Journal of New Music Research, Vol. 23 (1994), pp. 000–000 0929-8215/94/2301-0000$6.00

Sound Segmentation Using Onsets and Offsets

Leslie S. Smith

ABSTRACT

Onset and offset filters inspired both by the neurobiology of the cochlear nucleus and the on-centre off-surround filters used in image processing are presented. These are used to segment single streams of sound (music or other). The techniques developed are applied to a plucked guitar, and to tongued and slurred saxophone and flute sounds, and the results are given. The limitations of these methods are discussed, and some further development directions for the methods employed are suggested.

1. INTRODUCTION

The aim of this work is the temporal segmentation of single-source sound, that is, the segmentation of a single stream of sound (in the sense of (Bregman 90)). The work here represents a small aspect of an overall auditory scene understanding or synthetic music comprehending system. The techniques used are entirely bottom-up: that is they are driven by the data itself, without explicit higher-level knowledge of the sound, or of what is expected in the way of segmentation, and are based on models of the inner ear and cochlear nucleus. They are applicable to any sound stream.

This begs the question: why segment at all? Given that the aim is to interpret or comprehend the sound, the alternatives would be either to process the sound stream all at once, or else in fixed-length sections. Most sound streams are too extended to permit the former approach, since one needs some results of the interpretation before the sound stream finishes. The latter approach would require the sections to be shorter than the shortest element of sound, and would present problems when a section contained parts of more than one sound element. It is the thesis of this paper that interpretation is carried out time-segment by time-segment, and that the time-segment endpoints are derivable from the gross structure of the sound stream.

The techniques described in section 2 are very general. Although inspired by the processing which appears to occur in the cochlea and cochlear nucleus, they are not neural models, in that their behaviour represents an abstraction of what the neural system appears to be doing rather than modelling the neurons themselves,

but nonetheless, they retain the general purpose nature of the neural original. Although they are very simple bottom-up techniques, they could easily be incorporated into a more complex system which was partly top-down. Section 3 applies the techniques to a variety of musical sounds. The results are discussed in section 4, and section 5 discusses how this work might be extended.

2. METHODS

The method used is outlined in Figure 1. The sound signal was acquired using an AKG D109 microphone, in an ordinary office environment. This was digitised at 22050 samples/second, 16 bits linear, using a Singular Solutions A/D64x. The resulting file was used as input to the AIM human auditory processing model's (Patterson and Holdsworth 1990) basilar membrane movement module. The parameters were set so as to give 32 bands of output, from 100 Hz to 10 kHz, with the audiogram switched off. This gives an output of 32 channels of digitally filtered signal, with the bands being approximately logarithmically spaced between the start and end frequencies. Each channel has a centre frequency, but is relatively wideband. The bands, and their widths are based on what is known of the human

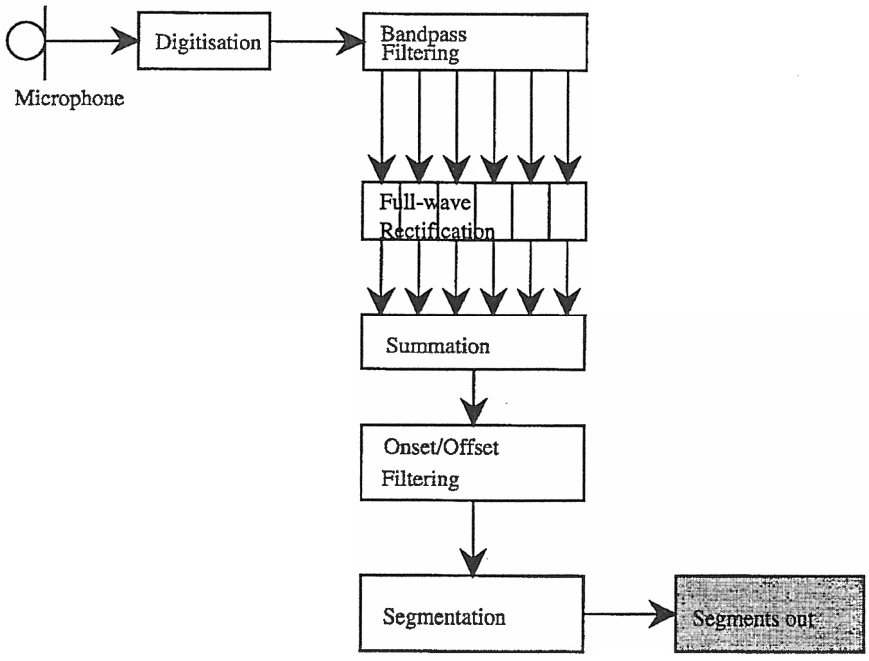


Fig. 1. Outline of the processing of the sound.

cochlea (see (Moore and Glasberg 1983)).

