990). Specifically, 100 filters were spaced equally in ERB-rate between centre frequencies of 50 Hz and 5 kHz, according to the relation

$$E(f) = 21.4\log_{10}(4.37f + 1) \tag{2}$$

where $E(f)$ is the number of ERBs at frequency f kHz. Subsequently, the activity in each filter channel is converted to simulated auditory nerve discharges by the Meddis (1986) model of inner hair cell transduction. The parameters of the model are configured to simulate an auditory nerve fibre with a high spontaneous firing rate (Meddis, 1988).

Cross-correlogram

There is good evidence from physiological and psychophysical studies that a coincidence or cross-correlation mechanism underlies the ability of listeners to detect interaural time differences (e.g. Yin & Chan, 1988; Jeffress, 1948). Here, a running cross-correlation

$$ccg(T, f, \tau_c) = \sum_{t=0}^{T} r(left, f, t)r(right, f, t - \tau_c)\exp\left(\frac{-t}{\Omega_c}\right) \tag{3}$$

is computed between the two ears with a time constant of integration Ω_c of 20 ms. Since interaural delays of more than 2 ms are unlikely to play a role in localisation (Blauert, 1983), cross-correlation functions were computed for values of the time lag τ_c between -2 ms (stimulus leading in the left ear) and +2 ms (stimulus leading in the right ear).

Location analysis

Recent psychophysical studies suggest that although the auditory system is able to compute location on a frequency-by-frequency basis, lateralisation decisions are based on a frequency

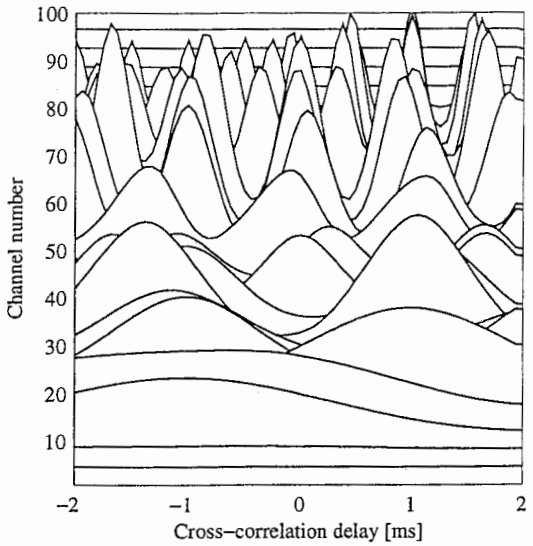

Figure 2: Cross-correlogram of a mixture of two complex tones. For clarity, only the response of every fourth channel is shown.

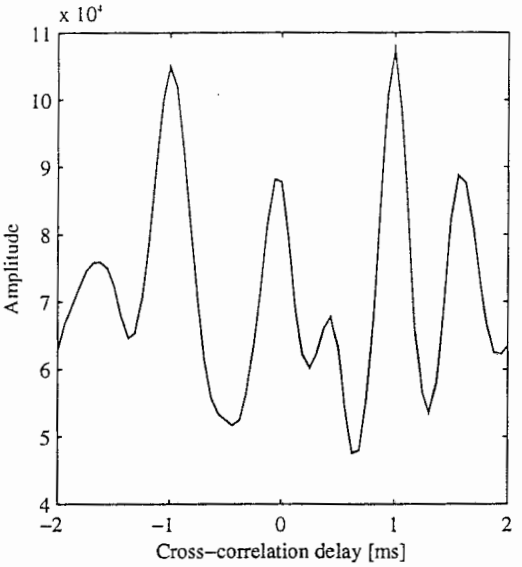

Figure 3: Summary cross-correlogram for a mixture of two complex tones. Peaks occur at correlation delays corresponding to the ITD of each complex (-1 ms and +1 ms).

region wider than a single critical band (Dye, 1990; Trahiotis & Stern, 1989). Indeed, computational models of lateralisation that incorporate across-frequency integration have been shown to give a close fit to psychophysical data (Stern *et al.*, 1988; Shackleton *et al.*, 1992).

Here, information in the cross-correlogram is integrated across frequency by computing a summary cross-correlogram

$$sum_{ccg}(t,\tau_c) = \sum_f ccg(t,f,\tau_c) \tag{4}$$

as suggested by Shackleton *et al.* (1992). A peak in this summary function occurs at a correlation delay corresponding to the interaural time difference of a binaural stimulus. For example, Figure 2 shows a cross-correlogram of a stimulus consisting of two complex tones. One tone has a fundamental frequency of 150 Hz and an ITD of -1 ms, the other has a fundamental frequency of 200 Hz and an ITD of +1 ms. The summary cross-correlogram shown in Figure 3 has a large peak at the correlation delay corresponding to each ITD.

Event formaton

Auditory events are identified by tracking peaks in the summary cross-correlogram through time. A birth-death peak tracking system, similar in principle to that described by Brown & Cooke (1994a), is applied to summary functions computed at 10 ms intervals. This tracking process is governed by the following three rules:

Rule 1: A peak in the summary cross-correlogram at time t and delay τ_1 is recruited to an existing event ending at time t-1 and delay τ_2 if $|\tau_1-\tau_2| < \alpha$. The value of α is not critical, but represents a compromise between many short events (small α) and long events that erroneously incorporate peaks from unrelated events (large α) . Here, an α value of 0.2 ms was used.

Rule 2: Existing events at time t-1 that are unable to recruit a peak at time t are terminated.

Rule 3: Peaks in the summary cross-correlogram that have not been recruited by any existing events become the start of a new event.

In practice, low amplitude peaks in the summary cross-correlogram do not need to be tracked, since these are unlikely to have a significant acoustic correlate. For the task considered here (tracking two sound sources) only the five largest peaks in each frame of the summary cross-correlogram were considered in the tracking process.

