Proceedings of the Institute of Acoustics

COMPUTATIONAL AUDITORY SCENE ANALYSIS: GROUPING SOUND SOURCES USING COMMON PITCH CONTOURS

G.J. Brown & M.P. Cooke

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

1. INTRODUCTION

Recently, a number of models of double (simultaneous) vowel perception have been described which are based on *correlograms*, an auditory representation in which frequency and pitch period are displayed orthogonally. Some of these schemes have been quite successful in predicting the performance of listeners on double vowel segregation tasks [1][10]. Additionally, correlograms belong to the family of place-time models of pitch perception, and are able to account qualitatively for many of the classical pitch perception phenomena [11]. However, it remains to be demonstrated that correlograms can serve as a basis for the segregation of *arbitrary* sound sources. No studies to date have addressed this issue.

In this paper, we present quantitative performance figures (measured as an improvement in signal-to-noise ratio) for two correlogram-based segregation strategies, evaluated over a large database of speech mixed with a variety of other sources. The first strategy, which is similar to those of Meddis and Hewitt [10] and Assmann and Summerfield [1], estimates the pitch of a source and then identifies the channels of the correlogram that match the candidate pitch. The second strategy operates in a somewhat inverse manner: the most likely pitch period is estimated for each channel of the correlogram, and channels which have a similar predicted pitch period are grouped. We show that this inversion can alleviate some of the problems associated with the first scheme, and extend the system to use pitch *contours* rather than single-frame estimates.

2. AUTOCORRELATION-BASED SEGREGATION SCHEMES

The correlograms shown in this paper were derived from a model of the auditory periphery, consisting of a bank of 128 bandpass gammatone filters followed by the Meddis [9] model of inner hair cell transduction. Filters were spaced equidistantly on the ERB-rate scale of Moore and Glasberg, between centre frequencies of 50Hz and 5kHz. The output of each channel of the model is a probabilistic representation of auditory nerve firing activity. The correlogram is obtained by computing the running short-term autocorrelation of the activity in each channel, as originally suggested by Licklider [7].

A correlogram for a mixture of the synthetic vowels /a/ (fundamental 112 Hz) and /e/ (fundamental 100 Hz) is shown in the upper panel of figure 1. The lower panel of the figure shows a *summary autocorrelation function* [10], which is obtained by averaging the autocorrelation functions over all channels of the auditory filterbank. Note that a peak occurs in the summary autocorrelation at the fundamental of each vowel.

Given the correlogram representation, previous approaches have attempted to identify the pitch of one or more sources from the summary autocorrelation function, and then partition the energy in the channels of the correlogram on the basis of the candidate pitch. Assmann and Summerfield (A&S) describe a technique of this form, in which the two largest peaks in the summary autocorrelation function are identified. The delays at which these peaks occur are assumed to correspond to the pitch periods of the two vowels. Subsequently, the spectrum of each vowel is estimated by sampling the channels of the correlogram at the delay corresponding to the vowel's pitch period. Hence, two 'synchrony spectra' are obtained, which indicate the degree of synchronisation to each vowel in the auditory nerve. By matching these spectra against reference templates, vowel identification performance can be quantified. The A&S scheme comes close to predicting the overall accuracy of listeners' responses. However, it is unable to replicate the find-

ing of Scheffers [12] that identification performance improves with larger differences in fundamental frequency between the two vowels.

A more successful strategy, in terms of predicting human performance, has been proposed by Meddis and Hewitt (M&H). Given that there are two vowels present with different fundamental frequencies, the M&H scheme partitions the correlogram into two mutually exclusive sets of channels. Initially, the largest peak in the summary autocorrelation is identified, and this is taken to be the pitch period of the dominant vowel. Channels with a peak in their autocorrelation functions at this delay are removed from the correlogram, and matched with a template. The remaining channels are assumed to belong to the second vowel, and are matched with a template in a similar manner. Hence, only the pitch of the most dominant vowel is estimated. This is advantageous, since the second pitch is often weak, and may be an unreliable cue for segregation (Meddis and Hewitt [10]). The M&H scheme is able to model Scheffers' findings quite closely, and shows an improvement in performance when the difference in fundamental frequency between the two vowels is increased.

