Proceedings of the Institute of Acoustics

MODELLING AUDITORY SCENE ANALYSIS: STRATEGIES FOR SOURCE SEGREGATION USING AUTOCORRELOGRAMS

Quentin Summerfield, Andrew Lea, and David Marshall

MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, UK.

1. INTRODUCTION

The autocorrelogram (ACG) of a sound is an informative representation in which pitch and timbre are displayed orthogonally. Interest in ACGs has arisen for two reasons. First, ACGs are generated when some "place-time" accounts of pitch perception [e.g. 1,2] are implemented as computational procedures [3,4,5]. Such models have been shown recently to account qualitatively for many of the classical phenomena of pitch perception. Second, ACGs offer a basis for explaining how listeners may use differences in pitch to segregate the sounds generated by competing periodic sources, such as pairs of talkers producing concurrent voiced speech sounds on different fundamental frequencies (f_0s) [6,7,8]. This paper is concerned with the second issue. It discusses strategies for segregating competing sources of sound using ACGs as the primary representation.

2. COMPUTATIONAL PROCEDURES

The ACGs illustrated in this paper were computed digitally in the following way. The waveform to be analysed was filtered by a set of 64 overlapping band-pass filters with centre frequencies (cfs) ranging from 63Hz to 4467Hz. The filters had the frequency responses of auditory filters measured psychoacoustically [9]. Their centre frequencies (cfs) were chosen to correspond to a constant spacing along the cochlear partition. The gain of the filters reflected the variation in absolute sensitivity with frequency found in listeners with normal hearing. Each filtered waveform was presented to the model of mechanical to neural transduction at the hair-cell synapse described by Meddis [10]. The output of this model is a time-varying probability of neural discharge in each channel reflecting the rectification, amplitude compression, and restricted range of phase-locking observed in primary auditory nerve fibres. The short-term autocorrelation function of a 20-ms segment of the waveform in each channel was computed. These "channel ACFs" show the strengths of the different periodicities present in each channel. Finally, the channel ACFs were plotted as a waterfall display with time (delay) on the x-axis and cf (place) on the y-axis.

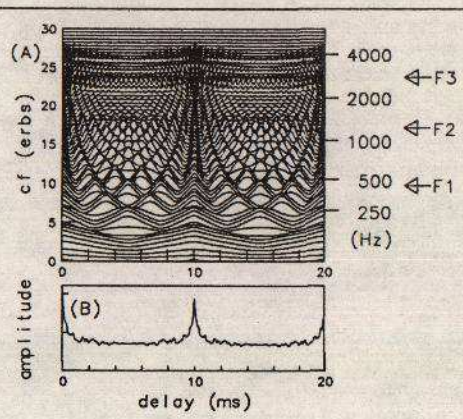

Fig. 1: (A) ACG of /ɜ/, f_0=100Hz. (B) Pooled ACG.

MODELLING AUDITORY SCENE ANALYSIS

3. THE ACG OF A SINGLE VOWEL

Fig. 1(A) shows the ACG of a steady-state synthetic vowel: an exemplar of /ɜ/ with f_0=100Hz, F1=450Hz, F2=1250Hz, and F3=2650Hz. The frequency axis has been scaled in units of the equivalent rectangular bandwidths (erbs) of the filters. One erb corresponds to a distance of about 0.85mm along the cochlear partition. The waveform in each channel is composed of harmonics of 100Hz and so repeats itself every 10ms, among other periods. As a result, there is a peak at a delay of 10ms in each channel ACF. These peaks form the vertical "spine" visible in the ACG. They reinforce each other if the channel ACFs are summed, giving a peak at 10ms, the period of the fundamental, in the resulting "pooled ACG", as shown in Fig. 1(B). The relative heights and positions of peaks in pooled ACGs provide good predictions of the pitches that listeners hear in complex sounds [3,4,5].

Resolved harmonics, and formants at higher frequencies, are marked by channels with greater energy. Thus timbre creates horizontal structure in the ACG while pitch creates vertical structure.