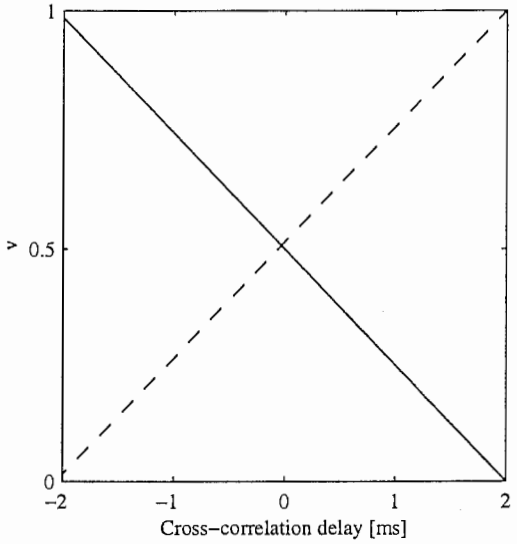

Figure 4: Scaling factors applied to the mask weights for the left ear (solid line) and right ear (dashed line). When the source is positioned on the median plane (zero cross-correlation delay), the contribution of each ear is weighted equally.

Sequential grouping by similarity of spatial location

Cross-correlation delays at each point along a location-time track are averaged over time to give a single estimate of the location of the auditory event. This approach is consistent with the observation that the auditory system is able to compute mean or compromise locations under some conditions, so-called 'summing localisation' (Blauert, 1983). Subsequently, events that have a similar location are allocated to the same source. Events are judged to be similar if there is a difference of less than 0.3 ms between their average cross-correlation delays.

Occasionally, gaps occur in the location-time tracks for a source, either because the source has been masked, or because of errors in the tracking procedure. The cross-correlation delays in these gaps are interpolated by using linear least squares analysis to fit a regression line through the points in the location-time track.

Evaluation

Processing of binaural signals by the system yields a location (cross-correlation delay) for each source at each time frame. The spectrum of a source can be recovered by using the location information to derive an appropriate weighting of the energy in each auditory filter. Initially, the weight $w(\varepsilon,t,f)$ for filter channel f of ear ε at time t is obtained by sampling the cross-correlogram at the correlation delay τ_s of the source:

$$w(\varepsilon,t,f) = v(\varepsilon,\tau_s) * ccg(t,f,\tau_s) \tag{5}$$

The scaling factor $v(\varepsilon,\tau_s)$ takes a value between zero and unity, as shown in Figure 4. This factor ensures that the weight for the left ear is greater when the source location is to the left side of the head, and vice versa. The weights derived in this manner are referred to as a *mask*. If segregation has been successful, the mask will have a high value at time-frequency regions dominated by the target source, and a low value at time-frequency regions dominated by the interfering source.

The performance of the system can be evaluated quantitatively by comparing the signal-to-noise ratio (SNR) before and after segregation (Brown & Cooke, 1994a). Here, the system was evaluated on a small data set of speech (the target source) mixed with one of seven intrusive sounds (the interfering source). Details of the test signals are given in the Appendix. The SNR at time frame t is given by

$$SNR(t) = \frac{2}{\pi}\tan^{-1}\left(\frac{\sum_{\varepsilon=left,right}\sum_{f} w(\varepsilon,t,f) * s(\varepsilon,t,f)}{\sum_{\varepsilon=left,right}\sum_{f} w(\varepsilon,t,f) * n(\varepsilon,t,f)}\right) \tag{6}$$

where $s(\varepsilon,t,f)$ and $n(\varepsilon,t,f)$ represent the RMS energy in ear ε and channel f of the auditory filterbank for the target source and interfering source respectively. This metric takes values between zero (all noise) and unity (all signal). The arctangent compression is employed because the denominator may be zero in some time frames.

The results are shown in Figure 5, expressed as mean $SNR(t)$ over all time frames t. The signal-to-noise ratio of the original mixture is calculated by setting all the weights $w(\varepsilon,t,f)$ in equation 6 to unity. After segregation by the system, there is an increase in mean SNR for each noise condition.

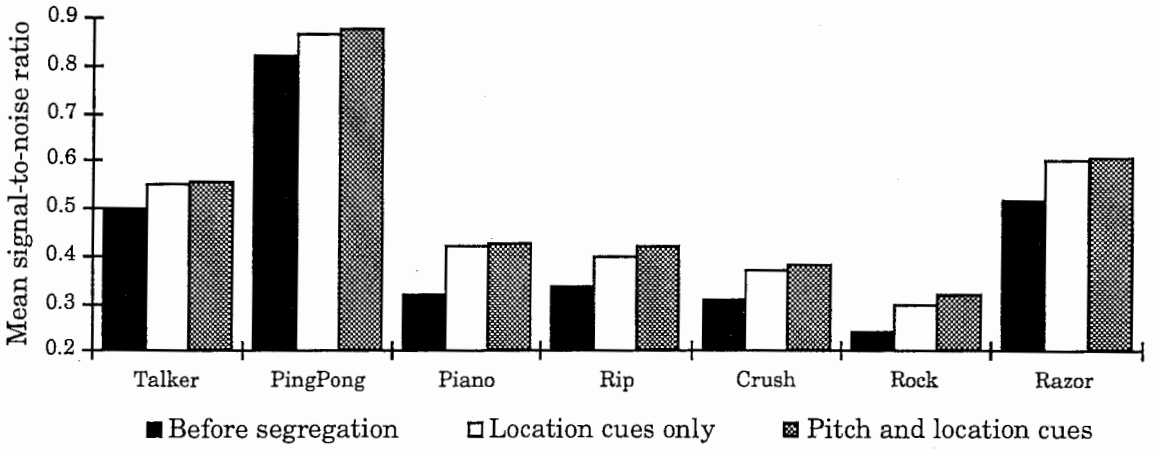

Figure 5: Comparison of mean SNR before and after segregation by two schemes. The details of the pitch and location scheme are given later in the paper.

Evaluation using the SNR metric alone is inadequate, since conditions may occur where only a small proportion of the target source has been recovered but all of the interfering source has been excluded; this would lead to a high SNR, but the reconstructed target source would have poor intelligibility. Hence, in addition to the SNR metric, we assess the *characterisation* of the target source by the segregation system, defined as

$$CHAR(t) = \frac{\sum_{\varepsilon=left,right} \sum_{f} w(\varepsilon,t,f) * s(\varepsilon,t,f)}{\sum_{\varepsilon=left,right} \sum_{f} s(\varepsilon,t,f)} \times 100\% \qquad (7)$$

This metric indicates the proportion of energy belonging to the target source that the system has retrieved from the mixture. Characterisation scores are shown in Figure 6. Typically, the segregation system recovers about 30% of the target voice.

