To appear in Proceedings of NEURAP95,
Marseilles, France, Dec 13-15, 1995.

Using an Onset-based Representation for Sound Segmentation

Leslie S. Smith
Centre for Cognitive and Computational Neuroscience
Department of Computer Science
University of Stirling
Stirling FK9 4LA
Scotland
email: lss@cs.stir.ac.uk

Abstract

We present a technique for using pre-processing based on mammalian early auditory processing to produce a segmentation of sound based on onsets and offsets. The sound signal is bandpassed and each band processed to enhance onsets and offsets. The onset and offset signals are compressed, then clustered both in time and across frequency channels using a network of integrate-and-fire neurons. A spike-based representation of onsets and offsets is produced, and the timing of these spikes used to segment the sound. By considering spikes in varying number of bands, a multi-level segmentation tree can be built. This tree is a purely data-driven representation of the segmental structure of the sound.

1 Background

Traditional techniques for speech interpretation are two stage processes. An internal representation is formed based on Fourier transforms and recoding of the spectrum; then a hidden Markov model or neural network interpretation stage is applied (Morgan and Bourlard 95). This approach has serious limitations both in interpreting continuous speech, and speech in the presence of noise. This has led to interest in more sophisticated internal representations (Ellis and Rosenthal 95), and in particular to front ends modelling biological auditory systems for speech interpretation systems (Ainsworth and Meyer 92; Cosi 93; Cole et al 95).

Auditory modelling systems use similar early auditory processing to that used in biological systems. Mammalian auditory processing uses two ears, and the incoming signal is filtered first by the pinna (external ear) and the auditory canal before causing the tympanic membrane (eardrum) to vibrate. This vibration is then passed on through the bones of the middle ear to the oval window of the cochlea. Inside the cochlea, the pressure wave causes a pattern of vibration to occur on the basilar membrane. This appears to be an active process using both the inner and outer hair cells of the organ of Corti. The movement is detected by the inner hair cells and turned into neural impulses by the neurons of the spiral ganglion. These pass down the auditory nerve, and arrive at various parts of the cochlear nucleus. From there, nerve fibres innervate other areas: the lateral and medial nuclei of the superior olive, and the inferior colliculus, for example. (See (Pickles 88)).

The representation used in traditional speech interpretation systems uses a form of bandpass filtering, following the biology at least as far as the cochlea. Generally, a Fourier transform is used to perform a calculation of the energy in each band over some time period, usually betwen 25 and 75 ms, producing an energy vector up to 100 times per second. This is not at all what the cochlea does. Auditory modelling front ends differ in the extent to which they follow animal early auditory processing, but the term generally implies at least that wideband filters are used initially, and generally that the initial internal representation maintains a high temporal resolution. This implies the use of filtering techniques, rather than Fourier transforms in the bandpass stage. Such filtering systems have been implemented by Patterson and Holdsworth (Patterson and Holdsworth 90; Slaney 93), and placed directly in silicon (Lazzaro and Mead 89; Lazzaro et al 93; Liu et al 93; Fragniere and van Schaik 94).

Some auditory models have moved beyond cochlear filtering, and have a number of different representational stages. The inner hair cell has been modelled by either simple rectification (Smith 94) or has been based on the work of (Meddis 88) for example (Patterson and Holdsworth 90; Cosi 93; Brown 92), leading to a represenation of the activity on the auditory nerve. Lazzaro has experimented with a silicon version of Licklider's autocorrelation processing (Licklider 51; Lazzaro and Mead 89). Others such as (Wu et al 1989; Blackwood et al 1990; Ainsworth and Meyer 92; Brown 92; Berthommier 93; Smith 94) have considered the early brainstem nuclei, and their possible contribution, based on the neurophysiology of the different cell types (Pickles 88; Blackburn and Sachs 1989; Kim et al 90). This provides a representation based on the features which appear to be detected in these nuclei.

Auditory model-based systems have yet to find their way

into mainstream speech recognition systems (Cosi 93). The work presented here uses auditory modelling up to onset cells in the cochlear nucleus. It adds a temporal neural network to clean up the onset representation produced. This part has been filed as a patent (Smith 95.1). Though the system has some biological plausibility, it is primarily aimed at producing effective data-driven segmentation. By making the segmentation part of the internal representation, we allow the later interpretation stage to work on the different segments independently, so that non-speech noise (such as a door slamming) can be identified and appropriately processed before an interpretation of a complete noisy utterance is attempted. By forming a segmentation tree, we allow several segmentation possibilities to be considered.

