Robust Sound Onset Detection using Leaky Integrate-and-fire Neurons with Depressing Synapses

Leslie S. Smith, *Member, IEEE,* and Dagmar S. Fraser

***Abstract*— A biologically inspired technique for detecting onsets in sound is presented. Outputs from a cochlea-like filter are spike coded, in a way similar to the auditory nerve (AN). These AN-like spikes are presented to a leaky integrate-and-fire neuron through a depressing synapse. Onsets are detected with essentially zero latency relative to these AN spikes. Onset detection results for a tone burst, musical sounds and the TIMIT corpus are presented.**

***Index Terms*— onset detection, depressing synapse, integrate-and-fire neuron**

I. Introduction

This work describes a biologically inspired technique for *onset detection*. Onsets occur at the start of certain perceptible changes in a sound. In [1] the term *onset detection* refers to the detection of discrete events in acoustic signals. Every sound source has an initial onset, and many have internal onsets (for example, animal vocalisations, such as human speech, or sequences of musical notes). Initial onsets are correlated with sudden increases in energy. Internal onsets also occur due to changes in spectral energy distribution, frequently including an increase in energy somewhere in the spectrum. Different sound sources have different onset characteristics. Onsets can exhibit differing magnitudes, spectral spreads and rates of energy increase. Magnitudes range from the just perceptible to the very large, with the pre-onset sound level having any possible level. Some onsets are wide-band, with sudden co-occurring increases in intensity in many parts of the audible spectrum: percussive sounds are a prime example. Others are narrowband, with the increase in energy being associated with some small area(s) of the audible spectrum, such as a note played on a flute. The flute note onset is a slow onset compared to say, the rapid onset in sound energy exhibited by a glass falling onto a stone floor. We approximate the perceptual onset by seeking certain characteristic types of increase in energy: see section III for details.

Mammalian auditory systems are strongly attuned to onsets from the earliest stage, with the auditory nerve responding more strongly to the start of a stimulus, and certain neurons in the cochlear nucleus spiking strongly at stimulus start [2], [3]. Therefore modelling aspects of the early auditory system (the cochlea, auditory nerve (AN) and cochlear nucleus) might offer engineering insight into early auditory processing. From an ecological perspective there are good reasons to believe that onsets provide a useful cue. The onset comes at the start of the sound (or at the beginning of some change in the

L. S. Smith and D.S. Fraser are with the Department of Computing Science and Mathematics, University of Stirling

sound), and is therefore useful for priming a response. Initial onsets are relatively undamaged by reverberation, since the first onset in the received signal will normally be from the direct path, and those onsets caused by reflections will generally be smaller. Indeed, these are normally ignored by animals when they estimate the location of a sound source. (This is known as the precedence effect, or law of the first wavefront [4].) Other cues such as offsets are severely smeared out in time in reverberant environments.

The aim of this work is to provide a signal that robustly indicates an onset. The signal is generated with low latency, during (rather than at the end of) the onset. We use this low latency to help group onsets in different parts of the spectrum. In earlier work, more critical use was made of the short latency time, when onsets were used to determine when to compute interaural time differences [5], [6]. The system detects onsets which may be either wide or narrow band, fast or (relatively) slow, large or (relatively) small, and starting from silence or some initial sound level. In this paper, we use the technique to detect musical note starts, and to detect certain phonemes in the TIMIT database [7].

The technique is described in section III. We present results in section IV.

II. Background

Onset detection systems have been used in music transcription [1], [8], where they are used for start-of-note identification. They have also been used for sound segmentation [9], lip synchronisation [10], monaural sound source separation [11], [12], and determining when to measure interaural time differences for sound direction finding [6] to avoid reflections. For off-line applications, the onset detection system can consider the sound after the onset as well as before, and the time taken for onset detection is unimportant. However, for on-line applications (e.g. real-time speech segmentation, real-time sound direction finding, or real-time music transcription), only the sound up to the time of onset is available, so that the latency of the detector becomes important.

In this, as in most onset detection work, we first bandpass the sound signal into many bands. This has two advantages: firstly, it allows onsets in some small part of the spectrum not to be overwhelmed by the overall signal strength, so long as they are not overwhelmed by the signal strength in the passbands that include their own frequencies. Secondly, it allows onsets found to be annotated with the bands in which they have been detected. This is important for music transcription and direction finding applications. The onset latency should be constant both for different signal strengths (unlike [9]) and independent of band. The bandpass filters themselves introduce a known fixed delay, and this needs to be taken into account when combining onsets from different bands. Given this, onsets in different bands may be grouped together, permitting onset detection in background noise.

Many different onset detection techniques have been used, in the context of segmenting either musical, or speech sounds. The simplest of these are based on signal energy, and are used in the context of segmenting hummed or sung notes [13], and are intended to improve the note differentiation capability of the early music transcription systems (such as [14]). Some more sophisticated techniques use simple first order difference based estimates, [8], [15], which take the maximum of the rising slope of the amplitude envelope as an index of onset. A more sophisticated variant of this is [1] which uses the relative difference, essentially calculating $\Delta\text{intensity}/\text{intensity}$. Another variant is [16] which uses

troughs in loudness to segment sung notes. A different approach uses filter based techniques: [10] uses a wavelet based filter and [9], [17], [18] use the difference between a long-term and a short-term average. Both Pont and Damper [19] and Smith [9], [17] use the coincidence detecting capabilities of leaky integrate-and-fire neurons. A related approach uses expectation based techniques [20] to detect sudden increases in intensity. Work by Hu and Wang [21] employs the peak of a firing rate derivative in tandem with coincidence across frequencies to indicate onset. The simplest techniques tend to find only the most prominent onsets, while techniques which rely on finding troughs can have a longer latency. Filter techniques can be optimised for particular source types and for particular reverberation characteristics, and can perform better, but require a convolution, and can have a long latency. The technique we describe uses depressing synapses in conjunction with leaky integrate-and-fire neurons, with the parameters set to detect increases in energy which correspond to onsets.