4. THE ACG OF A "DOUBLE" VOWEL

Now consider the case where two talkers are speaking together, each producing a periodic vowel. Fig 2(A) shows the ACG of a mixture of two vowels: the /ɜ/ with f_0=100Hz shown in Fig 1(A) has been mixed with an exemplar of /a/ with f_0=102.9Hz (0.5 semitones above 100Hz) and with F1=650Hz, F2=950Hz, and F3=2950Hz. Scheffers [11] was the first to show that listeners often hear two talkers producing vowels on different pitches when listening to such "double vowels". Compatibly, the pooled ACG (Fig. 2(B)) contains a peak at the period of each f_0 (10.0ms and 9.7ms) and the ACG shows evidence of the formants of both vowels.

How should the ACG be partitioned to recover the two vowels, thereby simulating the processes of perceptual grouping that underlie the experience of listening to this stimulus?

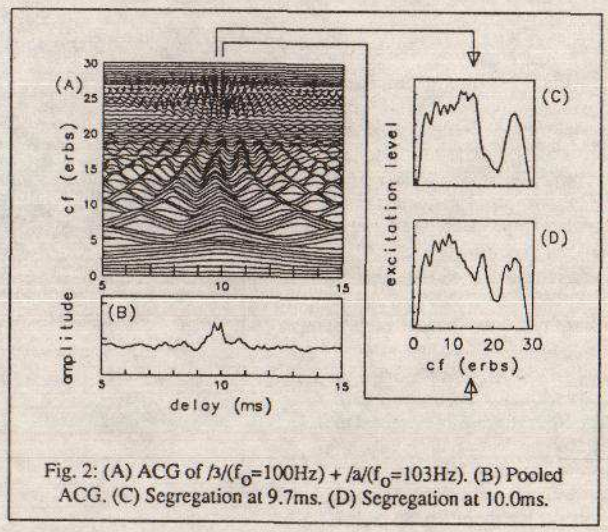

Fig. 2: (A) ACG of /ɜ/(f_0=100Hz) + /a/(f_0=103Hz). (B) Pooled ACG. (C) Segregation at 9.7ms. (D) Segregation at 10.0ms.

5. ASSMANN AND SUMMERFIELD'S SEGREGATION STRATEGY

Assmann and Summerfield (A&S) [7] noted that competing vowels are overlapping broad-band signals, thus each channel is likely to contain some evidence of both vowels. Accordingly, they implemented an example of the principle of "conjoint allocation" [12] in which some energy in each channel is assigned to each voice. A&S's strategy involved three stages: (i) locate the delay of the largest peak in the pooled ACG; (ii) plot the amplitudes of the channel ACFs at this delay as a spectrum, as in Fig 2(C); (iii) repeat steps (i) and (ii) for the second largest peak, generating a second spectrum, as in Fig 2(D). In this example, the strategy works well. The spectrum in Fig 2(C) contains peaks near the F1 (12.1 erbs) and F2 (14.9 erbs) of the /a/ while the spectrum in Fig. 2(D) contains peaks near the F1 (9.6 erbs) and F2 (17.1 erbs) of the /ɜ/.

Fig. 3 shows an example where the strategy is less successful. The stimulus is composed of the same two phonemic vowels as before. Once again the f_0 of the /3/ is 100Hz, but the f_0 of the /a/ is now 112.2Hz, 2 semitones above 100Hz. The two most prominent peaks in the pooled ACG (Fig. 3(B)) are found at the periods of the vowels, 8.9ms and 10.0ms. Segregation at the 8.9-ms period of the /a/ provides a satisfactory reconstruction of the spectrum of that vowel (Fig. 3(C)). However, segregation at the 10.0-ms period of the /3/ (Fig. 3(D)) produces a spurious formant at about 14.5 erbs (900Hz).

The spurious formant is the result of the problem of "overlapping harmonics" [7]. The 9th harmonic of 100Hz and the 8th harmonic of 112.2Hz more-or-less overlap. The segregation strategy does not possess the intelligence to deduce that nearly all of the energy in the region of 900Hz derives from the source with the period of 8.9ms. Inappropriately, the strategy assigns equal energy to each source, resulting in the spurious formant in the reconstruction of the /3/.