2 Techniques used

The overall architecture is illustrated in figure 1. Digitised sound was applied to an auditory front end which uses a Gammatone filterbank, (Patterson and Holdsworth 90), which bandpassed the sound into a number of channels each with bandwidth $24.7(4.37F_c+1)$Hz, where F_c is the centre frequency (in KHz) of the band (Moore and Glasberg 83). These were rectified, modelling the effect of a set of inner hair cells. The signals produced bear some resemblance to those in the auditory nerve: although there are far more channels in the real system, each nerve channel carries only spike-coded information so that the coding used models the signal in a population of neighbouring auditory nerve fibres.

2.1 The onset/offset filter

The signal present in the auditory nerve is stronger near the onset of a tone than later (Pickles 88). This effect is much more pronounced in the firing patterns of certain cell types of the cochlear nucleus: some of these fire strongly just after the onset of a sound in the band to which they are sensitive, and are then silent. This emphasis on onsets was modelled by convolving the signal in each band with a filter which computes two averages, a more recent one, and a less recent one, and subtracts the less recent one from the more recent one. One biologically possible justification for this is to consider that a neuron is receiving the same driving input twice, once excitatorily, and the other inhibitorily; the excitatory input has a shorter time-constant than the inhibitory input. Both exponentially weighted averages, and averages formed using a Gaussian filter have been tried (Smith 94), but the former place too much emphasis on the most recent part of the signal, making the latter more effective.

The filter output for input signal $s(x)$ is

$$O(t,k,r) = \int_0^t (f(t-x,k) - f(t-x,k/r))s(x)dx \quad (1)$$

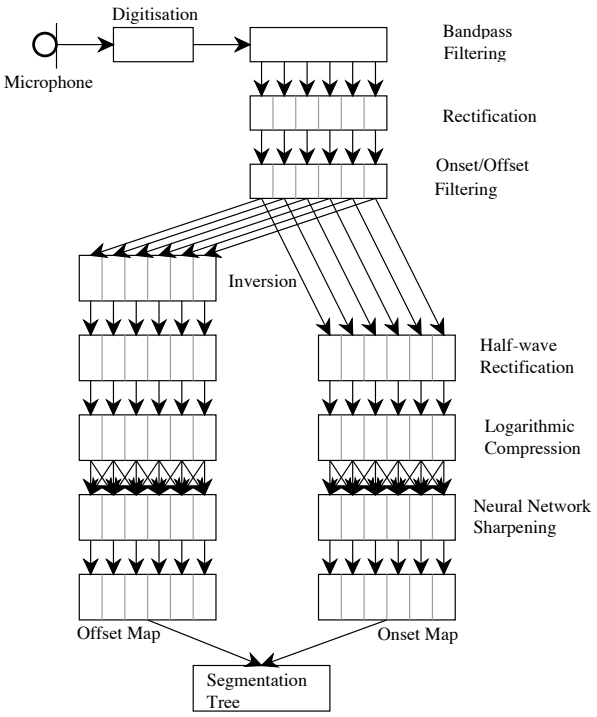

Figure 1: The overall sound segmentation system

where $f(x,k) = \sqrt{k}\exp(-kx^2)$. k and r determine the rise and fall times of the pulses of sound that the system is sensitive to. We used $k = 1.2$, $r = 1000$, so that the SD of the Gaussians are 24.49ms and 22.36ms. The convolving filter has a positive peak at 0, crosses 0 at 22.39ms, and is then negative. A positive onset/offset signal implies that the bandpassed signal is increasing in intensity, and a negative onset/offset signal implies that it is decreasing in intensity. The convolution used is a sound analog of the difference of Gaussians operator used to extract black/white and white/black edges in monochrome images (Marr and Hildreth 80). With these values, the transformed representation is sensitive to energy rises and falls which occur frequently in everyday sound. In (Smith 94) we performed sound segmentation directly on this representation.