Signal coding is important in permitting the system to work with a wide dynamic range. Most sound systems use a coding system which is either linear (which provides more resolution at high signal levels than is required) or logarithmic. We have used a spike based coding similar to that of [22] which in turn has much in common with the coding used on the auditory nerve. It requires multiple spike trains per bandpassed channel, but preserves the precise signal timing (critical in interaural time difference calculation [5]), and provides the advantages of log compression while permitting analysis of low-level signals. Unlike the auditory nerve coding, the onset and steady-state responses are identical.

The physiological mechanism of the auditory system's onset response is not entirely clear. It starts at the auditory nerve, and this aspect appears to be related to the depletion of the neurotransmitter reserves at the synapse between the inner hair cell (in the Organ of Corti) and the spiral ganglion neuron (see [23]). The onset response is much stronger in the auditory brainstem cochlear nucleus, where there are a number of cell types (octopus, and some bushy and stellate cells) which respond specifically to onsets [2], [24]. How this onset response is mediated is not known: it may be due to their synaptic innervation (many AN fibres converge on these cells), or to the nature of the synapses themselves, or to the form of the leakiness of these cells (i.e. to the particular ion channels expressed in their membrane), or to their morphology, or to some combination of these. Nor is it known precisely how the outputs of these onset cells are used: they appear to innervate the medial superior olive and lateral superior olive [24], both implicated in sound direction finding, as well as other auditory brainstem areas.

A number of different models for depressing synapses have been put forward [25]–[27]. The primary effect of all of them is that the first few spikes to arrive at a depressing synapse have a much larger effect than those that follow soon after. This is a form of onset enhancement. Hewitt and Meddis [23] suggested a form of depressing synapse at the inner hair cell to spiral ganglion dendrite synapse. We are not aware of work suggesting depressing synapses in the cochlear nucleus, but depressing synapses are very common in mammalian neural systems.

III. The onset detection technique

The overall technique is illustrated in figure 1. Sound from a microphone (or a sound file) is bandpass filtered. Multiple trains of spikes are generated from each band, and these are applied to depressing synapses on leaky integrate-and-fire neurons. The spiking outputs of these

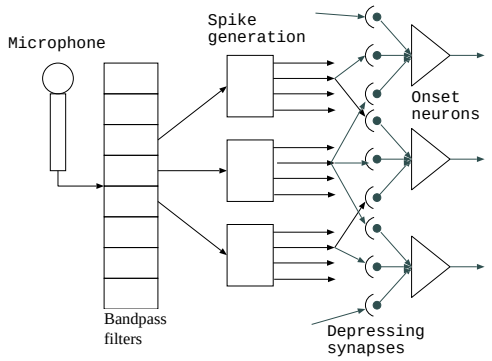

Fig. 1. Onset spike generation system. Note that spike generation is shown for only three bands, and that depressing synapses and onset generation is shown for a single level for three bands.

neurons signal onsets.

A. Bandpass filtering

Sounds (from a number of different sources and formats, but all sampled at a rate of at least 16Ksamples/second, 16 bits linear) were bandpass filtered using a Gammatone filterbank [28]. The Gammatone filterbank has a response similar to that of the basilar membrane in the cochlea: that is, the 6dB down point bandwidth is approximately 20% of the centre frequency. The filter density was chosen to ensure considerable overlap between adjacent filters. For differing signals, different bands were used, 15 in the case of a simple tone pulse, but normally 32 or greater, with center frequencies ranging from less than 250Hz to greater than 6kHz. An important issue in the design of the filters is delay: since the output of each filter is employed in conjunction with adjacent filters, ideally the insertion delay should be similar for all the filters. The Gammatone filter delay is proportional to the reciprocal of the bandwidth [29], and this delay has been corrected when combining onsets from different bands. Other filters, such as Butterworth have a more constant delay.

B. Spike generation

The representation used has some similarity with that of the mammalian auditory nerve. One advantage of the spike based representation we use is that it enables the system to work over a wide dynamic range through the use of multiple spike trains coding the output of each channel. Each spike codes a positive-going zero crossing. Each spike train S_i, for $i = 1 \ldots N$, (where N is the number of spike trains generated from a single bandpass channel) has a minimum mean voltage level E_i that the signal must have reached prior to crossing zero during the previous quarter cycle (where the cycle is assumed to be at the filter centre frequency). If there are N spike trains, these E_i are set by

$$E_i = D^i E_0 \tag{1}$$

for $i = 1 \ldots N$, for some E_0 fixed for all frequency bands. D was set either to 1.414 or 2, providing a 3dB or 6 dB difference between the energies required in each band. Note that if a spike is generated in band k, then a spike will also be generated in all the bands k' for $0 \leq k' \leq k$. This technique is similar to that used in [22], where Ghitza noted that it led to an improvement in automatic speech recognition in a noisy environment.