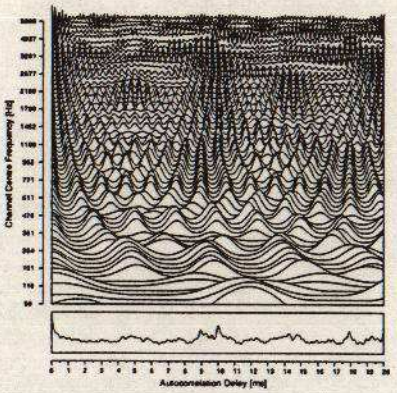

Figure 1: Correlogram and summary correlogram for the double vowel /a/ and /e/.

3. LIMITATIONS OF PREVIOUS APPROACHES

Although the A&S and M&H strategies are able to model the identification of double vowels quite closely, they suffer from a number of disadvantages. In particular, the schemes do not generalise in situations where several arbitrary sound sources are active at the same time. For example, both the A&S and M&H strategies require prior knowledge of the number of sound sources that are present. Consider the A&S scheme, which attempts to find the pitch period of each source by identifying peaks in the summary autocorrelation function. For stimuli more irregular than synthetic double vowels, this is a non-trivial problem. Generally, non-speech noise intrusions will generate a multitude of peaks in the summary autocorrelation, so that it is difficult to identify the number of sources present and assign a pitch period to each one. The M&H scheme overcomes the problem of multiple peaks by identifying the pitch period of the dominant source, and partitioning the channels of the correlogram into two mutually-exclusive sets. But what if there are more than two sound sources present? The M&H strategy does not easily generalise in this case.

A related criticism of the M&H scheme is that listeners can often hear both pitches in a double vowel, and are able to indicate which vowel has the higher pitch and which has the lower pitch (Summerfield *et al.* [13]). Similarly, Beerends and Houtsma [2] have found that listeners are often able to correctly identify the pitches of concurrent two-tone complexes, for differences in fundamental frequency of two semitones or more. It seems unlikely, therefore, that segregation is based only on the most dominant pitch.

Another point concerns the relationship between the pitch system and perceptual grouping mechanisms. The A&S and M&H segregation strategies assume that the pitch of a source is identified first, and then this pitch is used to group the components of the source together. However, Bregman [3] notes that this is unlikely to be the case:

> *"The pitch system acts to group harmonically related partials. We might conclude that this grouping is then used to derive other properties of the now segregated partials. This description implies a one-way transaction, the pitch system influencing the grouping system and not vice versa. However, this appears not to be true. There is evidence that the pitch that is calculated can depend on cues other than harmonicity, cues that we might think of as operating outside the pitch system."* (page 247)

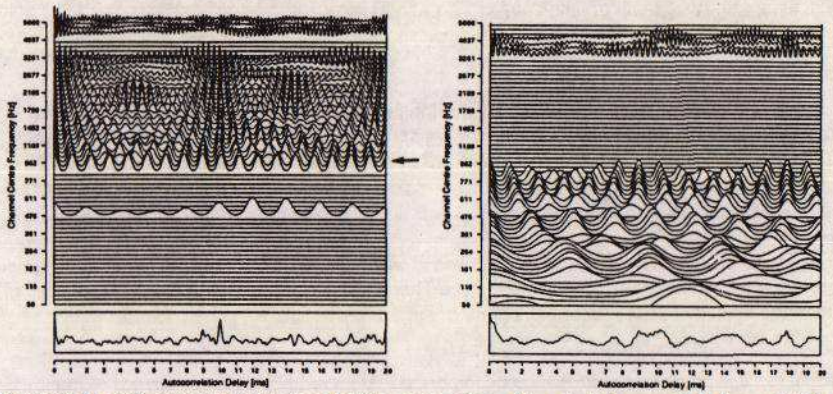

Figure 2: Meddis and Hewitt strategy for the separation of the double vowels in figure 1. Channels belonging to the /e/ are shown on the left, channels belonging to the /a/ are on the right.

This point has been demonstrated by McAdams [8], using a paradigm in which the odd and even harmonics of an oboe sound were separated and sent to different speakers. When the two sets of harmonics were coherently frequency modulated, a single source was heard with a single pitch. However, when the odd and even harmonics were incoherently modulated, two sounds were heard that had different pitches. Hence, it appears that perceptual grouping determines pitch, rather than vice versa. This conclusion may also be supported by the finding of Darwin and Ciocca [6], which indicates that a harmonic with a different onset time makes a reduced contribution to the pitch of a complex tone.