For the work reported here, the onset/offset representation was divided into two positive-going signals, an onset signal consisting of the positive-going part, and an offset signal consisting of the positive-going part of the inverted onset/offset signal. Both were compressed logarithmically (where $\log(x)$ was taken as 0 for $0 \leq x \leq 1$). This increases the dynamical range of the system, and models compressive biological effects. This compressed onset signal models the output of a population of onset cells. This technique for producing an onset signal is related to that of (Wu et al 1989; Cosi 93), differing primarily in how the system approximates the biology.

2.2 The integrate-and-fire neural network

To provide a clean representation of the sound using the onset and offset data, this data needs to be integrated across frequency bands and across time. This temporal and tonotopic clustering was achieved using a network of integrate-and-fire units. An integrate-and-fire unit accumulates its weighted input over time. The activity of the unit A, is initially 0, and alters according to

$$\frac{dA}{dt} = I(t) - \gamma A \tag{2}$$

where $I(t)$ is the input to the neuron and γ, the dissipation, describes the leakiness of the temporal integration. When A reaches a threshold, the unit fires (i.e. emits a pulse), and A is reset to 0. After firing, there is a period of insensitivity to input, called the refractory period. Such neurons are discussed in detail in, e.g. (Mirolla and Strogatz 90; Gerstner 95).

One integrate-and-fire neuron was used per channel: this neuron received input either from a single channel, or from a set of adjacent channels, all with equal positive weightings. The output of each neuron was fed back to a set of adjacent neurons, again with a fixed positive weight, one time step (here 0.5ms) later. Because of the leaky nature of the accumulation of activity, excitatory input to the neuron arriving when its activation is near threshold has a larger effect on the next firing time than excitatory input arriving when activation is lower. Thus, if similar (but non-identical) input is applied to a set of neurons in adjacent channels, the effect of the inter-neuron connections is that when the first one fires, its neighbours fire almost immediately. This allows a network of such neurons to cluster the onset or offset signals, producing a sharp burst of spikes across a number of channels and so providing an unambiguous representation of onsets or offsets.

The external and internal weights of the network were adjusted so that onset or offset input alone allowed neurons to fire, while internal input alone was not enough to cause firing. The refractory period used was set to between 50 and 75ms for the onset system, and 5ms for the offset system. For the onset system, the effect was to produce sharp onset firing responses across adjacent channels in response to a sudden increase in energy in some channels, thus grouping onsets both tonotopically and temporally. This is appropriate for onsets, as these are generally brief and clearly marked. The output of this stage we call the onset map. Offsets tend to be more gradual. This is due to physical effects: for example, a percussive sound will start suddenly, as the vibrating element starts to move, but die away slowly as the vibration ceases (see (Gaver 93) for a discussion). Even when the vibration does stop suddenly, the sound will die away more slowly due to echoes from sound-reflecting surfaces. Thus we cannot reliably mark the offset of a sound: instead, we reduce the refractory period of the offset neurons, and produce a train of pulses marking the duration of the offset in this channel. We call the output of this stage the offset map.

2.3 Segmenting the sound

Both the onset and offset maps consist of nearly-vertical lines of spikes across a number of channels (see figure 2b). This allows the onsets to be used for segmentation. The simplest technique is to divide up the continuous speech at each onset; additionally, one can use the offset map to find the end of a segment. However, one needs to ensure that onsets which occur near to each other in time do not result in very short segments and that the occasional onset in a single channel does not confuse the system. To achieve this we set the minimum segment length to 25ms, and counted the number of onsets (or offsets) which took place inside 10ms. We could then vary the sharpness of the segmentation by specifying the minimum number of onset (offset) spikes which had to occur in the 10ms window before that onset or offset line was taken to signal a segment start (end).

3 Results

As the technique is entirely data-driven, it can be applied to sound from any source. It has been applied to both speech and musical sounds. Figure 2 shows the effect of applying the techniques discussed to a short piece of speech. The straight vertical lines in figure 2b (compared to the more random spikes in figure 2a) show the effect of the neural network integrating the onset timings across all the channels.