C. Onset generation

The auditory nerve-like representation described above does not enhance onsets, unlike the real mammalian auditory nerve. However, the manner in which it codes the signal can be used to build a neurally inspired onset detection system.

The spikes generated are passed through a depressing synapse, to a leaky integrate-and-fire neuron (onset neuron). The synapse model employed is a 3 reservoir model used in [23], [26] in the context of inner hair cell to auditory nerve fibres, and later in [25]

to model rat neocortex synapses. The model has three interconnected populations of neurotransmitter: M, the presynaptic neurotransmitter reservoir (available), C, the amount of neurotransmitter in the synaptic cleft (in use), and R, the amount of neurotransmitter in the process of reuptake (i.e. used, but not yet available again). These neurotransmitter reservoir levels are interconnected by first order differential equations as follows:

$$\frac{dM}{dt} = \beta R - gM \tag{2}$$

$$\frac{dC}{dt} = gM - \alpha C \tag{3}$$

$$\frac{dR}{dt} = \alpha C - \beta R \tag{4}$$

where α and β are rate constants, and g is positive during a spike, and zero otherwise. These are calculated each sample time (and α, β, and γ adjusted for the sample rate). We do not model loss or manufacture of neurotransmitter. We take the amount of post-synaptic depolarisation to be directly proportional to C.

For a strong enough signal, AN-like spikes will arrive at approximately F_c spikes per second, where F_c is the centre frequency of the bandpass channel. However, an evoked post-synaptic potential (EPSP) will only be generated for the first few spikes. The recovery time is set by the rate of transfer from the cleft to the reuptake reservoir, and from the reuptake reservoir to the pre-synaptic reservoir. If these are set low, then there will need to be a considerable gap in AN signals before a new onset is marked. Yet if they are set too high, the post onset EPSP (i.e. the EPSP produced by an indefinite train of AN spikes) will be relatively high, resulting in unwanted onset neuron firing. The model parameters (which are the rates of transfer between each reservoir) are set so that the first few spikes arriving result in near total depletion of the presynaptic reservoir. For simplicity, we set the maximal weight on each depressing synapse to the same level. We have set $g = 1100$: each spike lasts one timestep, so this high value reflects the total transfer of transmitter from M to C; $\alpha = 100$, resulting in rapid depletion of the cleft transmitter level, and $\beta = 9$, resulting in relatively slow reuptake of the used transmitter.

As can be seen from figure 1, each onset cell is innervated by a number of auditory nerve-like spike trains. These arrive from a number of adjacent bandpass channels, but all have the same sensitivity (i.e. value of i as described in section III-B). Thus the input to the onset neuron for sensitivity i in bandpass channel b is

$$I_{b,i}(t) = \sum_{j=b-m}^{j=b+m} wC_{j,i}(t) \tag{5}$$

where w is the weight associated with each synapse (here these are all the same), and $C_{j,i}(t)$ is the neurotransmitter in the cleft associated with the synapse from AN-like fiber from bandpass channel j at sensitivity level i. The value of m defines the size of the neighborhood innervating the onset neuron, and has to be adjusted if the bandpass channel spacing is altered. Each single post-synaptic potential is insufficient to make the onset neuron fire, ensuring that spikes on more than one auditory nerve-like input are required. The neurons used are leaky, so that these spikes need to be nearly co-incident in time.

We model the onset neurons using leaky integrate-and-fire neurons. Leaky integrate-and-fire neurons are the simplest model neurons which maintain any semblance of the temporal behaviour of real neurons. They are a single compartment model whose below threshold behaviour is described by a first order differential equation

$$\frac{dV}{dt} = -V/\tau + I(t) \tag{6}$$

where V is the voltage-like state variable of the neurons, τ is the membrane time constant, and $I(t)$ is the external driving input. When V reaches the threshold θ (set to

1 here) from below, the neuron fires and V is reset (to 0 here). There is a refractory period following this during which the neuron is unable to fire. Many different versions of the leaky integrate-and-fire neurons have been described: see [30], chapter 14 for a useful review.

This approach tends to reduce the effects of noise (which might result in occasional but uncorrelated firing in auditory nerve-like inputs in adjacent channels). Onset neuron firing is always the result of the most recent post-synaptic potential: post synaptic potentials are much larger before the synapse depresses, so that onset spikes generally occur as a result of (and at the same time as) AN-like spikes coding increases in energy in the bandpassed signal. In this sense, there is zero latency between the bandpassed signal onset and the onset spike.

The degree of leakiness of the onset neurons determines the degree of coincidence of incoming excitatory spikes required to cause firing. Some experimentation has been done with this leakiness: because the spike rate on the AN fibres is proportional to F_c, the neurons with lower F_c receive fewer incoming spikes, and hence have lower leakage than those with higher F_c's. However, we have found that making the leakage directly proportional to F_c does not work well: at low frequencies, the leakiness is too low, causing firing to occur too often, and at high frequencies, the leakage is too high, resulting in onsets being missed. As a result, we set the leakage proportional to frequency (leakage $= 1/\tau = 0.15 * f_c$) for only a part of the frequency range, usually between 500Hz and 1000Hz, making the leakage constant below 500Hz, and above 1000Hz.