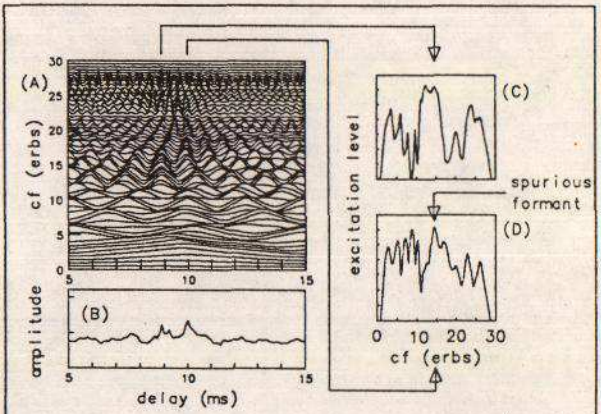

Fig. 3: (A) ACG of /3/(f_0=100Hz) + /a/(f_0=112Hz). (B) Pooled ACG. (C) Segregation at 8.9ms. (D) Segregation at 10.0ms.

The subsequent stages of A&S's model used a formant-based template-matching strategy to label segregated spectra. The model correctly predicted the accuracy of listeners' identification responses when the constituents of double vowels had f_0s separated by 0.5 semitones (as in Fig. 2), but underestimated their performance when the difference was 2 semitones (as in Fig. 3). It is likely, therefore, that listeners can solve the problem of overlapping harmonics. How might they do it?

6. MEDDIS AND HEWITT'S SEGREGATION STRATEGY

A segregation strategy with several attractions compared to A&S's procedure has been developed and evaluated by Meddis and Hewitt (M&H) [8]. It is based on two important insights. First, the second pitch in some double vowels is weak and so provides an unreliable basis for segregation. Second, when sounds with peaked spectra are mixed, energy from one or other source generally dominates each channel. As a result, any estimate of the contribution of the non-dominant source to that channel is likely to be unreliable. Accordingly, M&H implemented an example of the strategy of "disjoint allocation" [12]: (i) locate the most prominent peak in the pooled ACG; (ii) group all those channels whose individual ACFs contain a peak at this delay and treat them as evidence of the dominant voice; (iii) treat all the remaining channels as evidence of the non-dominant voice.

A weakness of M&H's strategy is that listeners often do hear two clear pitches in a double vowel and can indicate correctly which constituent has the higher and which the lower pitch. Nonetheless, the strategy correctly assigns formants to voices in many cases where A&S's strategy generates spurious formants. The subsequent stages of M&H's model use a template-matching strategy based on the low-time part of the pooled ACG to predict listeners' identification responses. Overall, in some important respects, their model comes closer to predicting the pattern of listeners' identification responses to double vowels than does A&S's model. However, the example in Fig. 3 has been chosen deliberately to illustrate a possible limitation of M&H's strategy. The largest peak in the pooled ACG (Fig.

3(B)) corresponds to the 10-ms period of the /3/. Fig. 4 shows the channels whose individual ACFs contain peaks at this delay. Channels near the F1 (450Hz) and F2 (1250Hz) of the /3/ have been selected, along with channels near 900Hz, once again representing the spurious formant.

The spurious formant arises here because the vowel with the dominant pitch (/3/ in this case) contains less energy at the frequency of the overlapped harmonics than the non-dominant vowel. The strategy groups the formant-dominated channels appropriately if the pattern of dominance is reversed by raising the level of the /a/ by 4dB to make it the dominant vowel. Channels excited by the F1 and F2 of the /a/, including channels close to 900Hz, are now grouped together, and the remaining channels mainly contain energy from the /3/.

Experience of listening to double vowels suggests that perception is not as dependent on the relative levels of the constituents as this example suggests it might be. Accordingly, we have explored a heuristic procedure for avoiding the problem of overlapping harmonics.