Segmenting the onset and offset map gives the results shown in figure 3. If too many near-coincident onsets are required in different bands, the first part of the word is lost; otherwise, the entire word is segmented. The original word, /naɪn/, has a stronger onset at the /a/ than at the /n/, as can be seen from figure 2b. The degree of segmentation depends on the number of bands required to contain a spike before a segment boundary is declared: if 25 bands are required, only one segment is produced, /aɪn/; if 13 bands are required, two segments are found, /n/, /aɪn/; if four bands are required, the segments are /n/, /na/, /aɪn/; if only two bands are required, we get /n/, /na/, /aɪ/, /ɪ/, /ən/. The concept of the segmental tree is described in (Cosi 93), and is also used in pyramidal descriptions of visual scenes (Asada and Brady 86).

A similar approach has been taken with longer continuous speech utterances, such as the author saying "Department of Computing Science". In this case, the segmentation found for differing numbers of bands is shown in figure 4. The segmentation found if we require 25 near-coincidental spikes was

/di/, /pa:mənt/, /ɔvkɔmpjutɪŋsajɛns/

and if we require 15 spikes, the segmentation was

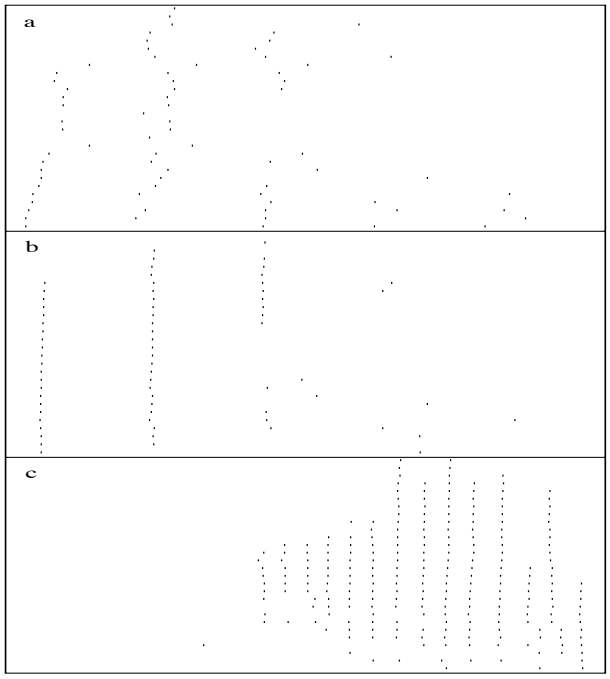

Figure 2: Onset and Offset maps from author saying "Nine". a: onset map, from 28 channels, from 80Hz-6KHz. Onset filter parameters as in text; one neuron per channel, with no interconnection. Neuron refractory period is 75ms. b: as a, but network has input applied to 12 adjacent channels, and internal feedback to 12 channels. c: offset map produced similarly, with refractory period 5ms.

/di/, /pa:/, /mənt/, /ɔv/, /kɔmpj/, /jut/, /ɹŋ/, /s/, /ajɛns/

and if we require 6 spikes, the segmentation was

/di/, /pa:/, /mən/, /t/, /ɔv/, /kɔm/, /p/, /ju/, /t/, /ɹŋ/, /s/, /aj/, /ɛn/, /s/

The technique has been successfully applied to many different utterances.

To investigate the relationship between phonemes and these segments, the technique was applied to some of the TIMIT database, a corpus of short (2-5 second) utterances of continuous speech. A subset consisting of 32 utterances was used (16 male, 16 female, two from each dialect region). Table 1 summarises the results found. It is clear that the phonemes and the segments found are different. The segmentation tends to start segments at stop consonants: even when the number of spikes required to signal a segment start is 12 (so that the number of segments is considerably less than the number of phonemes), 69% of stops are at the start of a segment. Finding phonemes and segmenting utterances are different tasks: the segmentation system breaks the utterance into sections whose start is signalled by a concurrent increase in energy in a number of channels, and whose end is signalled by either the concurrent decrease in energy in a number of channels, or by the start of the next segment. Phonemes, on the other hand, are the smallest units of speech that serve to distinguish one utterance from another: although some do coincide with major increases in energy in some parts of the spectrum, many do not. When the number of spikes required to signal a segment start is such that the number of segments is considerably less than the number of phonemes, segment startpoints are generally near phoneme startpoints as can be seen in table 2.