The output of each band was then full-wave rectified, and the outputs from all the bands summed. Rectification computes a measure of the energy in each band. It can be thought of as modelling in outline the effect of a population of inner cochlear hair cells. Summing all the bands gives a measure of the total signal energy, modelling, again in outline, the total activity in the auditory nerve. This summary signal was used as input to the onset/offset filters, described in section 2.1, and the output of this filter used to segment the signal, as discussed in section 2.2. It was possible to return to the original sound signal, and to hear what each segment actually sounded like. The ability to listen to each segment was critical in the development of the segmentation techniques, and in their assessment.

2.1. The onset/offset filter

The onset/offset filters used are an abstraction of cochlear nucleus cells responsive to onsets and offsets (Pickles 1988, Blackburn and Sachs 1990, Blackwood et al. 1989, Brown 92). Onset has also been found to be an important psychophysical grouping criterion (reviewed in (Brown 1992)). Brown suggests that the biological onset and offset cells are organised into a two-dimensional map, with one dimension being frequency, and the other excitatory (for onset) or inhibitory (for offset) delay. For this paper, we shall deal with a much simpler system, one in which the onset/offset filter is applied to the whole signal. As will be discussed in Section 5, this is really taking simplification too far.

The filters used are based on the ideas applied to visual processing in (Marr and Hildreth 1980). Two different filters have been experimented with, namely a difference of exponentials filter (DoE) and a half difference of Gaussians (HDoG). Both filters produce their results by taking the difference between an average over recent time, and an average over a longer sample of recent time. This approach is more noise-immune than one which uses differences directly.

Each average is computed by convolving the filter with the summary signal $s(x)$ described above:

$$A_z(t,k) = \int_0^t f_z(t-x,k)s(x)dx$$

where $f(x,k)$ is the convolving function. z is either E or G, depending on whether a DoE or an HDoG is being used. The k parameter of f defines the particular moving average being used. For the DoE filter:

$$f_E(x,k) = k\exp(-kx)$$

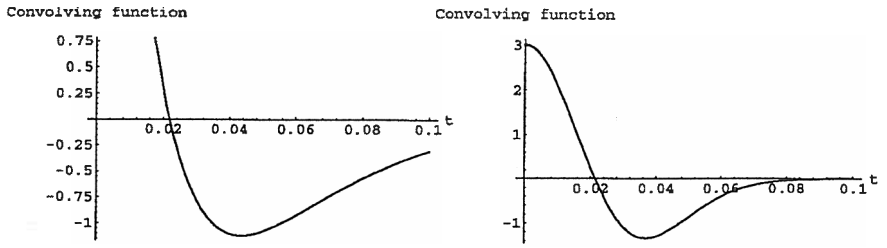


Fig. 2. Left shows graph of DoE convolving function ($f_E(t,k) - f_E(t,k/r)$), with $k = 50$ and $r = 1.2$. Right shows HDoG convolving function ($f_G(t,k) - f_G(t,k/r)$) with $k = 1200$ and $r = 1.2$.

and for the HDoG filter:

$$f_G(x,k) = \sqrt{k}\ \exp(-kx^2)$$

These particular functions have been chosen so that $\int_0^\infty f(x,k)dx$ is a constant independent of k. Both functions have their maxima at $x = 0$, and tail away rapidly towards 0 as x increases. Thus $A_z(t,k)$ is a moving average of $s(x)$, most strongly influenced by the value of $s(x)$ near t.