IV. Results

We first consider a 440 Hz tone burst in a noisy background. Figure 2 shows the stimulus waveform, the coding produced from this, and the onsets generated. The bandpass filter had 15 bands, from 250 to 750 Hz, and 15 different levels of sensitivity were used in the AN-like spike coding. The onset cells were innervated by a maximum of 11 bands (so that those near the middle were more strongly stimulated). It is clear that the AN-like representation captures both the white noise and the tone burst, and that the onset cells spike over all the bands at the start of the white noise, and in a small number of localised low sensitivity bands at the start of the tone burst. The high sensitivity onset detectors do not fire again because the noise has kept the synapses to these onset cells depressed. Although the envelope of the white noise varies widely, there are no stray onsets detected. This reflects the sensitivity of the onset detector to envelope modulations that are perceived as onsets. Testing the system with a regular pulse train, each pulse is detected as a single onset when the pulses are spaced apart by about 60ms or more: below this separation, the system treats the pulse train as a single entity, with an onset only at the start. Around 60ms, the first few pulses are detected as onsets, but not subsequent ones.

Next we show that onsets are generated with low latency (excluding the filter delay), and that the latency is largely independent of the signal strength. Figure 3 shows the onset of a 6kHz signal of maximal intensity, and the onset times found for varying signal strengths. The actual onset time is 0.0148 seconds. At high signal level, the onset is found at 0.0158 seconds, a delay of 1ms. The filterbank delay is $D = (n-1)/2\pi b$ [29] where n is the filter-bank order (here 4), and b is the bandwidth. At 6kHz, the equivalent rectangular bandwidth is $0.108 \times 6000 + 24.7 = 672.7$Hz [31], so that $D = 0.71$ms. The additional delay may be due to the fact that AN-like spikes are emitted at positive-going zero crossings where the pre-crossing level exceeds some value. From figure 3 the first of these occurs 1 cycle after

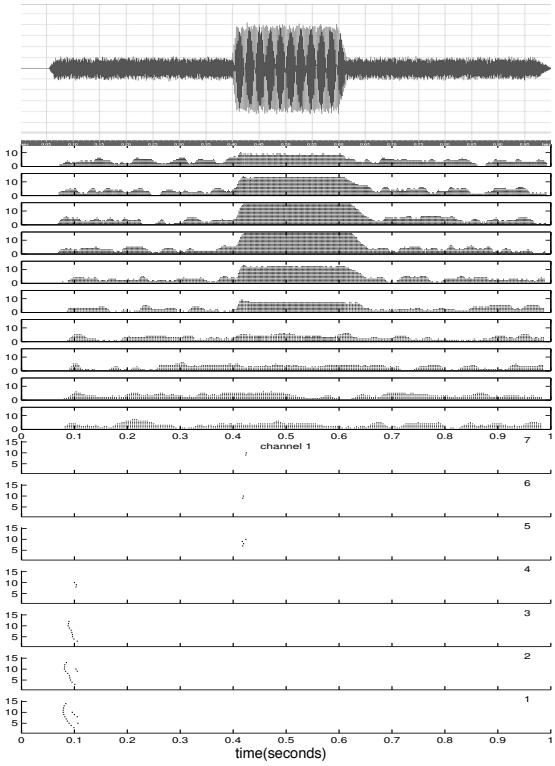

Fig. 2. Tone burst at 440Hz in white noise (Tone is 12dB louder). Top shows the sound waveform, middle shows the AN-like spike output, and bottom shows the onset cell firings. In the middle image, each subgraph shows the firings for one frequency band of the AN-like cells. The subgraph for the lowest frequency band is at the bottom. Within each subgraph, the dark area is made up from a number of horizontal spike trains. Spikes from the highest sensitivity channel are at the bottom, and spikes from the lowest sensitivity channel are at the top. In the bottom image, each subgraph shows the onsets found at a single sensitivity level, for all frequency bands. The subgraph for the highest sensitivity is at the bottom. The top subgraph shows the least sensitive level at which onsets were found. Inside each subgraph, the lowest frequency channel is at the bottom, and the highest at the top.

the stimulus onset, at 0.015 seconds. This suggests that the 1ms delay is made up of the 0.71ms filter-bank delay, the 0.17ms cycle time of the signal, plus another 0.12ms due to the actual latency of the onset detecting system. We note that it is not until the signal is 24dB attenuated that the onset time changes. This delay, and the delay in the onset up to an attenuation of 36dB are due to the AN-like spike not being generated until the signal is stronger: that is, these delays are due to the onset not being detected until the initial signal has become stronger. At lower signal levels, the off-centre frequency response of adjacent filters is small, so that it takes a number of AN-like spikes before the post-synaptic activity at the onset cell exceeds threshold. Below 54dB attenuation, no AN-like spikes are generated.

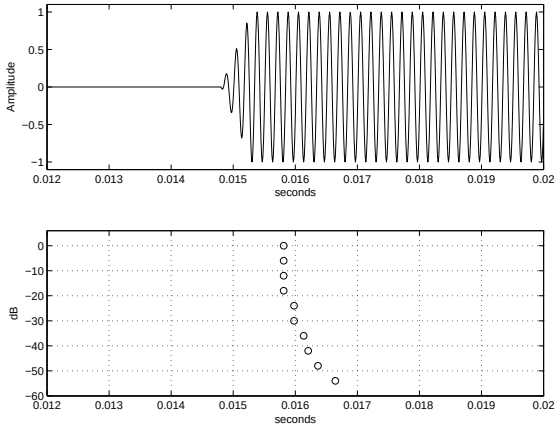

Fig. 3. Onsets caused by a 6kHz stimulus with 3 cycle (0.5ms) rise time. Top shows the stimulus. Bottom shows the onset time found for attenuation varying in 6dB steps.