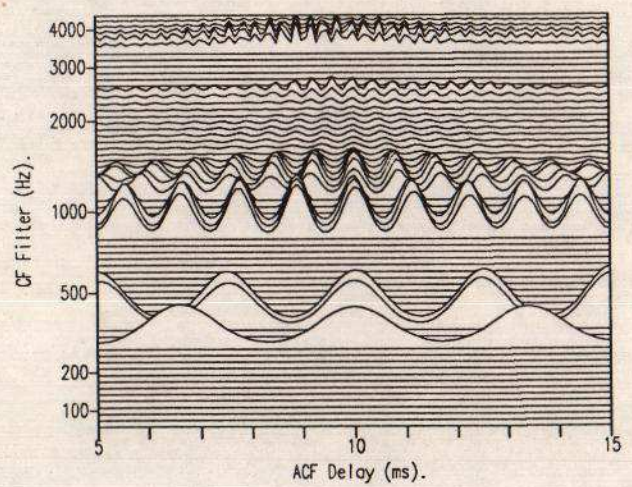

Fig. 4 Channels contributing to the "dominant pitch" segregated from Fig. 3(A) using the strategy described by Meddis and Hewitt.

7. PARTITIONING ACGS WITH GABOR FUNCTIONS

The rationale for the procedure can be understood by considering again the ACG of the single vowel /3/ shown in Fig. 1(A). The "architecture" of the ACG consists of the spine mentioned in Section 3 and a set of curved "arches" [13] which flank the spine. We refer to this patterning as the "tree structure". The spine is located at a delay of $1/f_0$. The arches arise because the harmonics in each channel necessarily have frequencies close to the cf of that channel. Thus, the waveform in each channel tends to repeat itself at intervals of approximately 1/cf, giving rise to a succession of peaks in the channel ACF at integer multiples of 1/cf. One of these peaks falls at $1/f_0$. Thus, the nearest arches to the spine are located at approximately $(1/f_0+1/cf)$ and at $(1/f_0-1/cf)$. The fact that the tree structure is displayed coherently across the entire frequency range in this example indicates that a single harmonic source is present.

Now consider the ACG shown in Fig. 5(A). The stimulus here was composed of the first 8 harmonics of 100Hz and the 8th to the 21st harmonics of 133Hz. The tree structure is disjoint with a break between the lower and upper parts of the trunk at about 1kHz. The stimulus gives the impression of two sources, one defined in the higher spectral frequencies with a higher pitch, the other defined in the lower spectral frequencies with a lower pitch. What computation can derive this description of the stimulus from the structure visible in the ACG?

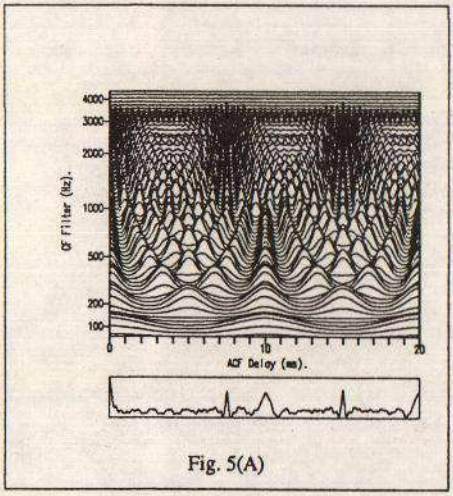

Fig. 5(A)

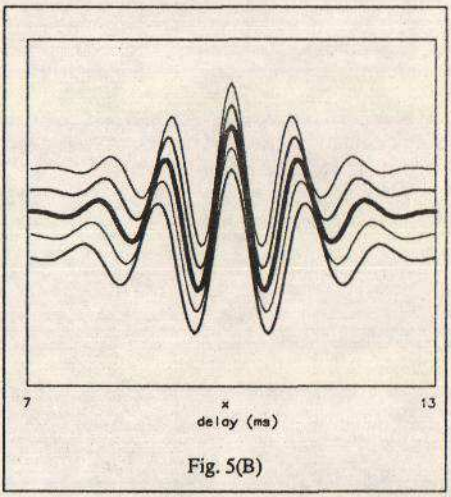

Fig. 5(B)

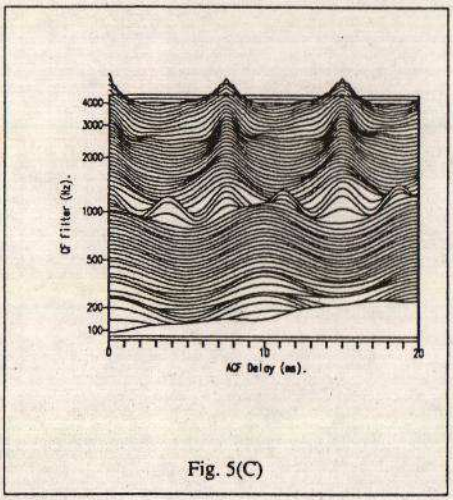

Fig. 5(C)