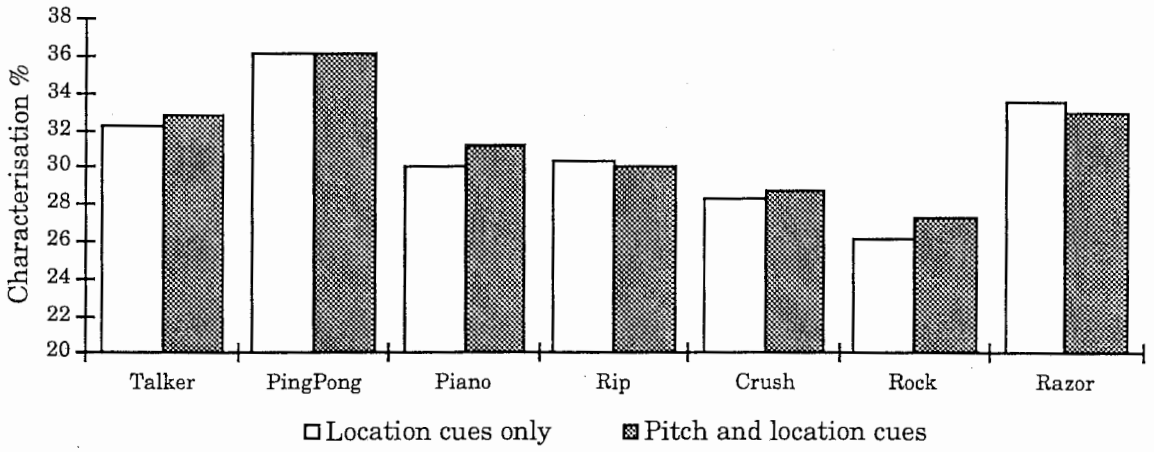

Figure 6: Comparison of characterisation scores after segregation by two schemes. The details of the pitch and location scheme are given later in the paper.

Discussion

The segregation system achieves modest improvements in SNR when tested on a small data set. Two observations suggest that the performance of the system could be improved.

First, although the principle of grouping spectral components with a common ITD fits elegantly into Bregman's framework for auditory scene analysis, this explanation may be flawed. In natural acoustic environments, several wideband sounds originating from different spatial locations may reach the ears at any one time. It is quite likely, therefore, that in a particular critical band the auditory response at one ear will be dominated by one source, and the response at the other ear will be dominated by a different source. As a result, spurious peaks may occur in the cross-correlogram that do not correlate with the ITD of a sound source. A similar observation has been made by Lyon (1988), although he does not suggest a solution:

> "*...local apparent direction decisions often do not correspond to any real sound source, but are the results of mixtures of signals...*" *(page 322)*

Second, recent psychophysical studies support the notion that the segregation of concurrent sounds according to their fundamental frequencies is a more robust strategy than using binaural cues (Culling *et al.*, 1994). These considerations prompted the development of a second system, in which location and periodicity cues are used together.

EXPERIMENT TWO: PERIODICITY AND LOCATION CUES

The block diagram of a second segregation scheme is shown in Figure 7. The scheme has a number of interesting characteristics. First, location is seen as a *derived property* of groups of spectral components that have been formed by the action of another primitive grouping cue, common periodicity. Second, pitch analysis and location analysis are performed by mechanisms with similar mathematical abstractions (autocorrelation and cross-correlation respectively). Both mechanisms also involve integration actoss frequency. Finally, the scheme proposes a novel auditory representation, the *pitch-azimuth-time cube*, which allows temporal continuity in the pitch and location domains to be exploited simultaneously. Peripheral auditory processing and cross-correlation are performed as described for the previous scheme; other components of the new system are described below.

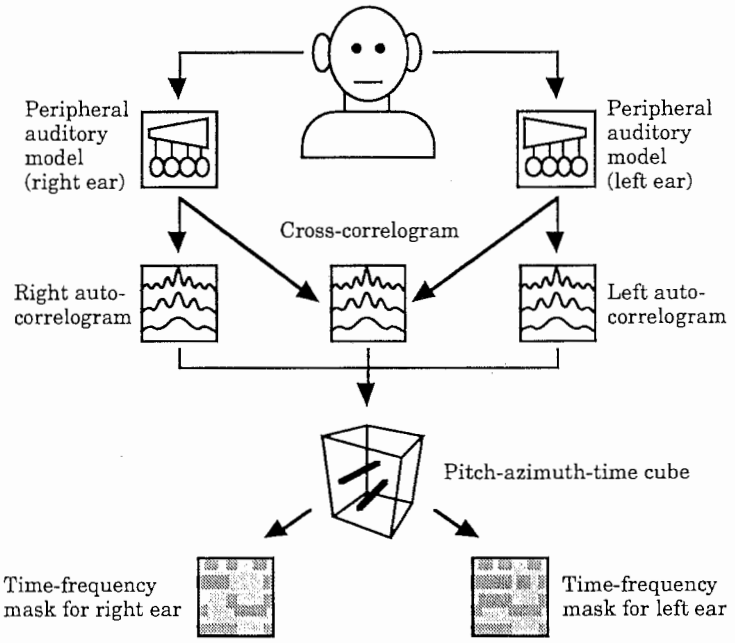

Figure 7: A segregation scheme using periodicity and location cues. The fundamental frequency and location of each auditory event is displayed within a three-dimensional auditory representation, the pitch-azimuth-time cube. The cube allows continuity in the pitch and location domains to be exploited simultaneously.