The same system has been used to segment sound from single musical instruments. Where these have silences between notes this is straightforward: in (Smith 94) correct (i.e. note by note) segmentation was achieved directly from

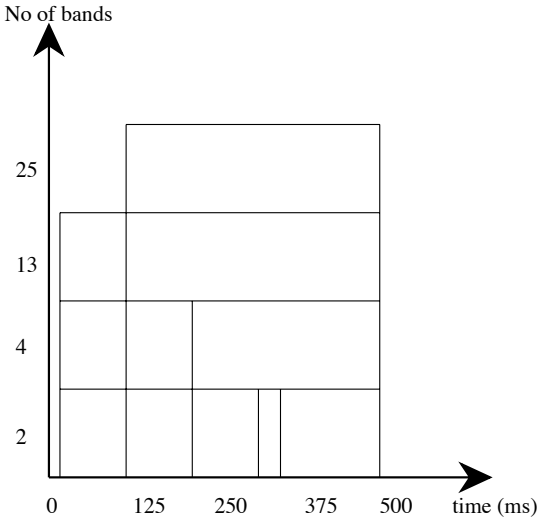

Figure 3: Segmental tree for utterance "nine" found by varying the number of near-coincident onset and offset signals in different bands required to signal a segment boundary.

Phon type	N=4			N=8			N=12		
	F	M	X	F	M	X	F	M	X
stop	73	24	3	72	25	3	69	29	3
affr	93	0	7	86	7	7	50	43	7
fric	32	49	19	27	61	12	17	74	9
sv, gl	44	20	35	31	44	25	21	67	12
vowel	53	20	27	45	31	24	32	49	19

Table 1: Percentage of different types of phoneme found (F) missed (M) or incorrectly found (X) for three values (4, 8, and 12) of number of spikes required to signal a segment start (N). Utterances were filtered into 28 bands from 100Hz to 6000Hz. Phonemes are counted as found if the segment start and the phoneme start are within 15ms of each other, as missed if the nearest segment start was assigned to a different phoneme, and as incorrectly found if the nearest segment boundary is more than 15ms away from the phoneme start.

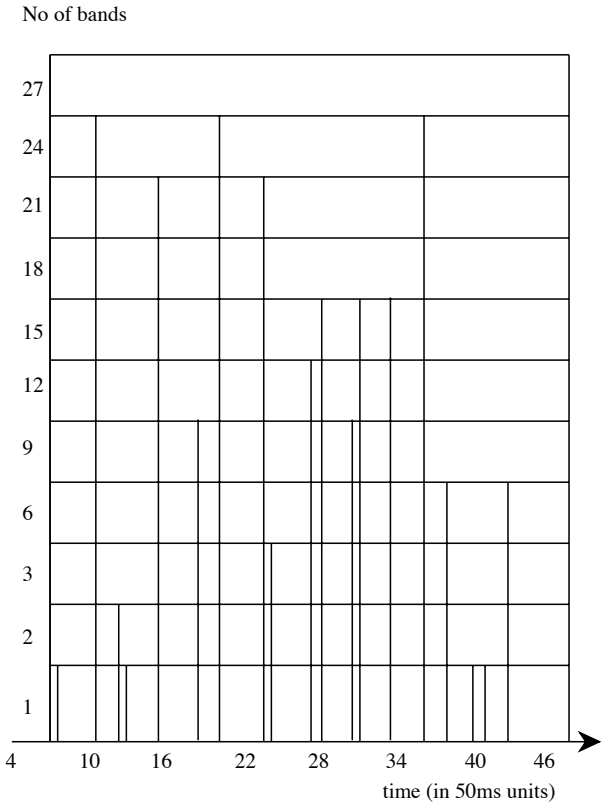

Figure 4: Segmental tree found for utterance "Department of Computing Science", using different numbers of near-coincident onset and offset signals in different bands to signal a segment boundary.

N	Segments starting near phonemes		
	$\Delta t \leq 15$ms	$\Delta t \leq 20$ms	$\Delta t > 20$ms
8	0.62 ± 0.13	0.69 ± 0.13	0.31 ± 0.13
12	0.70 ± 0.16	0.76 ± 0.15	0.22 ± 0.15

Table 2: Fraction of segment startpoints (in each utterance) near phoneme startpoints for varying number of spikes required to signal a segment start (N) and time interval between segment and phoneme start (Δt).

the onset/offset signal. This was not achieved for slurred sounds, in which the note changes without an intervening silence. For musical sounds, better results were obtained when the input to the network was not spread across channels.