We define the onset/offset operator as the difference between a pair of averages. Such an operator could be defined in terms of three parameters, t, k_1 and k_2: however, noting that the positive average will always be a shorter-term average than the negative average, and because we want to consider families of such onset/offset operators, we define the operator as:

$$O(t,k,r) = A(t,k) - A(t,k/r) = \int_0^t (f_z(t-x,k) - f_z(t-x,k/r))\, s(x)dx$$

where $r > 1$. Thus, k defines the short-term average, and k/r defines the longer-term average. This filter has the appropriate property of giving a 0 output for constant input. The convolving functions are illustrated in Figure 2. Although they have the same intercept, the tail of the DoE filter is much longer.

2.2. Segmentation using the onset/offset filter output

The output from the filter rises at the start of a sound, and falls when the sound is decreasing. However, exactly how this should be used to perform segmentation of a sound is not immediately clear. One could consider a sound to start when the

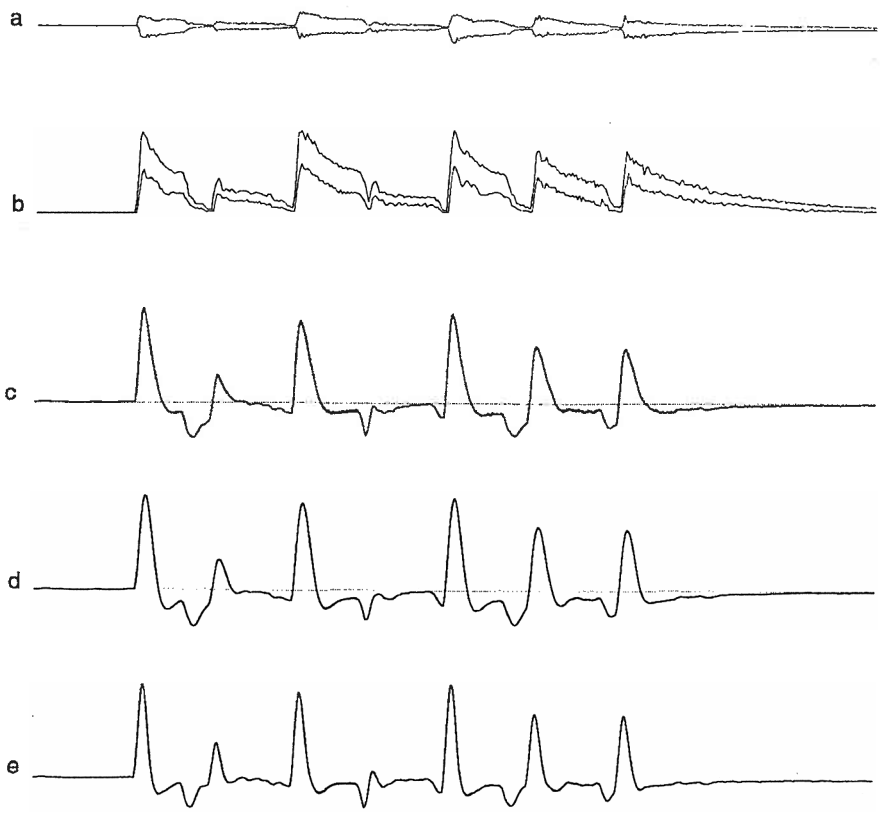


Fig. 3. The plucked guitar sound: a: the outline of the original signal. b: the signal after rectification and summation. c,d and e: the results of various onset/offset filters: c uses a DoE filter, with $r = 50$ and $k = 1.2$, d and e use an HDoG filter, with $k = 1.2$ and $r = 600$ and $r = 1200$ respectively.

output crosses 0: however, this makes the system very sensitive to any extraneous noise, and since, as is clear from Figure 3c to e, the rise in output is very steep for a sound of reasonable volume, we prefer to identify the start of a sound as being when the output exceeds some threshold.

Identifying the end of a sound is more difficult. Although some sounds have a clear finish (such as a note on a flute or a saxophone), many do not, and one often considers the end of the sound as being the start of a new sound, certainly when the notes are played quickly. Examples are notes played on a glockenspiel, or single plucked notes on a guitar. Indeed, the exact placing of the end of a sound is often rather arbitrary, particularly for short sounds in an echoing environment, such as a handclap in a bare room.