Next we consider some musical sounds. We have experimented with two single note instruments: a saxophone sound and a flute sound: the results are similar. We report the saxophone sound here. Two styles of playing were considered: tongued and slurred. In the tongued style, each note has a distinct start: the musician uses their tongue to interrupt the flow of air before the start of each note. In the slurred style, there is no break between notes. For the tongued saxophone sound (figure 4), one can see the brief breaks between the notes in the spectral view (top part). In the AN-like representation (middle part), one can see the individual notes, and the (brief) breaks between them. Note that the energy between notes does not go down to zero, and, indeed, does not fall

much in the lower frequency bands at all. As a result, apart from the very first onset which is detected in all the bands, further onsets are detected in the higher frequency bands. All the onsets are detected (including the last one which is relatively quiet).

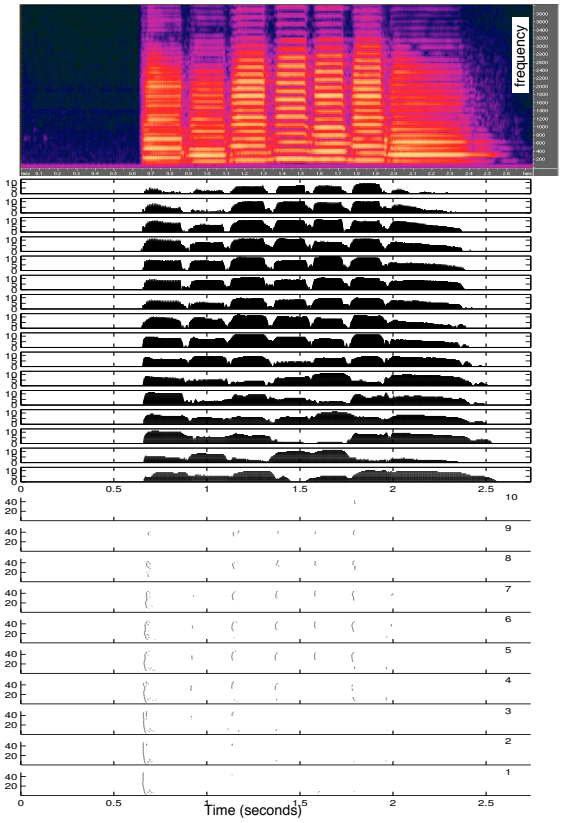

Fig. 4. Tongued saxophone sound. Top shows the spectral view between 0 and 3kHz (dark areas are low energy, light areas high energy). Middle shows the AN-like representation, for every third band of 48 bands with centre frequencies between 300 and 3000Hz. Bottom shows the onsets found with 15 bands 3dB apart. Middle and bottom organization are as in figure 2.

With the slurred style, the notes run into each other, and this is visible in the AN-like representation (see figure 5, top). As a result not all note onsets are found: the first one is found easily, but the 2nd, 5th and 8th are missed. Considering the extent to which the harmonics overlap, and considering that the bandpass channels we are using are wideband, this is not surprising. The starts of the notes missed result from relatively small changes in the energy distribution, without any preceding dip in overall local energy. The technique used requires that detectable onsets have a distinct increase in energy in a number of adjacent bands. The change in energy between adjacent slurred notes is not always sufficient to trigger onset detection. If we use narrower bands, for example, using a bandwidth one half of the original bandwidth, and use 18 bands per octave, all the note onsets are found. Alternatively, a pitch movement detector can be used. Such detectors were initially used in music transcription without onset detection [14], and are generally mixed with onset detection in more modern systems [13], [16].

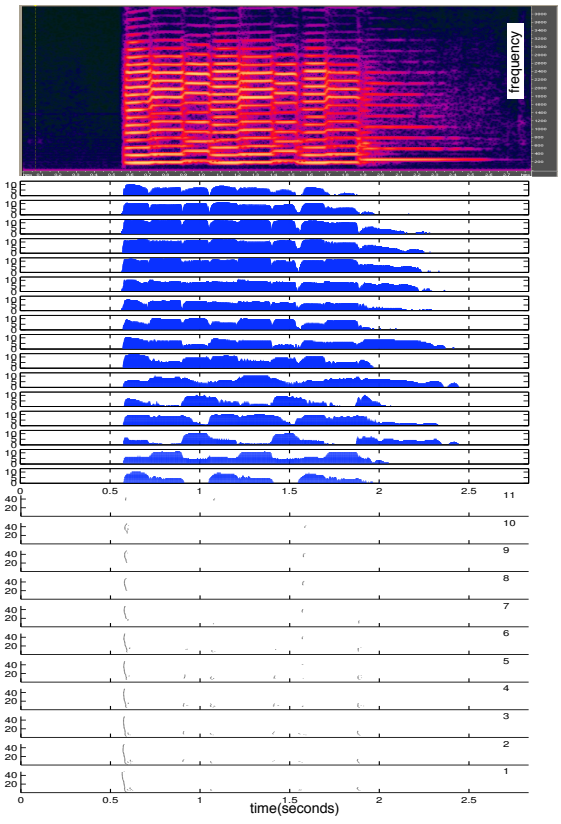

Fig. 5. Slurred saxophone sound. Top shows the AN-like representation, for every third band of 48 bands with centre frequencies between 300 and 3000Hz. Bottom shows the onsets found with 15 bands 3dB apart. Middle and bottom organization are as in figure 2.