Finally, the A&S and M&H segregation strategies both suffer from the problem of *overlapping harmonics* (Assmann and Summerfield [1]). Consider the correlogram of the double vowel /a/ (fundamental 112 Hz) and /e/ (fundamental 100 Hz) shown in figure 1. The channel of this map with centre frequency 898 Hz is dominated by the eighth harmonic of the /a/, which has a frequency of 896 Hz. Peaks occur in the autocorrelation function at the period of this harmonic (1.12 ms) and at integer multiples of this period. A large peak occurs at a delay of eight periods (8.93 ms), corresponding to the pitch period of the /a/. However, there is also a smaller peak at a delay of nine periods (10.04 ms), which is close to the pitch period of the /e/ (10.0 ms). Since the /e/ is the dominant vowel in the mixture, the M&H strategy initially removes the channels of the correlogram which have a peak at a delay of 10.0 ms. Consequently, the channels dominated by the 896 Hz harmonic of the /a/ are incorrectly assigned to the /e/ (see arrowed channel in figure 2). A similar error is made by the A&S strategy, since the peak in the channel autocorrelation function at 10.04 ms is almost as large as the peak at 8.93 ms. Hence, the 'synchrony spectrum' sampled at the pitch period of the /e/ contains spurious energy in the region of 896 Hz.

4. A NEW STRATEGY

In this section, a new autocorrelation-based segregation strategy is presented which avoids many of the limitations of the A&S and M&H schemes.

The summary autocorrelation of a periodic sound has peaks at integer multiples of the pitch period, as well as a peak at the pitch period itself. In order to reduce the influence of these 'false' pitch 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:

$$s_w[t, \Delta t] = \frac{w[\Delta t]}{M} \sum_{j=1}^{M} acm[t, f, \Delta t] \qquad (1)$$

is computed, where the weighting function $w[\Delta t]$ is defined by

$$w[\Delta t] = 1 - 0.9\frac{\Delta t}{\Delta t_{max}} \tag{2}$$

as suggested by Weintraub [14]. Here, Δt_{max} is the longest autocorrelation delay and $acm[t,f,\Delta t]$ is the autocorrelation function of channel f at time t and delay Δt. 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 pitch period is larger than the peaks at integer multiples of the pitch period.

The weighted summary autocorrelation $s_w[t, \Delta t]$ is an average measure of the periodicities present in the correlogram. As such, it indicates the likelihood of a pitch period Δt occurring in the correlogram at time t. Similarly, the channel autocorrelation functions $acm[t, f, \Delta t]$ indicate the likelihood of a particular pitch period occurring in a channel of the correlogram. Therefore, the product of these two quantities gives an estimate of the likelihood that a channel f belongs on a pitch period Δt at time t,

$$Pr[t, f, \Delta t] = acm[t, f, \Delta t]\, s_w[t, \Delta t] \tag{3}$$

From equation (3), it is possible to predict the pitch period that a channel is most likely to belong on. Specifically, the predicted pitch period $p[t, f]$ is given by the autocorrelation delay at which $Pr[t, f, \Delta t]$ is highest

$$p[t, f] = \max_{\Delta t} Pr[t, f, \Delta t]\,. \tag{4}$$

Here, $p[t, f]$ is computed for values of Δt between 2 ms and 20 ms, corresponding to pitches in the range 50 Hz to 500 Hz. Segregation can now be achieved by application of the following grouping principle:

> *Channels of the correlogram are grouped together if they have the same predicted pitch period.*

This strategy is illustrated in figure 3, for the double vowel /a/ and /e/ shown in figure 1. The two largest groups found by this process, which account for 80% of the channels in the correlogram, are shown in the figure. The group on the left of the figure has a pitch period of 10.0 ms, and corresponds to the /e/. Similarly, the group on the right has a pitch period of 8.93 ms, and corresponds to the /a/. The remaining channels of the correlogram form small groups, or fail to group at all.