Using the output of the onset/offset filter, the immediately obvious choices for ending a segment are where the filter output recrosses 0, or where it has a negative minimum. The former marks the point at which the filter declares that the two averages are the same, which could be because sound is constant, or because the degree of onset matches the degree of offset (e.g., when a sound which had a recent onset is now decreasing in intensity), and the latter marks the maximum rate of offset. The former is clearly too early: the sound is still present, and the latter may well also be too early, since the maximum rate of offset of a sound may occur while the sound is still present. This is particularly true for sounds which have a rapid attack and decay at their start, followed by a slower sustained period, and a slow final decay. Many plucked and percussive sounds have this form (see Figure 3a and b).

To overcome this problem, a slightly more sophisticated approach was taken. Since the perceived end of a segment can appear to depend on what follows it, we considered pairs of adjacent segments produced using the negative minimum of the onset/offset filter to mark the segment end. These were generally not contiguous. To make them contiguous, the program must decide whether to extend the end of first segment forwards, extend the beginning of the second segment backwards, or to insert another "quiet" segment between the two segments. If there is a large gap between the two segments, it is reasonable to insert another segment to mark this gap. We need only decide on how large a gap we should consider to be large enough: we will write G_{quiet} for this gap length, in ms. After some experimentation, the following was found to be a reasonable compromise. Firstly, look forward from the minimum marking the initial estimate of the end of a segment to the crossing of a value which marks the start of the next segment, seeking out additional minima. Choose the last minimum which is within a certain predefined ratio (K_{minmin}) of the largest minimum and extend the first segment to this point. If the time between this point and the start of the next segment is less than G_{quiet}, then extend the next segment backwards to this point. Otherwise, insert an additional "quiet" segment from this new end of the first segment to the start of the second segment. This technique was used with $G_{\text{quiet}} = 50$ ms and $K_{\text{minmin}} = 0.4$.

3. EXPERIMENTAL RESULTS

We report here the results of applying these techniques to some single-note musical instruments. The instruments used are a single note at a time plucked guitar, a flute and a saxophone, played both tongued and slurred.

Figure 3 shows the effect of applying the processing technique to a plucked guitar. It is clear that the notes are plucked with varying intensities, though this is less obvious when listening. The transition from note to note is quite clear from

Table 1. The result of segmenting plucked guitar sound. ~~Left~~ table shows the sound segments as found by ear. ~~Right~~ table shows segmentation produced for one DoE filter, and for two different HDoG filters. Columns (a) show the segmentation start and end times using only the simple segmenting technique, and columns (b) show the use of the more sophisticated technique discussed in section 2.2. (1) marks those segments which are created by the addition of a "quiet" segment between two "note" segments. Times are in 0.5 ms units.

Filter	Segment	Start (a)	End (a)	Start (b)	End (b)
	0	631	870	0 (1)	631
DoE	1	1108	1259	631	989
$k = 50$	2	1602	1838	989 (1)	1108
$r = 1.2$	3	2538	2766	1108	1585
	4	3062	3271	1585	2520
	5	3617	3876	2520	2971
	6			2971	3563
	7			3563	4259
	0	637	845	0 (1)	637
HDoG	1	1116	1270	637	1003
$k = 600$	2	1610	1817	1003 (1)	1116
$r = 1.2$	3	2548	2742	1116	1587
	4	3070	3270	1587	2525
	5	3627	3827	2525	2982
	6			2982	3583
	7			3583	4185
	0	633	792	0 (1)	633
HDoG	1	1103	1226	633	982
k = 1200	2	1602	1778	982 (1)	1103
r = 1.2	3	2093	2169	1103	1585
	4	2539	2688	1585	2054
	5	3058	3238	2054	2520
	6	3614	3763	2520	2961
	7			2961	3561
	8			3561	4172

Note	Start	End
0	600	1020
1	1060	1560
2	1580	2020
3	2020	2480
4	2500	3000
5	3040	3580
6	3580	5180