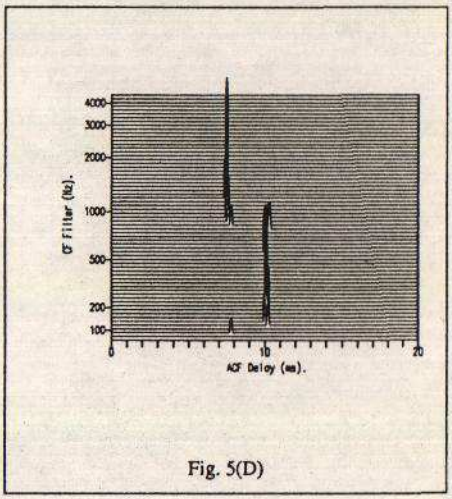

Fig. 5(D)

The strategy we have explored is derived from work by Gabor [14]. The ACG is convolved with an operator which describes the local appearance of the tree structure. The aim is to reduce the complex image of the ACG to a set of straight lines that trace out the spine of the tree.

To achieve this goal, the operator consists of five sinusoids each shaped by a Gaussian window (Fig. 5(B)). The vertical spacing of the sinusoids is the same as the spacing of the channels in the ACG. The frequency of each sinusoid is the cf of the channel with which it is aligned. The standard deviation of the Gaussian is 1/cf. Thus, the operator describes the local shape of the tree structure. When the operator is aligned with the spine of the tree, the convolution gives a large product. When it is misaligned, the product is smaller.

As described so far, the procedure is flawed since it also produces relatively large values when the operator is aligned with an arch rather than a spine, thereby producing appreciable ripple in the results of the convolution. The ripple can be removed by convolving the ACG with two different operators, one composed of sine functions and the other of cosine functions. If the results of the two convolutions that are computed at each point are squared and summed, the relationship, $\sin^2+\cos^2=1$, ensures that the result is a smooth function with peaks located on the spines of the tree.

The result of this double convolution is shown in Fig. 5(C). A peak picker has then selected the first major peak in each channel to generate the "simplified ACG" plotted in Fig 5(D). The amplitude of each peak is the amplitude at the corresponding point in the ACG shown in Fig. 5(A). The simplified ACG contains two lines of peaks. Their horizontal locations indicate the f_0s present in the stimulus. Their vertical ranges indicate the spectral frequencies over which each f_0 is defined. Thus, the distribution of "local pitches" in the stimulus has been computed and could be used to group the channels that contribute to each f_0.

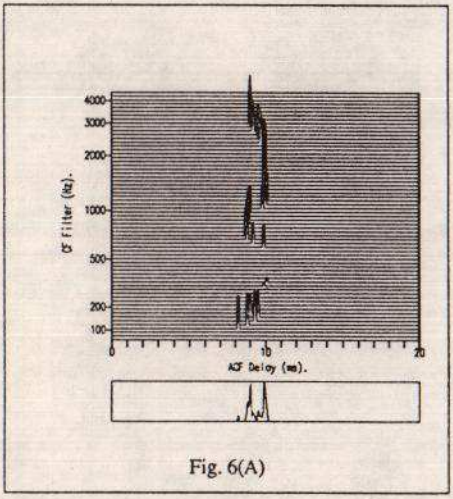

Fig. 6(A)

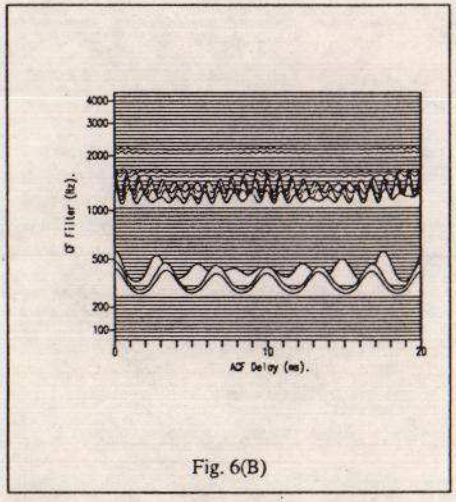

Fig. 6(B)