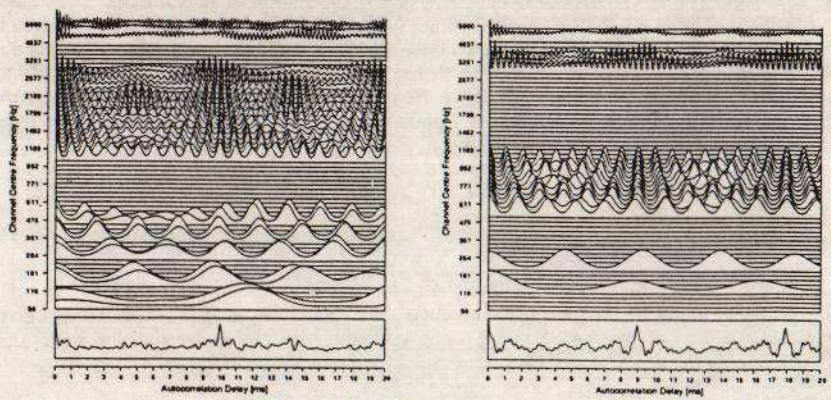

Figure 3: Segregation of the double vowels in figure 1 by the new strategy. Channels belonging to the /e/ are shown on the left, channels belonging to the /a/ are shown on the right.

This approach has a number of advantages when compared with the A&S and M&H 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 pitch periods. Secondly, because the strategy determines the *most likely* pitch period for each channel, it tolerates small irregularities in the channel autocorrelation functions. For example, a comparison of figures 2 and 3 shows that the new strategy has correctly assigned several low-frequency channels to the /e/ that were incorrectly grouped with the /a/ by the M&H scheme. Thirdly, 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 correlogram, 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. Additionally, the strategy is robust in situations where there are many spurious peaks in the summary autocorrelation function.

It is also apparent from figure 3 that the new strategy can solve the problem of overlapping harmonics. The channels in the region of 896 Hz have been assigned to the /a/, as required. Note that this result is not due to any change in the dominance of the two vowels caused by the weighting of the summary autocorrelation function. If the M&H scheme were to use the weighted summary autocorrelation, it would still produce the groups shown in figure 2, since the largest peak still occurs at the 10.0 ms pitch period of the /e/. Rather, the new strategy is able to solve the problem of overlapping harmonics because of two factors. Firstly, channels of the correlogram are allocated exclusively to one source. Secondly, the strategy uses information about the height of the pitch period peak in the summary autocorrelation *and* in the channel autocorrelation. The A&S and M&H schemes each take *one* of these factors into account, but not *both*.

Although the new strategy solves the problem of overlapping harmonics in many situations where the A&S and M&H schemes fail, it is not guaranteed to do so in every case. Again, consider the double vowel /a/ and /e/. The channels near to 896 Hz are correctly assigned by the new strategy because the product of the summary and channel autocorrelation functions at the pitch period of the /a/ is larger than the product at the pitch period of the /e/. However, if the pitch period peak of the /a/ in the weighted summary autocorrelation was much smaller than the pitch period peak of the /e/, the strategy would fail and the channels near to 896 Hz would be incorrectly grouped with the /e/. This problem could be minimised by exaggerating the differences in peak height in the channel autocorrelation functions. One way of achieving this would be to square the autocorrelation function in each channel.

Another strategy for solving the problem of overlapping harmonics has been proposed by Summerfield *et al.* [13]. They attempt to identify local pitches in an correlogram by convolving adjacent channels with Gabor functions. However, this approach is computationally expensive and fails at low frequencies where harmonics are resolved.

5. GROUPING PITCH CONTOURS

In Brown [4], we describe a strategy for characterising the auditory scene as a collection of time-frequency auditory objects. Primitives for object formation (and subsequent grouping processes) are provided by physiologically-motivated models of higher auditory organisation, called *auditory maps*. Channels of the correlogram that are responding to the same spectral dominance are identified by a *cross-correlation map*, and combined into explicit groups of channels called *periodicity groups*. Additionally, a map of frequency-transition sensitive cells is used to extract information about the movement of spectral dominances. Auditory objects are formed by using the frequency transition information to track periodicity groups across time and frequency.

The implementation of the new segregation strategy described in this section exploits the fact that temporal continuity has been made explicit in the auditory object representation. Rather than comparing predicted pitch periods at each time frame, a temporally-extensive *pitch contour* is computed for each object in the auditory scene. Subsequently, objects are grouped if their pitch contours are similar.