The same analysis was employed with a multi-note instrument, namely a Spanish guitar. Unlike the saxophone, a guitar can play up to six notes simultaneously. In the fragment of music analysed, only one note is started at a time, but new notes overlap existing notes (as is generally the case with guitar music). In this case, it is difficult to see from the AN-like representation where notes start or end. The note information is shown in table I.

Sound number	1	2	3	4	5
Sound onset time	0.53	0.6	0.71	0.83	0.92
Sound type	m	f	s	f	m
Found (full band)	Y	N	N	Y	Y
Found (low freq.)	Y	N	Y	N	Y
Sound number	6	7	8	9	10
Sound onset time	1.08	1.26	1.32	1.43	1.53
Sound type	s	m	f	s	s
Found (full band)	Y	N	Y	Y	Y
Found (low freq.)	Y	Y	Y	N	Y
Sound number	11	12	13	14	15
Sound onset time	1.6	1.76	1.86	1.93	2.07
Sound type	m	s	f	m	s
Found (full band)	N	Y	Y	Y	N
Found (low freq.)	Y	N	Y	Y	Y
Sound number	16	17			
Sound onset time	2.18	2.27			
Sound type	g	m			
Found (full band)	Y	Y			
Found (low freq.)	N	Y			

TABLE I

TABLE OF NOTE TIMES AND NUMBERS FOR GUITAR SOUND. KEY: m IS MAIN NOTE, s SUBSIDIARY NOTE, g GRACE NOTE, AND f FINGER NOISE

The guitar sound is complex. The sounds themselves have been grouped into main notes (the strongest notes in the sample), subsidiary notes (less strong notes), grace notes (one very short note), and finger noises (noises made by the guitarist but not resulting in musical notes). The AN-like spike trains show quite rapid onsets and slow offsets, in the lower frequency bands, characteristic of the vibration of an undamped plucked guitar string. From the AN-like spike trains, (figure 6, top), it is clear that the higher harmonics do not last as long. Note 1 is visible in all channels. However, subsequent notes are not. Note 15, for example is only visible in some lower frequency channels, and note 16 is most visible in the higher frequency channels. In listening, the most salient notes are 1, 5, 7, 11, 14, and 17. Two different analyses have been applied. In the first of these, 60 bands between 75 and 3kHz were used, and the onsets are shown in figure 6. In this analysis, onsets are found for most of the notes: notes 3, 7, 11, and 15 are missed. Most of the finger noises are also found. In the second analyses, 24 bands between 75 and 300 Hz were used, with the bandwidth set to one quarter of the original width. In this analysis, nearly all the notes were found, the only exception being notes 9 and 12, and a very brief grace note, note 16. Fewer finger noises are found. Because of the delay of the gammatone filter, the onsets are found rather later with the second technique than with the first one, and this delay has been taken into account. Table II shows the results in terms of hits etc. It is important to note that although the main notes do have clear onsets,

TABLE II

ANALYSIS OF GUITAR NOTE ONSETS. THERE ARE 17 SOUND ONSETS, 4 OF WHICH ARE FINGER NOISE. LF IS LOW FREQUENCY.

	All sounds			All except finger noise		
Filter	Full	LF only	Both	Full	LF only	Both
Hits	12	12	16	9	10	13
Deletions	5	5	1	4	3	0
Insertions	1	2	3	4	4	6

the subsidiary notes start from a high background sound level, hence finding these onsets is non-trivial.

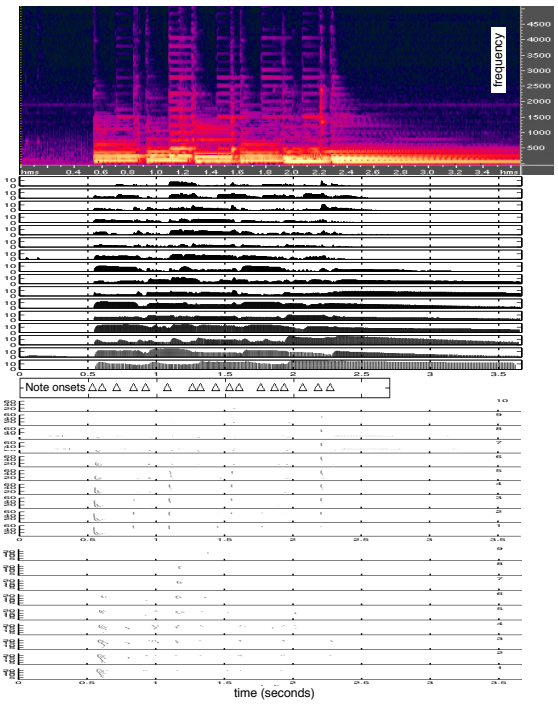

Fig. 6. Spanish guitar sound. Top shows the spectral view. Next shows the AN-like representation, for every third band of 60 bands with centre frequencies between 75 and 3000Hz. The onsets of notes are marked by the triangles below this. Second bottom shows the onsets found with 16 levels 3dB apart. Bottom shows onsets found for the frequency range from 75Hz to 300Hz (24 bands). AN-like spikes and onset spikes organization are as in figure 2. See text for details.

Turning to speech, we present results for speech from the TIMIT dataset [7]. We have applied our technique to the entire training segment of the TIMIT corpus. AN-like output and onsets are shown for example male and female utterances in figures 7 and 8. The wideband nature of the speech is very visible in the AN-like output, as are the areas of near silence inside the continuous speech. The onsets are also spread across the bands. In addition, onsets tend to start in the bands of highest sensitivity, and then to occur slightly later in the lower sensitivity bands. This is because some onsets have a slow attack time.