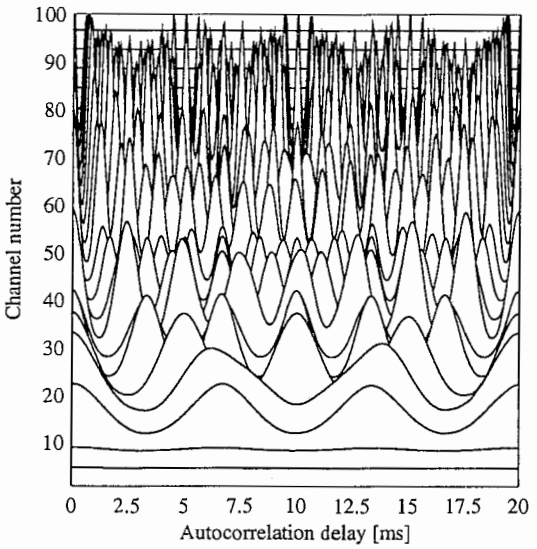

Figure 8: Autocorrelogram of a mixture of two complex tones. For clarity, only the response of every fourth channel is shown.

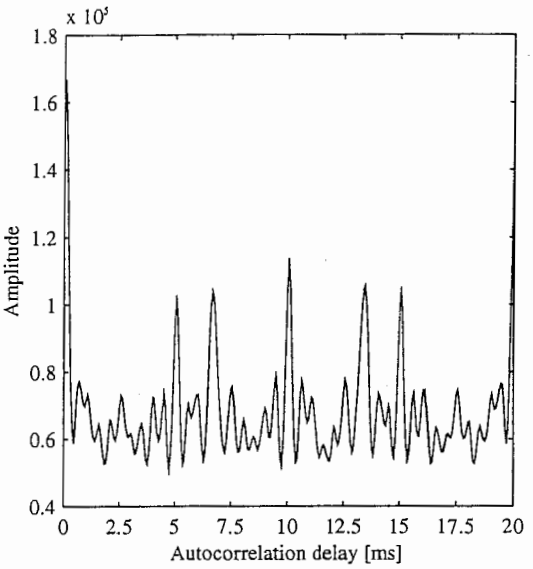

Figure 9: Summary autocorrelogram for a mixture of two complex tones. Peaks occur at correlation delays corresponding to the pitch period of each complex (5 ms and 6.6 ms).

Autocorrelogram

Periodicity information is extracted from the auditory nerve by the familiar 'autocorrelogram' model of auditory pitch analysis (e.g. Slaney & Lyon, 1990; Meddis & Hewitt, 1991; Brown & Cooke, 1994a,b), a computational implementation of Licklider's Duplex theory (Licklider, 1954). In this scheme, a running autocorrelation of the auditory nerve response at time T is computed for each filter channel f as follows:

$$acg(\varepsilon,T,f,\tau_a) = \sum_{t=0}^{T} r(\varepsilon,f,t)r(\varepsilon,f,t-\tau_a)\exp\left(\frac{-t}{\Omega_a}\right) \tag{8}$$

Here, $r(\varepsilon,f,t)$ is the probability of a spike in the auditory nerve, derived from the Meddis hair cell model, for ear ε at time t. The time constant of integration Ω_a is set to 20 ms. Autocorrelation functions are computed for values of the time lag τ_α between 0 ms and 20 ms; the latter was considered to be a reasonable upper limit for the period of voiced speech. A left ear autocorrelogram for a mixture of two complex tones with fundamental frequencies 200 Hz and 150 Hz is shown in Figure 8 (the corresponding cross-correlogram is shown in Figure 2).

Pitch analysis

The pitch periods of periodic sources are identified by integrating activity in the autocorrelogram over frequency, as suggested by Meddis & Hewitt (1992). Here, we assume that the autocorrelation functions from both ears are pooled to give a summary autocorrelogram:

$$sum_{acg}(t,\tau_a) = \sum_f acg(left,t,f,\tau_a) + \sum_f acg(right,t,f,\tau_a) \tag{9}$$

The pooling of periodicity information from both ears is consistent with psychophysical evidence that binaural integration of spectral energy precedes pitch identification (e.g., Houtsma & Goldstein, 1972). Large peaks in the summary autocorrelogram occur at the pitch period of a stimulus. For example, Figure 9 shows the summary function for the stimulus used in Figure 8. Large peaks occur at the pitch periods of the 200 Hz complex (5 ms) and the 150 Hz complex (6.6 ms). Note however, that peaks also occur at integer multiples of these periods.

For the task of tracking the pitches of two concurrent sources, only large peaks in the summary autocorrelogram need to be considered as candidate pitch periods. Here, the five largest peaks in the summary autocorrelogram are employed in subsequent stages of processing.

Simultaneous grouping by common periodicity

Auditory filters that are synchronised to the periodicity of a particular pitch period can be identified by sampling the channels of the autocorrelogram at the corresponding autocorrelation delay, as suggested by Assmann & Summerfield (1990). Here, the autocorrelograms for each ear are sampled to give separate synchrony spectra for the left ear *p(left,t,f)* and right ear *p(right,t,f)*.

Location analysis

The spatial location corresponding to a group of components with common periodicity is identified by weighting the channels of the cross-correlogram by the synchrony spectrum for the group. Specifically, a weighted summary cross-correlogram

$$sum_{wgt}(t,\tau) = \sum_f ccg(t,f,\tau) * \left[\frac{p(left,t,f) + p(right,t,f)}{2} \right] \tag{10}$$

is computed. The cross-correlation delay at which the largest peak occurs in this summary function is taken to be the ITD of the group:

$$itd(t) = \max_\tau sum_{wgt}(t,\tau) \tag{11}$$

An example of this procedure is shown in Figures 10 and 11, for the mixture of two harmonic complexes whose unweighted summary cross-correlogram function is shown in Figure 3. In Figure 10, the summary cross-correlogram has been weighted by the synchrony spectrum for the complex with fundamental frequency 150 Hz. The largest peak in the weighted summary function occurs at a correlation delay corresponding to the ITD of this complex (-1 ms). Similarly, the summary cross-correlogram shown in Figure 11 has been weighted by the synchrony spectrum for the complex with fundamental frequency 200 Hz, and exhibits a peak at the corresponding ITD (+1 ms).