As before, the probability of each pitch period is predicted by computing the product of the channel and summary autocorrelation functions. However, auditory objects generally occupy more than one channel of the correlogram at each time frame. Therefore, a *local summary autocorrelation* is computed, which

averages the channel autocorrelation functions over the frequency spread of the object. For an auditory object which occupies channels f_1 to f_2 of the autocorrelation map at time t, the local summary autocorrelation $I[t, f_1, f_2, \Delta t]$ is given by

$$I[t, f_1, f_2, \Delta t] = \frac{1}{f_2 - f_1 + 1} \sum_{f=f_1}^{f_2} acm[t, f, \Delta t] . \quad (5)$$

Now, using the same rationale as for equation (3), the probability of the object belonging on a particular pitch period Δt at time t is given by

$$Pr[t, f_1, f_2, \Delta t] = I[t, f_1, f_2, \Delta t] \, s_w[t, \Delta t] \quad (6)$$

As described in the last section, the most likely pitch period could be estimated from equation (6) in a frame-by-frame manner. However, this approach does not take advantage of temporal continuity. Instead, $Pr[t, f_1, f_2, \Delta t]$ is computed at every time frame occupied by the auditory object, and the best path through this series of functions is found by a dynamic programming algorithm. The choice of pitch period at each time frame depends upon the probability of the new pitch period and its distance from the previous pitch period.

A dynamic programming score is computed for each initial pitch period, and the pitch 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 $Pr[t, f_1, f_2, \Delta t]$ in order to determine the pitch contour. This process is repeated for each object in the auditory scene.

Pitch contours for the objects in a mixture of speech and noise are illustrated in figure 4. Two distinct groups are visible, corresponding to the pitches of the speech and siren. Additionally, a small number of contours occur at twice the pitch period of the speech due to sub-octave errors in the tracking procedure.

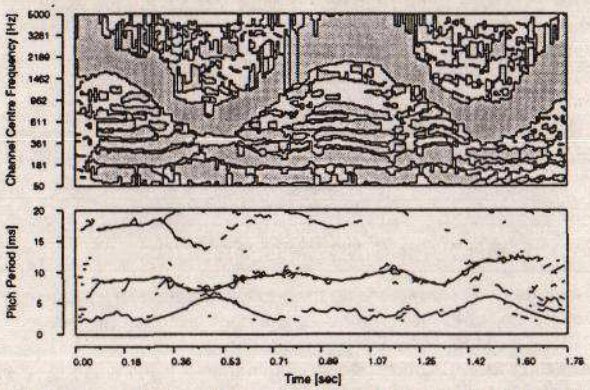

Figure 4: Auditory object representation of speech and siren intrusion (upper panel) and pitch contours for each object (lower panel).

6. QUANTITATIVE EVALUATION

We have developed a method for quantifying the signal-to-noise ratio of an acoustic mixture before and after segregation. The method relies on the linearity of a resynthesis path from the segregated representation; details of the technique are presented in Brown [4]. Here, we concentrate on a comparison between a strategy like the one proposed by Meddis and Hewitt and the new approach described earlier.

The evaluation is based on the 100 acoustic mixtures used by Cooke [5], obtained by adding each of 10 voiced sentences with each of 10 acoustic sources. These 10 'intrusions' included synthetic signals such as noise bursts, sirens and wideband noise, in addition to music, laboratory noise and other speech.

6.1 The frame-based scheme

Initially, pitch 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, sub-octave errors were manually corrected. Subsequently, these pitch contours were used to inform the segregation of the utterances from the noise intrusions. A correlogram was computed at each time frame, and channels of the map which had a peak at the given pitch 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 pitch period of the speech at each time frame. As such, the results represent the *optimum* performance of a frame-based autocorrelation segregation strategy on the test set.

6.2 The new scheme

Given a predicted pitch contour for each object in the auditory scene, segregation is achieved in the new scheme by application of the following grouping principle:

> *Auditory objects which overlap in time are grouped together if their predicted pitch contours are sufficiently similar.*

For two objects that overlap in time, the similarity of their pitch contours (calculated using the method described in the previous section) can be quantified by a Gaussian-weighted similarity metric. Two objects are allowed to form a group if their similarity score exceeds a threshold value (a value of 0.9 is used here). Clearly, this process groups auditory objects in a pairwise manner. Brown [4] describes a strategy for combining these pairwise groupings into larger groups. The grouping regime operates under extremely tight constraints: a new object is recruited to a group only if its pitch contour is sufficiently close to those of all other objects in the group.

6.3 Results

Figure 5 shows the mean SNR after segregation, for the object-based and frame-based correlogram strategies. The performance of the object-based scheme is better for every intrusion except n9, for which it is the same. The poorer performance of the frame-based strategy probably arises from the fact that frame-based schemes do not exploit temporal continuity.