Bottom
Top

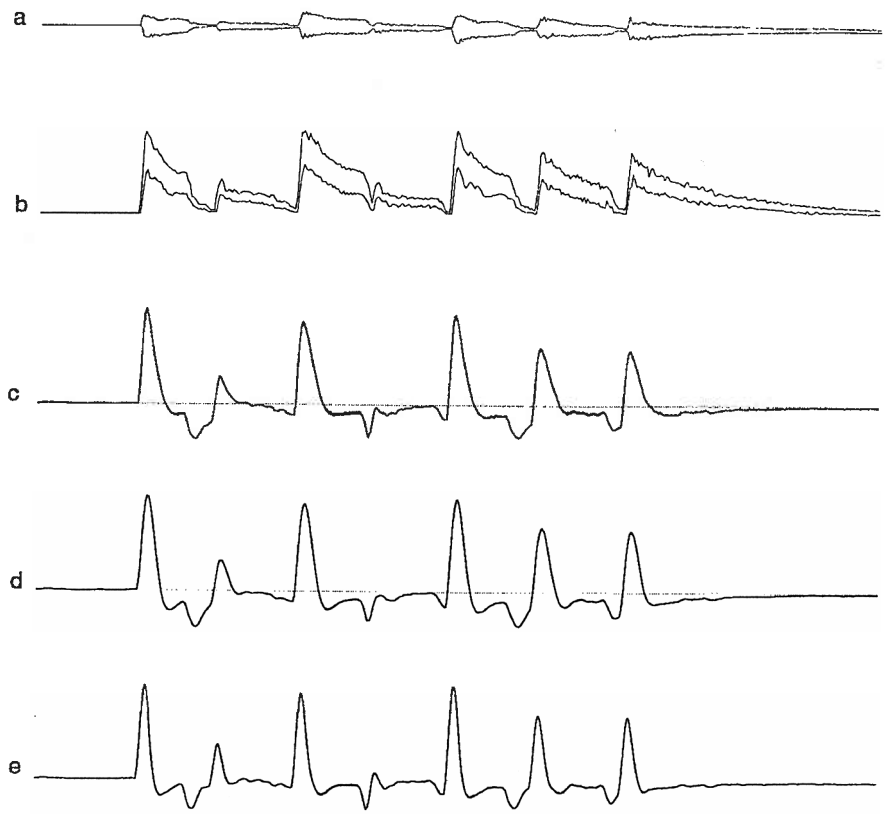


Fig. 3. The plucked guitar sound: a: the outline of the original signal. b: the signal after rectification and summation. c,d and e: the results of various onset/offset filters: c uses a DoE filter, with $r = 50$ and $k = 1.2$, d and e use an HDoG filter, with $k = 1.2$ and $r = 600$ and $r = 1200$ respectively.

the rectified summary signal: all the notes except note 3 result in a strong onset signal at the note's start, allowing the segmentation algorithm to pick out the notes. The onset of note 3 tends to get lost in the offset of note 2, unless a faster filter is used, as in Figure 3e. Table 1 shows the segmentations that result from the filters used in Figure 3c, d, and e. The first one uses a DoE filter with an intercept of 21 ms, and the second uses an HDoG with an intercept of 30 ms. These have very similar outputs, partly because of the extended tail of the DoE. The third filter uses an HDoG with an intercept of 21 ms. The short tail of the HDoG makes this a faster filter than the DoE with the same intercept, and as a result it picks out the onset of note 3, which otherwise gets lost in the offset of note 2.

Figure 4 shows the waveforms for tongued and slurred saxophone sounds. The ends of the tongued sounds are clearly visible, and this is reflected in the

Table 2. Segmentation of the saxophone sound by ear and by HDoG filter, with $r = 600$ and $k = 1.2$. (1) excludes a "quiet" segment from 1459 to 1712 and from 3449 to 3628. (2) uses HDoG filter with $k = 1200$, as this note was not found using the slower filter. Times are in 0.5 ms units.