We present results for two different instruments: a flute and a spanish guitar. For the flute, there is a clearly correct result: namely, each note should be one segment. However, one might also want the segmentation to be able to reflect phrasing, so that some notes might be grouped together. In addition, notes on a flute are not necessarily constant once started, so that one might also expect some notes to be oversegmented if the segmentation is made too sensitive. Figure 5 shows the onset map produced from a brief snatch of slurred flute. Using the onset and offset map, a segmental

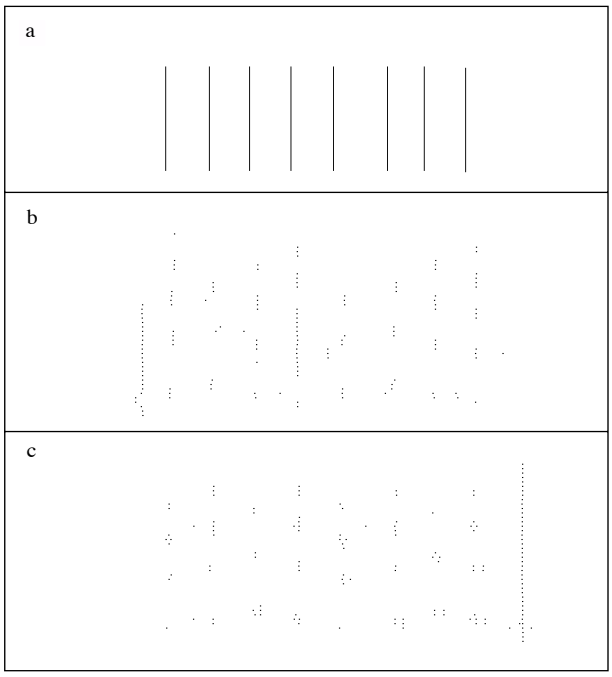

Figure 5: a: envelope of original flute sound. b: onset map. Bands are from 200 to 2500 Hz. Bandpass tuning is narrower than in text. c: offset map.

tree was produced, and is shown in figure 6. When 15-20 near-coincindent onset or offset spikes are required to signify a segment boundary, the sound breaks up into two phrases: when only 4 or 6 are required, the sound is broken up into its constituent notes. When 2 or less are required, the some of the notes are oversegmented.

For the spanish guitar music, the appropriate segmentation is more debatable. The different notes have different volumes, and notes do not always stop when a new note starts. Here, the segmental tree shows up different aspects of the sound when different numbers of near-coincident onset and offset spikes are used. Figure 7 shows the onset and offset maps. The onsets are very sharp; however, offsets are more gradual, starting almost immediately the note is produced, and continuing until the note dies away. This form of onset and offset is also typical of percussive sounds. As can be seen from figure 8, the segmentation produced depends on the number of coincident onsets or offsets required to signify a boundary. For large numbers, the phrasing is important: at smaller numbers, the main notes are segmented; at still smaller numbers, each note shows up as a different segment, sometimes including segments made up from the noise of the fingers plucking the string.

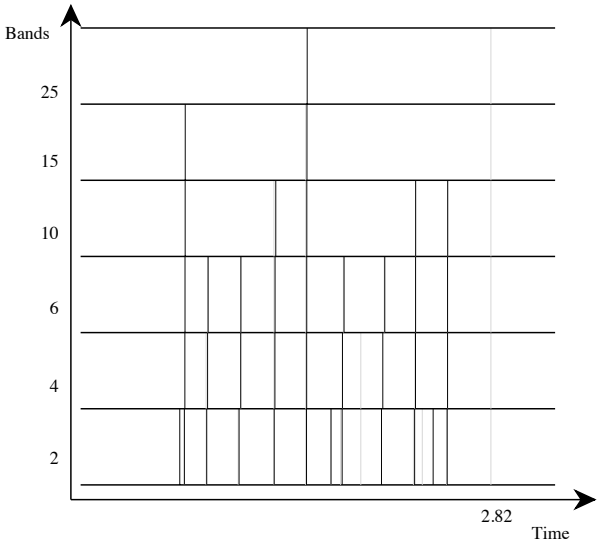

Figure 6: Segmental tree for the snatch of flute music. Grey lines mark ends of segments.