Note that in two conditions (n1 and n8), the frame-based strategy *degrades* the mean SNR after segregation. This might be expected for the random noise intrusion (n1), since it causes many peaks to occur in the channel autocorrelation functions. If a spurious peak in a channel dominated by the noise intrusion coincides with the pitch period of the speech source, the channel will be inappropriately grouped. The problem of overlapping harmonics may contribute to the poor performance on n8, the male speech intrusion. From the pitch tracks of the speech intrusions and the

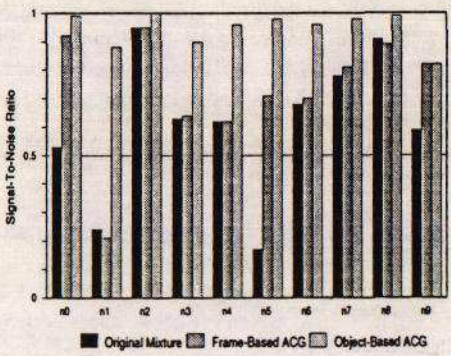

Figure 5: Signal-to-noise ratios for each of 10 intrusive sources (averaged across 10 voiced sources).

voiced utterances, the frequencies of the first 10 harmonics were calculated at each time frame and compared for overlap. This informal analysis suggests that, on average, overlapping harmonics occur more frequently for condition n8 (5.3% of time frames) than for n7 (1.3%) or n9 (2.5%).

7. SUMMARY AND DISCUSSION

A new correlogram-based segregation strategy has been described. The main innovation of the new scheme is that it allows pitch to arise as a result of grouping, rather than as a determinant of it. This view is supported by several psychophysical observations, which suggest that grouping cues such as common onset have an influence on the pitch system.

The performance of the object-based scheme is generally much better than that of a frame-based correlogram segregation scheme. Several factors may contribute to this result. Firstly, temporal continuity is made explicit in the object representation, whereas conventional correlogram strategies operate on a frame-by-frame basis. Secondly, the strategy used here is tolerant of variations in the position of peaks in the channel autocorrelation functions, so that the groups found by the model tend to be larger. Finally, the strategy is able to solve the problem of overlapping harmonics on many occasions. It may be possible to isolate some of these factors, in order to assess their contribution to the overall performance of the object-based scheme. For example, the segregation strategy described in section 4 could be implemented in a frame-based manner. This would allow the contribution of temporal continuity to be quantified.

8. REFERENCES

[1] P ASSMANN & Q SUMMERFIELD, 'Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies', J Acoust Soc Am 88, 680-697 (1990)

[2] JG BEERENDS & AJM HOUTSMA, 'Pitch identification of simultaneous diotic and dichotic two-tone complexes'. J Acoust Soc Am, 85, 813-819 (1989)

[3] AS BREGMAN, 'Auditory Scene Analysis'. Cambridge, MA: MIT Press (1990).

[4] GJ BROWN, 'Computational Auditory Scene Analysis: A Representational Approach', Ph. D. Thesis, University of Sheffield (1992).

[5] MP COOKE, 'Modelling Auditory Processing and Organisation', Cambridge University Press, forthcoming.

[6] CJ DARWIN & V CIOCCA, 'Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component', J Acoust Soc Am, 91, 3381-3390 (1992)

[7] JCR LICKLIDER,'A duplex theory of pitch perception',Experientia,7,128-134 (1951)

[8] S McADAMS, 'Spectral fusion, spectral parsing and the formation of auditory images', Ph.D. Thesis, Stanford University (1984)

[9] R MEDDIS, 'Simulation of mechanical to neural transduction in the auditory receptor'. J Acoust Soc Am, 83, 1056-1063 (1986)

[10] R MEDDIS & MJ HEWITT, 'Modelling the identification of concurrent vowels with different fundamental frequencies', J Acoust Soc Am, 91, 233-245 (1991)

[11] R MEDDIS & MJ HEWITT, 'Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. pitch identification', J Acoust Soc Am, 89, 2866-2882 (1992)

[12] M SCHEFFERS, 'Sifting vowels: Auditory pitch analysis and sound segregation'. Ph.D. Thesis, University of Groningen (1983)

[13] Q SUMMERFIELD, A LEA & D MARSHALL, 'Modelling auditory scene analysis: Strategies for source segregation using autocorrelograms', Proc Inst Acoust, 12, 507-514 (1990)

[14] M WEINTRAUB, 'A Theory and Computational Model of Monaural Auditory Sound Separation', Ph.D. Thesis, Stanford University (1985)