Pitch-azimuth-time cube

At each time frame, the pitch periods identified in the summary autocorrelogram are plotted together with their corresponding cross-correlation delays to form a three-dimensional auditory representation, the *pitch-azimuth-time cube*. The motivation for this representation is that it allows continuity in the pitch and location domains to be exploited simultaneously.

Points in the pitch-azimuth-time cube are shown in Figure 12 for a stimulus consisting of a mixture of two pulsed complex tones. Each pulse has a duration of 200 ms and has the same characteristics as the mixture used previously (a component with F0 150 Hz and ITD -1 ms and a component with F0 200 Hz and ITD +1 ms). Note that spurious peaks occur at integral multiples of the pitch period of each source, but at the same cross-correlation delay.

Event formation

Auditory events are identified by tracking points in the pitch-azimuth-time cube. This procedure is similar to the birth-death tracking process described previously, except that points are recruited to an event only if their pitch and location are sufficiently similar to that of the candidate event. Event formation therefore proceeds in a three dimensional space, rather than the two-dimensional space used previously. Specifically, the first rule of the peak tracking scheme is modified as follows:

Rule 1: A peak in the pitch-time-azimuth cube at time t, autocorrelation delay τ_{a1} and cross-correlation delay τ_{c1} is recruited to an existing event ending at time $t-1$, autocorrelation delay τ_{a2} and cross-correlation delay τ_{c2} if $|\tau_{a1} - \tau_{a2}| < \alpha_a$ and $|\tau_{c1} - \tau_{c2}| < \alpha_c$. Here, α_a is 0.6 ms and α_c is 0.2 ms. As before, the values of these thresholds are not critical, but represent a compromise between many short events and long events that incorporate unrelated peaks.

The auditory events formed by tracking the peaks in Figure 12 are shown in Figure 13. Events that exist for less than two time frames are removed, since these are unlikely to have a significant acoustic correlate.

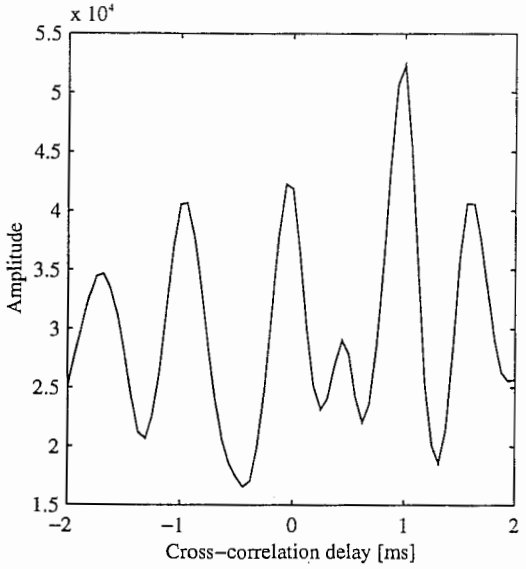

Figure 10: Cross-correlogram for the mixture of two complex tones weighted by the pitch synchrony spectrum for the 150 Hz complex. The largest peak occurs at the ITD of the complex (-1 ms).

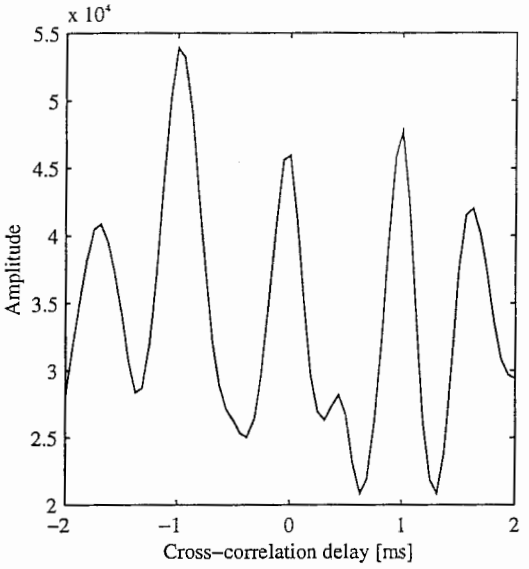

Figure 11: Cross-correlogram for the mixture of two complex tones weighted by the pitch synchrony spectrum for the 200 Hz complex. The largest peak occurs at the ITD of the complex (+1 ms).

Sequential grouping by similarity of pitch and spatial location

As in the first scheme, the cross-correlation delays at each time frame of an auditory event are averaged to give a single estimate of location. Autocorrelation delays are averaged in a similar manner, giving a mean pitch. Events that have a similar pitch period and location are allocated to the same source. Events are judged to be similar if there is a difference of less than 0.3 ms between their average cross-correlation delays and a difference of less than 3 ms between their mean pitch periods. The sequential grouping of auditory events according to the similarity of their pitches is in agreement with psychophysical studies (e.g. Blokx & Nooteboom, 1982).

This grouping process is illustrated in Figures 14 and 15, which show groups derived from the pitch-azimuth-time cube for the mixture of two harmonic complexes, shown in Figure 13.

Again, breaks may occur in the tracking of events. These gaps occur when the source being tracked is not periodic, and also because of errors in the pitch and location analyses or because the source has been masked. The cross-correlation delays in these gaps are interpolated using linear least squares analysis.

Evaluation

For each source, the system identifies a stream of events. When the source is periodic, events in the stream are defined by a pitch (autocorrelation delay) and a location (cross-correlation delay) at each time frame. In these cases, the spectrum of the source is recovered by sampling the autocorrelogram at the correlation delay τ_p corresponding to the pitch period:

$$w(\varepsilon,t,f) = v(\varepsilon,\tau_s) * acg(\varepsilon,t,f,\tau_p) \qquad (12)$$

As before, the scaling factor $v(\varepsilon,\tau_s)$ shown in Figure 4 weights the spectra for the left and right ears according to the cross-correlation delay τ_s.

When the souce is not periodic, events are defined only by an (interpolated) location. In these cases, the spectrum of the source is recovered by sampling the cross-correlogram, as described for the first scheme.