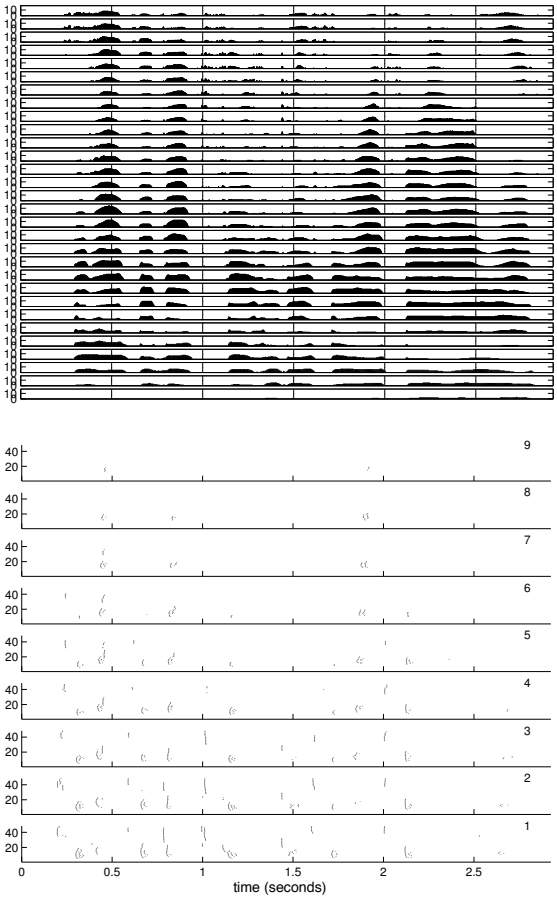

Fig. 7. Female TIMIT utterance. Top shows the AN-like representation, for every band of 48 bands with centre frequencies between 100 and 7500Hz. Bottom shows the onsets found with 9 levels 3dB apart. Organization is as in middle and bottom part of figure 2.

We have correlated the onsets found with the phonetic transcription supplied with the TIMIT dataset. To achieve this, we first had to find the actual times of onsets from the onset spikes. We took the onsets at each sensitivity level, and clustered them by demanding a gap of at least 0.01 seconds between groups. Using a gap overcomes possible problems caused by the different delays in the filterbank. As can be seen from the bottom halves of figures 7 and 8, the onset spikes are already quite tightly grouped. This is partly a result of the filter

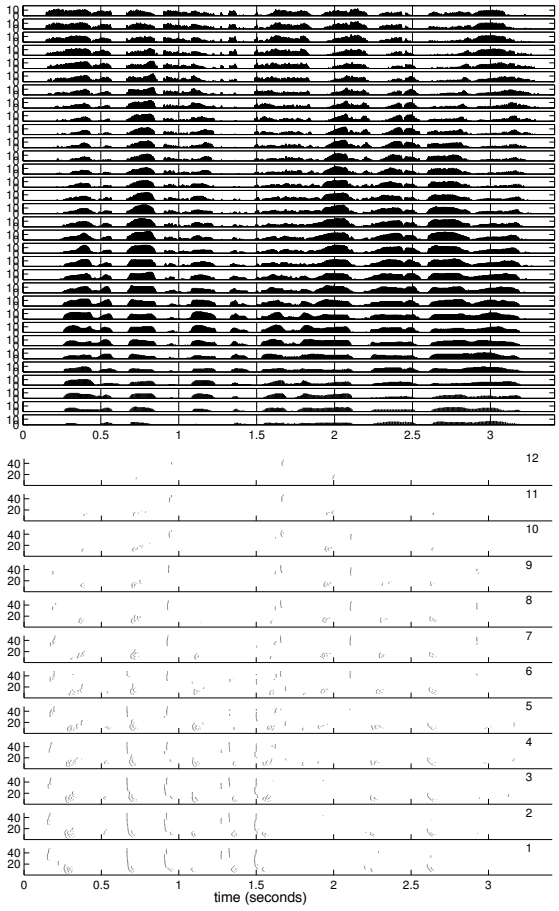

Fig. 8. Male TIMIT utterance. Top shows the AN-like representation, for every band of 48 bands with centre frequencies between 100 and 7500Hz. Bottom shows the onsets found with 12 levels 3dB apart. Organization is as in middle and bottom part of figure 2.

delay changing little from band to band, partly of the low latency of the onset detection (excluding filter delay), and partly a result of the latency of the onset detection being near independent of the actual signal level inside the filtered band. This tight grouping allows the simple grouping mechanism above to work. Grouping results in a sequence of onset intervals for each sensitivity level. We then integrated the results from the different sensitivity levels by replacing any overlapping intervals by a single interval. This resulted in a new overall sequence of intervals, each representing the existence of an onset. This onset coincidence across frequencies is akin to that used in [21]. We do not examine the structure of the onset pulses for classification: utilising this structure for onset labelling is part of ongoing research. The results are shown for males in table III and females in table IV.