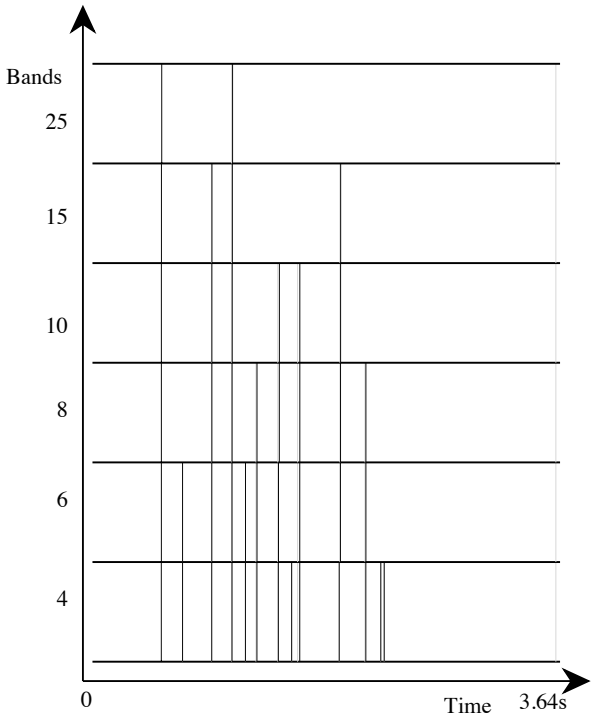

Figure 8: Segmental tree for the snatch of spanish guitar music.

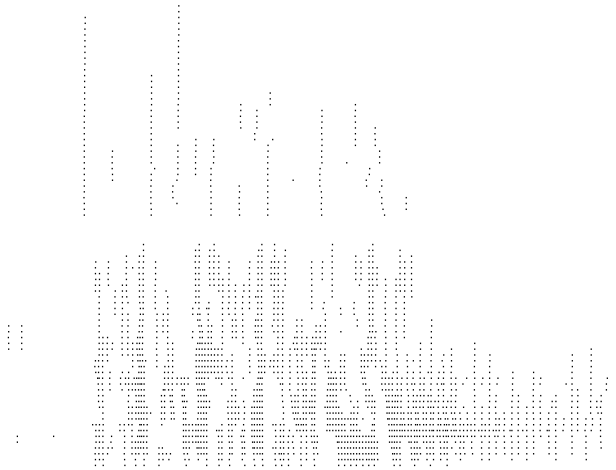

Figure 7: Top: envelope of original guitar sound. Middle: onset map. Bands are from 200 to 2500 Hz. Bottom: offset map.

4 Conclusions and further work

An effective data driven segmentation technique based on onset feature detection and using a network of integrate-and-fire neurons has been demonstrated. Although not discussed here, the system is also relatively immune to broadband noise. Segmentation is not an end in itself: the effectiveness of any technique will depend on the eventual application. From the experiments with the TIMIT dataset, it is clear that this data-driven segmentation technique tends to break the sound up at points which sometimes coincide with phoneme boundaries (particularly stops), but which often do not. This technique is probably most useful for an intiial breaking down of continuous speech into manageable segments, prior to further interpretation. It is important to note that the phoneme boundaries supplied with the TIMIT data were produced by a mixture of top-down and bottom-up processing, including some human intervention (Senneff and Zue 88).

The appropriateness of a segmentation will depend on the application, so that the capacity to produce a segmental tree is useful as it allows the same technique to be used to process sounds for different applications. The segmentation tree allows the selection of segment boundaries to be influenced by top-down processing. Further work is required to examine the interaction between the parameters of the onset/offset filter, the parameters of the integrate-and-fire network, and the precise way in which the segmental tree is derived.

Not all the information in the onset and offset maps is being used. In particular, the information on precisely which bands the onsets and offsets occur in is ignored. This could be useful where segments can overlap, as in the spanish guitar case above. We are extending this work by combin-

ing the segmentation described here with work segmenting sound using the presence of amplitude modulation across a number of wideband filtered bands to detect voiced sounds (Smith 95.2). The will allow us to extract sound segments from some subset of the bands, allowing segmentation and a simple form of streaming to run concurrently. The author is hoping to implement some of the techniques described directly in silicon, in conjunction with the Department of Electrical Engineering at the University of Edinburgh.