The system was evaluated on the mixtures and speech and intrusive noises described previously. Mean SNRs for each noise condition are shown in Figure 5. Performance of the system using pitch and location cues is slightly better for each noise condition that performance using location

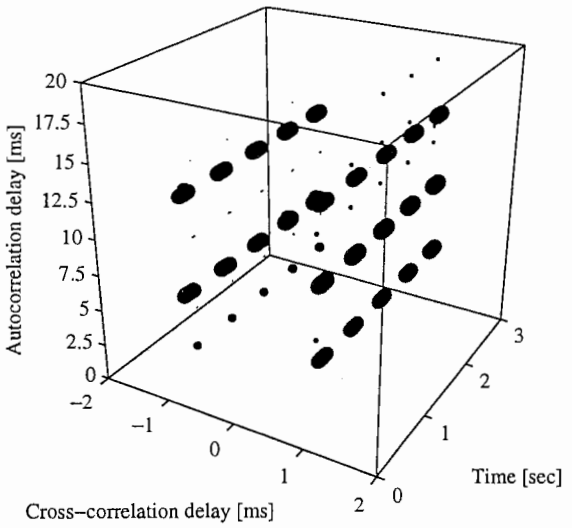

Figure 12: Points in the pitch-azimuth-time cube for a stimulus consisting of a mixture of two pulsed complex tones. One component of the mixture has a F0 of 150 Hz and an ITD of -1 ms, the other has a F0 of 200 Hz and an ITD of +1 ms. Each pulse has a duration of 200 ms.

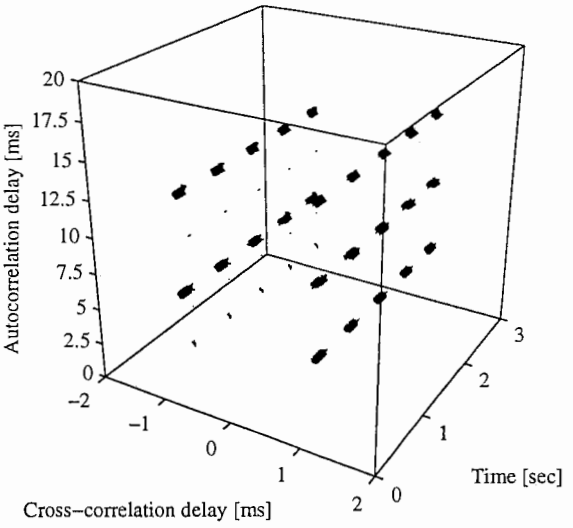

Figure 13: Auditory events for the mixture of two pulsed complex tones, derived by tracking the points in Figure 12 through time. Events occur at the pitch period of each source, and also at integer multiples of the pitch period.

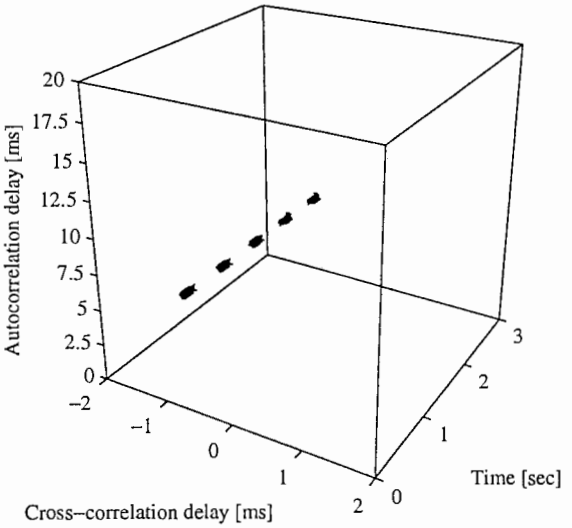

Figure 14: Pitch-azimuth-time stream for the component of the mixture with fundamental frequency 150 Hz and ITD -1 ms. The stream was derived by sequentially grouping the events in Figure 13 according to the similarity of their pitch (autocorrelation delay) and location (cross-correlation delay).

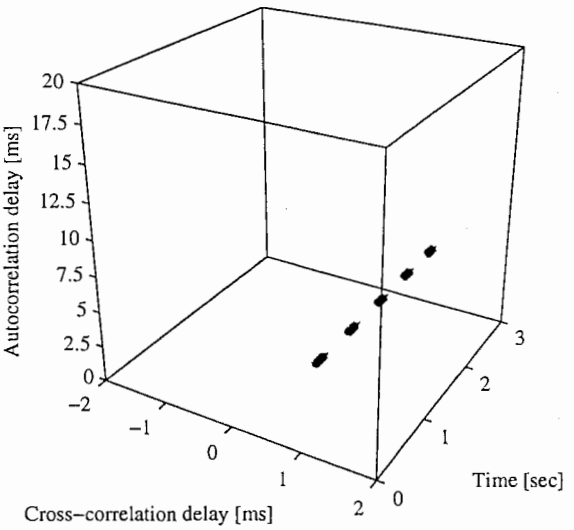

Figure 15: Pitch-azimuth-time stream for the components of the mixture with fundamental frequency 200 Hz and ITD +1 ms, derived in the same manner as the stream shown in Figure 14.

cues alone. Similarly, Figure 6 indicates that the characterisation of the target source is improved for most noise conditions by using both cues rather than location cues only.

Discussion

The results of the second experiment suggest that a segregation scheme in which pitch and location cues are used together has a performance advantage over a scheme that uses location cues only.

Although the emphasis of this study has been on investigating the way in which different auditory grouping cues can be employed together — rather than on constructing a high performance segregation system – it should be remarked that both systems only achieved modest improvments in SNR, and the improvement obtained by the second scheme relative to the first was also small. A possible reason for this was that neither system employed a principle of exclusive allocation (Bregman, 1990) in which all of the energy in a particular frequency band is assigned to a single source. Our previous work (Brown, 1992) suggests that such a scheme might give rise to a higher baseline performance. However, incorporating a principle of exclusive allocation into the current scheme was problematic, because it was not clear how to compute the cross-correlogram in frequency bands where the energy at the left and right ears had not been allocated to the same source.

It should be noted that the computational load of the second scheme is approximately three times that of the first scheme, since two autocorrelograms must be computed in addition to the cross-correlogram. If computational efficiency is an important consideration, therefore, it is debatable whether the small improvment in SNR obtained by the second system merits the computational cost of the additional processing.