Tongued saxophone					Slurred saxophone				
Note	Ear		HDoG		Note	Ear		HDoG	
	Start	Finish	Start	Finish		Start	Finish	Start	Finish
0	1280	1720	1316	1804	0	1120	1420	1146	1459
1	1800	2180	1804	2259	1	1420	1800	1712 (1)	1842
2	2200	2620	2259	2694	2	1800	2080	1842	2130
3	2700	3120	2694	3131	3	2080	2420	2130	2453
4	3120	3460	3131	3543	4	2420	2780	2453	2800
5	3540	3920	3543	3952	5	2780	3010	2840	3126
6	3920	5040	3952	4849	6	3020	3400	3126	3449
					7	3400	3720	3628 (1)	3824
					8	3760	4880	3810 (2)	4739 (2)

segmentation produced in Table 2. The slurred saxophone sound is more difficult to segment, both visually and by the techniques here: however, as can be seen in Table 2, the system does perform reasonably well.

Figure 5 shows the waveforms for tongued and slurred flute sounds. As with the saxophone, the tongued segments are quite visible, unlike those from the slurred notes. The system performs well on the tongued notes, and quite well on the slurred notes, except for the last two notes. For these, it is guided by the variations in the envelope, which do not correspond to the note changes.

4. DISCUSSION

The techniques used here can be applied to any sound. They have also been applied to speech in (Smith 1993). We do not pretend that this is how sound is really segmented in the human ear: although there is a basis in the biology for this approach, there are many more onset and offset cells and many other cells with different types of responses in the cochlear nucleus (Blackburn and Sachs 1989). What this work represents is the simplest possible application of an onset/offset based technique to the sound segmentation problem.

This approach is purely data-driven. It is completely ignorant of any higher-level information that might help to drive such a segmentation (such as prosody in speech (Cutler 1990), or information about the particular instrument). Such an approach may appear to be throwing out far too much that might be helpful:

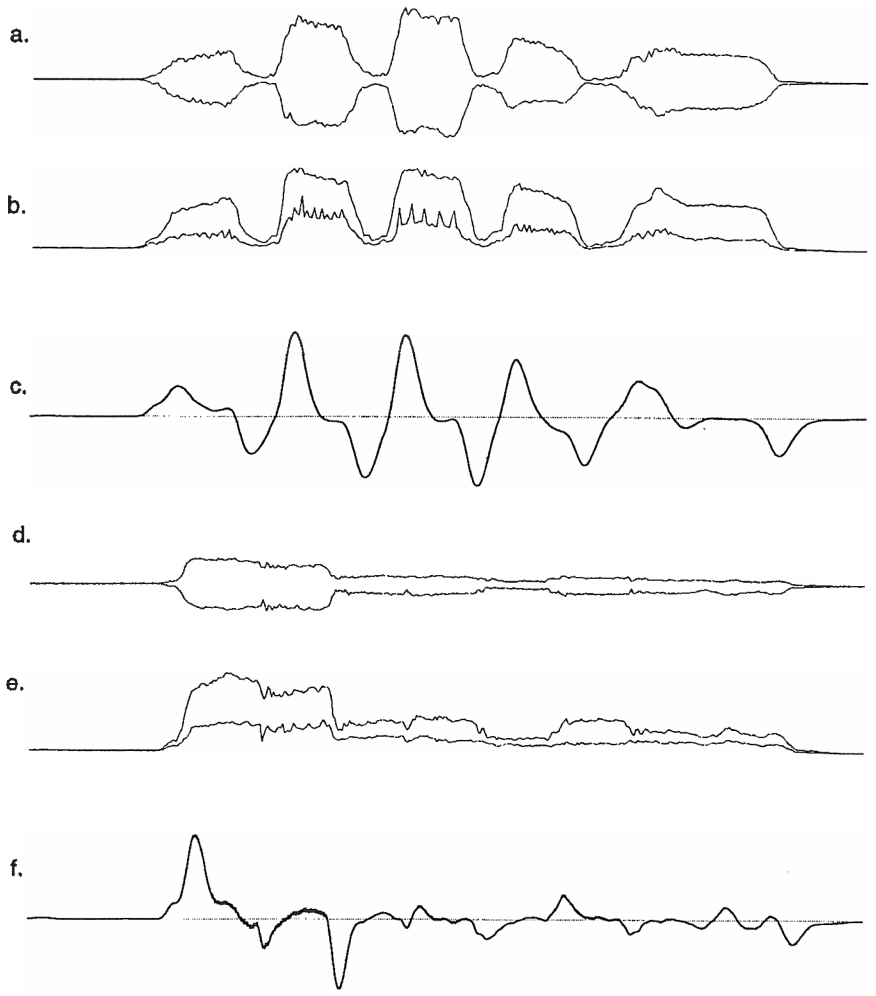


Fig. 5. Flute sound. a: original tongued flute sound. b: summary rectified signal from a. c: onset/offset signal output from HDoG filter with $k = 600$ and $r = 1.2$. d: original slurred flute sound signal. e: summary rectified signal from d. f: onset/offset signal output from HDoG filter with $k = 600$ and $r = 1.2$.