Acknowledgements

Many thanks are due to the members of the Center for Cognitive and Computational Neuroscience at the University of Stirling.

References

Ainsworth W, Meyer G. Speech analysis by means of a physiologically-based model of the cochlear nerve and cochlear nucleus, in *Visual representations of speech signals*, Cooke M, Beet S, eds, 1992.

Asada H., Brady M., The curvature primal sketch, *IEEE Trans Pattern Analysis and Machine Intelligence*, PAMI-8, 1, 2-14, 1986.

Berthommier F., Modelling neural responses of the intermediate auditory system, in *Mathematics applied to biology and medicine*, Demongeot J, Capasso V, Wuertz Publishing, Canada, 1993.

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

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

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

Cole R., et al, The challenge of spoken language systems: research directions of the 90's, *IEEE Trans Speech and Audio Processing*, 3, 1, 1995.

Cosi P. On the use of auditory models in speech technology, in *Intelligent Perceptual Models*, LNCS 745, Springer Verlag, 1993.

Ellis D., Rosenthal D., Mid-level representations for computational auditory scene analysis, to be presented at the workshop on computational auditory scene analysis, IJCAI, Montreal, 1995.

Fragniere E., van Schaik A., Linear predictive coding of the speech signal using an analog cochlear model, MANTRA Internal Report, 94/2, MANTRA Center for Neuro-mimetic systems, EPFL, Lausanne, Switzerland, 1994.

Gaver W.W. What in the world do we hear?: an ecological approach to auditory event perception, *Ecological Psychology*, 5(1), 1-29, 1993.

Gerstner W. Time structure of the activity in neural network models, *Physical Review E*, 51, 1, 738-758, 1995.

Kim D.O. , J.G. Sirianni, S.O.Chang, Responses of DCN-PVCN neurons and auditory nerve fibres in unanesthetized decerebrate cats to AM and pure tones: analysis with autocorrelation/power-spectrum, *Hearing Research*, 45, 95-113, 1990.

Lazzaro J., Mead C., Silicon modelling of pitch perception, *Proc Natl. Acad Sciences*, USA, 86, 9597-9601, 1989.

Lazzaro J., Wawrzynek J., Mahowald M., Sivilotti M., Gillespie D., Silicon auditory processors as computer peripherals, *IEEE Trans on Neural Networks*, 4, 3, May 1993.

Licklider J.C.R., A Duplex theory of pitch perception, *Experentia*, 7, 128-133, 1951.

Liu W., Andreou A.G., Goldstein M.H., Analog cochlear model for multiresolution speech analysis, *Advances in Neural Information Processing Systems 5*, Morgan Kaufmann, 1993.

Marr D., Hildreth E. Theory of edge detection, *Proc. Royal Society of London B*, 207, 187-217, 1980.

Meddis R., Simulation of auditory-neural transduction: further studies, *J. Acoust Soc Am*, 83, 3, 1988.

Mirollo R.E., Strogatz S.H. Synchronization of pulse-coupled biological oscillators, *SIAM J. Appl Math*, 50, 6, 1990.

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

Morgan N., Bourlard H., Continuous speech recognition, *IEEE Signal Processing Magazine*, 12, 3, 25-42, 1995.

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

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

Senneff S., Zue V.W., *Transcription and alignment of the TIMIT database*, Second symposium on the advanced man-machine interface through spoken language, Oahu, Hawii, 1988.

Slaney M., An efficient implementation of the Patterson-Holdsworth auditory filter bank, Apple technical report No 35, Apple Computer Inc, 1993.

Smith L.S. Sound segmentation using onsets and offsets, *J of New Music Research*, 23, 1, 1994.

Smith L.S. *Onset/offset coding for interpretation and segmentation of sound*, UK patent no 9505956.4, March 1995.

Smith L.S., *A simplistic model of amplitude modulation detection*, (abstract), British Journal of Audiology, in press, 1995.

Wu Z.L., Schwartz J.L., Escudier P. A theoretical study of neural mechanisms specialized in the detection of articulatory-acoustic events, *Proc Eurospeech 89*, ed Tubach J.P., Mariani J.J., Paris, 1989.