GENERAL DISCUSSION

Relationship to Psychoacoustical Studies

Our results suggest that the performance of a segregation system, in terms of SNR improvement, is better if periodicity and location cues are used together than if location cues are used alone. The reason for this result is that an analysis of location is more reliable if the contribution of a source to the activity in each ear is determined first by other primitive grouping principles (in this case, common periodicity). Essentially, the second scheme considers periodic regions of a source to be the optimum points at which to compute its location. During nonperiodic regions, the location of the source is interpolated between these 'islands' of reliable location estimates.

The notion that pitch cues are more robust than location cues finds some support in the literature. A number of psychophysical studies have recently suggested that pitch-based segregation may be more reliable than location-based segregation, particularly in reverberant environments (e.g., Culling *et al.*, 1994). Similarly, it is well known that spatial location cues are subservient to pitch cues in some situations, as illustrated in Deutch's (19xx) scale illusion.

Additionally, the second scheme described here is compatible with the finding of Shackleton & Meddis (1992) that the ability of listeners to exploit spatial location cues in order to segregate concurrent vowels is improved when the vowels have a different fundamental frequency.

Limitations of the Model

Perhaps the most significant limitation of the current model is the omission of interaural intensity difference (IID) cues. However, recent psychophysical findings suggest that interaural time differences (ITD) are the dominant cue for sound localisation. For example, Wightman & Kistler (1992) asked listeners to localise complex sounds that were synthesized in such a way that ITD and IID cues were placed in conflict. They concluded that

> "*...with wideband stimuli, interaural intensity and spectral shape cues appear to play a secondary role in cueing apparent source azimuth and elevation...the interaural time cue may be used primarily to establish the locus of possible source directions. Interaural intensity and spectral cues are then analysed to resolve confusions...*"

Such confusions occur because a given ITD is insufficient to uniquely define the position of a sound source in space. Rather, there is a 'cone of confusion' around each ear, such that any sound source positioned on the surface of the cone will give rise to the same ITD (Mills, 1972). However, the segregation system described here was limited to the task of localising sounds in the frontal horizontal plane. Hence, cone of confusion ambiguities did not need to be considered, and it is doubtful whether the inclusion of IID cues would have significantly improved the performance of the system.

Clearly, it would be necessary to implement IID cues if the system was required to segregate sounds from arbitrary positions in auditory space. The IID at characteristic frequency f and time t can be computed in decibels as follows:

$$iid(f,t) = 10\log_{10}\left(\frac{ar(right,f,t)}{ar(left,f,t)}\right) \tag{13}$$

Here, $ar(\varepsilon,f,t)$ is a leaky sum of the auditory nerve firing rate, defined by

$$ar(\varepsilon,f,t) = \sum_{0}^{t} r(\varepsilon,f,t)\exp\left(\frac{-t}{\Omega_3}\right) \tag{14}$$

where Ω_3 is the time constant of integration. A similar technique has previously been described by Macpherson (1991). Time and intensity differences could be combined in our system by relating ITD, IID and source location though a pre-learned mapping. Alternatively, IID could be treated as an additional dimension of the pitch-azimuth-time cube.

A second limitation arises from the fact that the system has only been applied to the segregation of sound sources in an anechoic environment. As a result, no attempt has been made to simulate the Haas (precendence) effect, which inhibits the contribution of reflected sound waves to the perceived location of a sound source (Zurek, 1980). Additionally, the tendency of listeners to weight certain spectral regions more heavily when making location judgments (Bilsen & Raatgever, 1973) and the bias of location judgments towards central rather than lateral locations (Stern *et al.*, 1988) have not been modelled.

SUMMARY

Two approaches to the segregation of speech from noise intrusions have been described. The first scheme grouped sound components according to their locations, whereas the second scheme employed pitch as well as location cues. In terms of improvement in signal-to-noise ratio, the second scheme has a performance advantage over the first.

ACKNOWLEDGMENTS

Thanks to Ray Meddis, Minoru Tsuzaki and Hideki Kawahara for comments on an earlier version of this paper and to Roy Patterson for stimulating discussions. Thanks also to Malcolm Crawford for software support and Inge-Marie Eigsti, James Magnuson and Kevin Lenzo for moral support.

APPENDIX

Binaural Recording Setup

Binaural recordings of a target source and interfering source were made in an anechoic chamber using a Brüel & Kjær head and torso simulator type 4128. The sources were positioned approximately 90 degrees apart, with the target source to the right side of the head (see Figure 16). Both sources were placed 1 metre from the head.

The left and right outputs from the head and torso simulator were amplified by a Yamaha HA8 microphone preamplifier, and sampled at a rate of 48 kHz with 16 bit resolution by a Macintosh IIfx computer with Digidesign ProTools software. Recordings were then downsampled to 16 kHz. Clearly, using the higher sampling rate would be preferable since this would give greater resolution in the autocorrelation and cross-correlation maps; however, the computational load at such a high sampling rate is prohibitive.

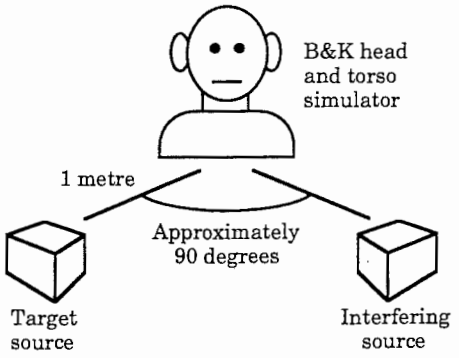

Figure 16. Schematic diagram of the binaural recording setup.

Note that because quantitative evaluation of the segregation system requires separate signal and noise waveforms, the target source and interfering source were recorded separately. Mixtures were obtained by adding the waveforms of the target sound and interfering sound after downsampling. Since the outer and middle ears are roughly linear, this technique gives a reasonable approximation to the waveform that would be obtained by recording the target and interfering sources simultaneously.