however, since the approach is entirely data-driven, it could be considered to provide a first approximation at segmentation, one which could later be modified using additional information. The nearest approximation to this system is in (Andre-Obrecht 1988), which uses changes in the statistical structure of the signal, as modelled by a parametric description of a sliding 8 ms window on the signal, and is applied to speech segmentation. Their system is again entirely data-driven,

but makes no attempt at all at biological plausibility. Unlike their system, we do not attempt to describe the statistical structure of the signal, concentrating purely on signal onsets and offsets. As such, a frequency glide will be invisible to this system: however, the initial results we have are, we believe, interesting enough to merit further research.

The system deals well with simple plucked sounds, and with tongued blown sounds. It is also successful with simple handclaps and a glockenspiel (Smith 1993). This is at least partly because it identifies the start and end of bursts of energy, and these percussive sounds are made up from well defined bursts of energy, with an envelope which is characterised by sudden onset, and more gradual (though frequently still quite rapid) offset. Given that the filter intercept approximates to the duration of the energy in the pulse, this technique works well. For sounds with a different envelope, the system may not work as well. Internal changes in intensity inside a musical note can divide up the note (as in Table 3 note (1)); however, if onset/offset filters with a range of intercepts were applied, then a filter whose intercept closely approximated the note length would not suffer from this problem. Certainly, when listening to a musical note, it is possible to be aware simultaneously that a note changes in intensity or tone, without considering it as more than one note, suggesting the existence of multiple concurrent timescales of interpretation.

The slurred saxophone and flute experiments point out one of the major problems with the simple system described here: a sequence of notes without any spaces between them will not be segmented correctly. This is because the system

Table 3. Segmentation of the flute sound by ear and by HDoG filter, with $k = 600$ and $r = 1.2$. (1) sound was actually broken into two segments, at $t = 808$ (tongued) and $t = 1129$ (slurred). (2) the segmentation breaks down completely here: the last two notes are broken into four pieces in a way which does not correspond to the actual notes. Times are in 0.5 ms units.

Tongued Flute					Slurred Flute				
Note	Ear		HDoG		Note	Ear		HDoG	
	Start	Finish	Start	Finish		Start	Finish	Start	Finish
0	480	960	492	971 (1)	0	800	1360	792	1392 (1)
1	1000	1440	1076	1467	1	1400	1890	1531	1824
2	1520	1920	1569	1964	2	1900	2180	2016	2217
3	2000	2360	1954	2417	3	2180	2640	2217	2687
4	2500	3220	2534	2852	4	2640	3010	2932	3020
					5	3010	3580	3020	3296 (2)
					6	3580	4460	3296	3928
								3928	4232
								4232	4457

considers only the whole spectrum sound. Clearly, replacing the single filter by a range of filters, each sensitive to some part of the frequency band would alter this, and it is clear that biological systems have a range of onset and offset detectors, innervated by different parts of the auditory nerve.

5. CONCLUSIONS AND FURTHER WORK

A prototype system which can segment some musical sounds played by single-note instruments has been demonstrated. The system is only a prototype in the sense that it uses only one onset/offset filter applied to the whole summary rectified signal. Realistically, a range of filters with varying intercepts should be applied. These would give a number of onset/offset filter outputs, reflecting the segmentation structure for different sizes of possible segment.

In addition, the filters should not be applied to the whole signal, but only to some part of the spectrum of the signal. The receptive field of each filter could be some continguous range of the spectrum, or could consist of a number of harmonically related small contiguous ranges. A claim for biological plausibility could be made for both of these receptive fields, since auditory nerve signals for nearby frequencies are frequently coactive, as are auditory nerve signals for harmonically related signals. Correlated activity in nerve fibers is believed to be an important organisational guide in cortical network organisation (von der Malsburg and Singer 1988), so that an onset or an offset cell is likely to be innervated by an appropriate subset of the auditory nerve fibers.