Fig. 6 illustrates the application of the strategy to the double vowel which gave rise to the problem of overlapping harmonics in Fig. 3. It is composed of an /ɜ/ with f_0=100Hz and an /a/ with f_0=112.2Hz. The main panel in Fig. 6(A) shows the simplified ACG. The panel beneath shows the "pooled simplified ACG". It contains prominent peaks close to the periods of the two f_0s. The next step is to group those channels which contain peaks in the simplified ACG at the delay of the most prominent period. That has been done to produce Fig 6(B) which contains channels near the F1 (450 Hz) and F2 (1250 Hz) of the /ɜ/. The final step is to group those channels containing peaks in the simplified

ACG at the delay of the second most prominent period. That has been done to produce Fig. 6(C) which contains channels near the f_0, F1 (650), and F2 (950) of the /a/.

Thus, by simplifying the ACG, it has been possible to group the channels containing the first two formants of the vowels appropriately, and to avoid creating spurious formants.

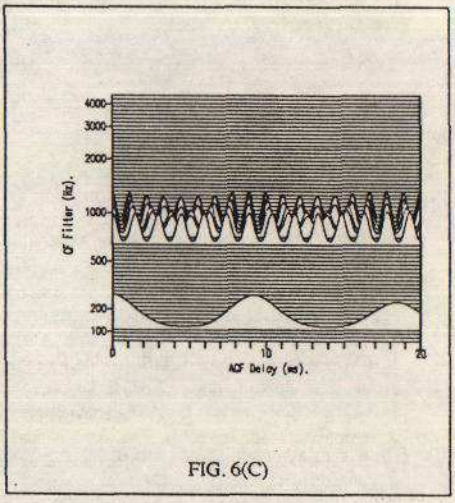

FIG. 6(C)

8. LIMITATIONS

To date, we have computed the simplified ACGs of only a subset of the double vowels whose perception was modelled by A&S and M&H. The procedure requires wider verification and it is likely that the parameter values specified above could be optimised further. A weakness of the procedure is its reliance on the simplifying assumption that the arches in ACGs follow smooth paths. This assumption is realised across channels excited by harmonics above about the 10th where filter bandwidths are broad in relation to the spacing of adjacent harmonics. It is not realised at lower frequencies where harmonics are resolved. The procedure works adequately when only one source is present in the low frequencies, as Fig. 5 shows. However, it sometimes fails to locate the F1 of one constituent of a double vowel if that formant is defined by only one or two resolved harmonics which are surrounded by harmonics of the competing vowel.

9. DISCUSSION

The strategy for segregating periodic sources by partitioning a simplified autocorrelogram is not intended as a description of how the auditory system might work. Rather, it is an exploration of the computational consequences of two assumptions about auditory analysis, both of which could be incorrect.

The first assumption is that ACGs underlie pitch perception and source segregation. One reason for caution in accepting this assumption is that ACGs are computationally expensive. Physiologically, their derivation would require the intervals between all spikes occurring in a channel to be measured [15]. Licklider [1] demonstrated that this computation could be carried out by a delay line and a set of coincidence detectors for each channel. Such neural machinery underlies binaural localisation [16,17], but has not been found in structures thought to be involved in pitch perception. Concern for these issues has prompted the search for alternatives to ACGs whose neural computation could be more feasible [18].

The second assumption is that the auditory system preferentially groups a component which could be a member of either of two concurrent harmonic series with components that are adjacent in frequency rather than remote. This assumption is plausible since it means that the adjacent harmonics which define a formant peak will generally be grouped together. However, it requires a solution to the problem of overlapping harmonics. The present paper demonstrates that procedures as elaborate as cross-channel convolution may be necessary to solve this problem. Therefore, the assumption may be wrong and auditory analysis may involve a different, simpler, strategy for partitioning channels.