Details of the Test Stimuli

Target source. The target source was a male native English speaker uttering the sentence "one, two, three, four, five". Note that the target sound therefore contained voiced and unvoiced segments; this is a departure from our previous studies (Brown, 1992; Cooke & Brown, 1993; Brown & Cooke, 1994a,b) which used voiced speech only.

Interfering sources. The waveform of the target sound was added to the waveform of each of seven interfering sounds, giving seven mixtures. The interfering sounds were selected as examples of typical environmental intrusions; their properties are given below:

ID	Description	Characteristics
Talker	Voice of male native English speaker	Wideband, continuous
Rip	Sheet of paper being ripped	Mid to high frequency, continuous
Crush	Sheet of paper being crumpled	Mid to high frequency, continuous
Rock	Rock music	Wideband, continuous/impulsive
Piano	Solo piano music	Wideband, continuous
PingPong	Table tennis ball dropped on bat	Wideband, impulsive
Razor	Buzz of electric razor	Wideband, continuous

The music intrusions were played from compact disc through a pair of small domestic hi-fi speakers. The *PingPong* intrusion was inspired by a similar stimulus used by Lyon (1988) to test his binaural segregation scheme.

REFERENCES

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

Baumann, U. (1992) Pitch and onset as cues for the segregation of musical voices. *Proceedings of the 2nd ICMPC*, Los Angeles.

Bilsen, F.A. & Raatgever, J. (1973) Spectral domainance in lateralisation. *Acustica*, **28**, 131-132.

Blauert, J. (1983) *Spatial hearing: The psychophysics of human sound localisation*. MIT Press, London.

Blokx, J.P.L. & Nooteboom, S.G. (1982) Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, **10**, 23-56.

Bregman, A.S. (1990) *Auditory scene analysis*. MIT Press.

Bregman, A.S. & Pinker, S. (1978) Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, **32**, 19-31.

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

Brown, G.J. & Cooke, M.P. (1994a) Computational auditory scene analysis. *Computer Speech and Language*, in press.

Brown, G.J. & Cooke, M.P. (1994b) Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research*, **23**, 107-132.

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.

Cooke, M.P. (1993) *Modelling auditory processing and organisation*. Cambridge University Press.

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

Culling, J.F., Summerfield, Q. & Marshall, D.H. (1994) Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. *Speech Communication*, **14**, 71-95.

Deutsch, D. (1974) An auditory illusion. *Nature*, **251**, 307-309.

Dye, R.H. (1990) The combination of interaural information across frequencies: Lateralisation on the basis of interaural delay. *Journal of the Acoustic Society of America*, **88**, 2159-2170.

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

Houtsma, A.J.M & Goldstein, J.L. (1972) The central origin of the pitch of complex tones: Evidence from musical interval recognition. *Journal of the Acoustical Society of America*, **51**, 520-529

Jeffress, L.A. (1948) A place theory of sound localisation. *Journal of Comparitive Physiology and Psychology*, **41**, 35-39.

Kashino, K. & Tanaka, H. (1992) A sound source separation system using spectral features integrated by the Dempster's law of combination. In *Annual Report of the Engineering Research Institute*, 67-72.

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

Lyon, R.F. (1988) A computational model of binaural localization and separation. In *Natural Computation*, edited by W. Richards, MIT Press, pp. 319-327.

Macpherson, E.A. (1991) A computer model of binaural localisation for stereo imaging measurement. *Journal of the Audio Engineering Society*, **39**, 604-621.

Marr, D. (1982) *Vision*. W.H. Freeman and Company, San Francisco.

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) Simultation of auditory-neural transduction: Further studies. *Journal of the Acoustical Society of America*, **83**, 1056-1063.

Meddis, R. & Hewitt, M.J. (1991) Vitual pitch and phase sensitivity of a computer model of the auditory periphery: I. phase sensitivity. *Journal of the Acoustical Society of America*, **89**, 2883-2894.

Meddis, R. & Hewitt, M.J. (1992) Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, **91**, 233-245.

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

Mills, A.W. (1972) Auditory localisation. In *Foundations of Modern Auditory Theory Volume 2*, ed. J.V. Tobias, Academic Press, New York.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J. and Rice, P. (1987) An efficient auditory filterbank based on the gammatone function. *Institute of Acoustics Speech Group Meeting on Auditory Modelling*, RSRE, December 14-15.

Shackleton, T.M. & Meddis, R. (1992) The role of interaural time difference and fundamental frequency difference in the indentification of concurrent vowel pairs. *Journal of the Acoustical Society of America,* **91**, 3579-3581.

Shackleton, T.M., Meddis, R. & Hewitt, M.J. (1992) Across frequency integration in a model of lateralization. *Journal of the Acoustical Society of America,* **91**, 2276-2279.

Slaney, M. & Lyon, R.F. (1990) A perceptual pitch detector. *Proceedings of ICASSP-90*, 357-360.

Stern, R.M., Zeiberg, A.S. & Trahiotis, C. (1988) Lateralisation of complex binaural stimuli: A weighted image model. *Journal of the Acoustical Society of America*, **84**, 156-165.

Todd, N.P. & Brown, G.J. (1994) A computational model of prosody perception. *Proceedings of ICSLP-94*, Yokohama, Japan, 18th-22nd September.

Trahiotis, C., & Stern, R.M. (1989) Lateralisation of bands of noise: Effects of bandwidth and differences of interaural time and phase. *Journal of the Acoustic Society of America*, **86**, 1285-1293.

Webster, F.A. (1951) Influence of interaural phase on masked thresholds. *Journal of the Acoustical Society of America*, **23**, 452-462.

Wightman, F.L. & Kistler, D.J. (1992) The dominant role of low-frequency interaural time differences in sound localisation. *Journal of the Acoustical Society of America*, **91**, 1648-1661.

Yin, T.C.T. & Chan, J.C.K. (1988) Neural mechanisms underlying interaural time sensitivity to tones and noise. In *Auditory Function: Neurobiological Bases of Hearing*, ed. G.M. Edelman, W.E. Gall & W.M. Cowan, Wiley, London, 385-430.

Zurek, P.M. (1980) The precedence effect and its possible role in the avoidance of interaural ambiguities. *Journal of the Acoustical Society of America*, **67**, 952-964.