From Figures 3, 4, and 5, the summary rectified signal appears (at least visually) to contain important information about the segments in the original sound. Other, less biologically justifiable, techniques could be applied to the summary rectified signal, such as the "weak string" of (Blake and Zisserman 1987) to perform segmentation. However, from the point of view of understanding how humans segment sound, techniques based on cochlear nucleus cell behaviour are more reasonable.

ACKNOWLEDGEMENTS

I thank Lawrence Gerstley for supplying the sounds, and the other members of the Centre for Cognitive and Computational Neuroscience for useful discussion, and the University of Stirling for a 6-month sabbatical during which much of this work was done.

REFERENCES

Andre-Obrecht R. (1988). A new statistical approach for the automatic segmentation of continuous speech signals. IEEE Trans Acoustics. *Speech and Signal Processing*, 36, 1.

Blackburn C.C. & Sachs M.B. (1989). Classification of unit types in the anteroventral cochlear nucleus: PST histograms and regularity analysis. *Journal of Neurophysiology*, 62, 6.

Blackwood N., Meyer G. & Aimsworth W. (1990). A Model of the processing of voiced plosives in the auditory nerve and cochlear nucleus. *Proceedings Institute of Acoustics*, 12, 10.

Blake A. & Zisserman A. (1987). *Visual Reconstruction*. MIT Press.

Brown G. (1992). *Computational Auditory Scene Analysis*, TR CS-92-22. Department of Computing Science, University of Sheffield, England.

Bregman A.S. (1990). *Auditory Scene Analysis*. MIT Press.

Cutler A. (1990). Exploiting prosodic probabilities in speech segmentation. In: G.T.M. Altmann (ed.) *Cognitive Models of Speech Processing*, MIT press.

von der Malsburg C. & Singer W. (1988). Principles of cortical network organisation. In: P. Rakic and W. Singer (eds.). *Neurophysiology of Neocortex*, John Wiley and Sons.

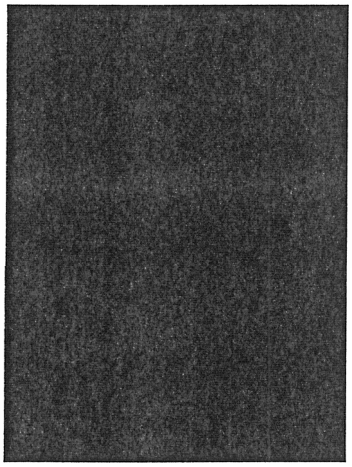
Marr D. & Hildreth E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London B*, 207, 187–217.

Moore B.C.J. & Glasberg B.R. (1983). Suggested formulae for calculating auditory-filter bandwiths and excitation patterns. *J Acoust Soc America*, 74, 3.

Patterson R. & Holdsworth J. (1990). *An Introduction to Auditory Sensation Processing*. In AAM HAP, Vol 1, No 1.

Pickles J.O. (1988). *An Introduction to the Physiology of Hearing*. 2nd Edition, Academic Press.

Smith L.S. (1993). *Temporal Localisation and Segmentation of Sounds Using Onsets and Offsets*. CCCN Technical Report CCCN-16, University of Stirling.



Leslie Smith
Centre for Cognitive and Computational Neuroscience
Department of Computing Science and Mathematics
University of Stirling
Stirling FK9 4LA
Scotland
email: lss@compsci.stirling.ac.uk

Born in 1952, in Glasgow, Scotland, he received his Ph.D from Glasgow University in Computing Science in 1981, after graduating B.Sc. in Mathematics in 1973. He was a founder member of the Centre for Cognitive and Computational Neuroscience at the University of Stirling, an interdisciplinary research group researching at the borders between Cognitive Psychology, Computer Science, and Neuropsychology. His research interests are in the design and analysis of neural networks, particularly from the viewpoint of sensory perception, and in the neural preprocessing of sensory data, particularly sound. A keen amateur musician, he performs in folk and jazz ensembles.