One possibility, which often achieves the desired partitioning, is M&H's strategy [8], described in Section 6. An alternative could be to reduce the computational cost of simplifying the entire ACG by first establishing candidate pitches from the pooled ACG and candidate formants from peaks in the distribution of excitation across channels. Computationally expensive processes, such as cross-channel convolution, could then be applied locally to establish the pitch underlying each peak.

10. REFERENCES.

[1] J C R LICKLIDER, 'A Duplex Theory of Pitch Perception', Experientia 7 pp128-133 (1951).
[2] B C J MOORE, 'An Introduction to the Psychology of Hearing', Academic Press, London (1989).
[3] M SLANEY & R F LYON, 'A Perceptual Pitch Detector', Proc. ICASSP-90, pp357-360 (1990).
[4] R MEDDIS & M J HEWITT, 'Virtual Pitch and Phase Sensitivity Studied Using a Computer Model of the Auditory Periphery', J. Acoust. Soc. Am. (in press).
[5] J LAZZARO & C MEAD, 'Silicon Modelling of Pitch Perception', Neurobiology (in press).
[6] P F ASSMANN & Q SUMMERFIELD, 'Modeling the Perception of Concurrent Vowels: Vowels with the Same Fundamental Frequency', J. Acoust. Soc. Am. 85, pp327-338 (1989).
[7] P F ASSMANN & Q SUMMERFIELD, 'Modeling the Perception of Concurrent Vowels: Vowels with Different Fundamental Frequencies', J. Acoust. Soc. Am. 88, pp680-697 (1990).
[8] R MEDDIS & M F HEWITT, 'Modelling the Identification of Concurrent Vowels with Different Fundamental Frequencies', Submitted to J. Acoust. Sc. Am.
[9] B C J MOORE & B R GLASBERG, 'Suggested Formulae for Calculating Auditory-filter Bandwidths and Excitation Patterns', J. Acoust. Soc. Am. 74, pp750-753 (1983).
[10] R MEDDIS, 'Simulation of Mechanical to Neural Transduction in the Auditory Receptor', J. Acoust. Soc. Am. 79, pp702-711 (1986).
[11] M T M SCHEFFERS, 'Sifting Vowels: Auditory Pitch Analysis and Sound Segregation', unpublished Ph.D. Thesis, University of Groningen, The Netherlands (1983).
[12] A S BREGMAN, 'The Meaning of Duplex Perception: Sounds as Transparent Objects', in "The Psychophysics of Speech Perception" (M E H Schouten, Ed.), Martinus Nijhoff, Dordrecht. pp95-111 (1987).
[13] R D PATTERSON, 'A Pulse-ribbon Model of Monaural Phase Perception', J. Acoust. Soc. Am. 82, pp1560-1586 (1987).
[14] D GABOR, 'Theory of Communication', J. Inst. Elec. Eng. 93, pp429-459 (1946).
[15] A R PALMER, 'The representation of the spectra and fundamental frequencies of steady-state single- and double-vowels in the temporal discharge patterns of guinea pig cochlear-nerve fibers', J. Acoust. Soc. Am. 88, pp1412-1426 (1990).
[16] M KONISHI, T T KAHASHI, H WAGNER, W SULLIVAN, & C E CARR, 'Neurophysiological and anatomical substrates of sound localisation in the owl. In "Auditory Function" (G M Edelman, W E Gall, & W M Cowan, Eds.) Wiley,: New York. pp721-745 (1988).
[17] T C T YIN & J C K CHAN, 'Neural Mechanisms Underlying Interaural Time Sensitivity to Tones and Noise'. In "Auditory Function" (G M Edelman, W E Gall, & W M Cowan, Eds.) Wiley,: New York. pp385-430 (1988).
[18] R D PATTERSON & J HOLDSWORTH, 'A Computational Model of Auditory Image Construction'. To appear in, "Cochlear Nucleus: Structure and Function in Relation to Modelling" (W. Ainsworth, Ed.) JAI Press: London.

11. ACKNOWLEDGMENTS

The idea of using Gabor functions to simplify ACGs was suggested to us by Roland Baddeley of the Computational Vision Group at the University of Stirling. Peter Assmann contributed helpful discussions and some of the programs used to create the figures.