Phoneme type	Phoneme events	Onsets detected	Sensitivity
affricative	1426	1361	0.95
closure	17020	2757	0.16
fricative	15135	11745	0.78
nasal	9978	2678	0.27
semivowel	14265	7839	0.55
stop	17866	13577	0.76
vowel	40468	29536	0.73

TABLE III

PHONEME TYPES IN THE 3260 MALE TIMIT UTTERANCES (99138 PHONEMES) PROCESSED, AND THOSE DETECTED (WITHIN 28MS OF ANNOTATED PHONEME START). SENSITIVITY IS DEFINED TO BE RATIO OF TRUE POSITIVES/POTENTIAL TRUE POSITIVES.

.

There is a clear correlation between the types of phoneme and the onsets found, and almost no variation between male and female. Phoneme onsets may be missed because the onset of this phoneme and the previous one overlap, or because that phoneme does not start with, or contain an onset. Many of the vowels, semivowels and nasals that are missed follow other voiced sounds, but sharper filtering (as suggested recently [32]) may allow these to be recovered. However 87% of the starts of sequences of voiced sounds (vowel, nasal and semivowel) are found. The fricatives missed are either just missed by a few milliseconds, or occur just beside a stop. Non-existent onsets may be found because a true onset is broken into multiple onsets. Envelope variations inside a phoneme are sometimes misidentified

Phoneme type	Phoneme events	Onsets detected	Sensitivity
affricative	640	611	0.95
closure	7153	1127	0.16
fricative	6334	4919	0.78
nasal	4144	1111	0.27
semivowel	5914	3284	0.56
stop	7511	5735	0.76
vowel	16911	12350	0.73

TABLE IV

PHONEME TYPES IN THE 1360 FEMALE TIMIT UTTERANCES (41454 PHONEMES) PROCESSED, AND THOSE DETECTED (WITHIN 28MS OF ANNOTATED PHONEME START). SENSITIVITY IS AS DEFINED IN TABLE III.

as onsets. This happens most frequently for vowels and results at least partly from the onset detector being confused by slow envelope modulation inside single vowels. The bulk of false positives, 83%, occur within vowels, with 12% inside sibilances. The remaining 5% occur in stops or at the beginning of the recording (due to extraneous recorded noise). Turning to stops, two particular stops, 'dx' and 'q' account for 75% of the missed stops: we believe that these stops are largely not associated with an increase in energy. If we consider the stop consonants ('b', 'd', 'g', 'p', 't', and 'k') as in [21] the sensitivity of the system is 0.97, compared to their result of 0.93 at 30dB SNR. The overall selectivity (the ratio of useful to total detections) is defined as (true positives)/(true positives + false positives). Here it has the value 0.75.

V. DISCUSSION AND FURTHER WORK

A neurally inspired technique for the robust detection of onsets in sound has been presented and tested with a simple tone burst, some musical instrument sounds and the TIMIT corpus. The spiking AN-like representation provides an effective early representation over a wide dynamic range, enabling onset detection. Because of the spiking nature of the system, the latency is essentially that of the filterbank: further, the onset pulses are essentially phase locked (see [6]). Importantly, the onsets detected fit with the informal definition of an onset as a rapid increase in the energy present in some part of the auditory spectrum.

The system has some similarity to its biological analogue, and exhibits some of the qualities of that system. Although neurally inspired in the use of cochlear filters, AN-like spike representation of the output of these filters, depressing synapses, and leaky integrate-and-fire neurons, it is certainly not a model of the neural system. The gammatone filters are modelled on the auditory filters, but their precise characteristic is not believed to be crucial to the system operation (although sharper filters do help with musical sounds). Indeed, recent work suggests that the gammatone filterbank is an oversimplified model, and that frequency selectivity is level dependent [32]. Although the spike generation system has some commonality with the auditory nerve, it is deterministic, and spikes are generated in phase on every cycle. This is unlike the auditory nerve where the spikes are stochastic, and have a maximum rate much lower than the highest F_c's we are working with.

In earlier work we used a different onset detector [9], [17]. This used a difference of averages technique, resulting in a 15-25ms latency. In addition, this latency was level dependent. The advantage of the system reported here is that it has essentially zero non-deterministic latency, whatever the signal level. That is, onsets are detected as soon as they occur in the bandpassed signal. This allows onsets to be used directly (that is, in near real time). One application of this is determining when to measure time delays and intensity differences between

sensors, important for determining the bearing of sound sources in reverberant environments [5], [20], [33].

The cells in the cochlear nucleus which we are modelling are sometimes called ”onset cells”. They get this name from their spiking behaviour when a simple stimulus (such as a tone pip) is used to stimulate the animal hearing. However, it is not clear what function they might have in the context of more complex (and ecologically realistic) sounds.

In the introduction, we characterised onsets informally. When we are dealing with musical notes, it is clear what is meant by the onset of a note. But what is meant by an onset in the context of continuous speech is much less clear. Obviously, the start of an utterance is an onset, as clearly is speech after a pause or epenthetic silence. But the amount of energy in continuous speech varies rapidly, and is distributed unevenly over the spectrum. This envelope variation occurs at a number of time-scales, with a period ranging from 40ms to 200ms (characteristic of phoneme production) down to more rapid modulations, with a period of less than 10ms (characteristic of amplitude modulation). We have implemented an onset detector which is inspired by the biological system, and whose time constants have been set to detect what we informally think of as an onset. By applying more complex sound signals to this system we have started to try to answer what onset cells might be useful for in more complex soundscapes. Onsets are still found in background sound, as can be seen from the guitar example (figure 6). We have analysed the annotated TIMIT corpus: fricatives, affricatives and selected stop consonants are almost all detected, as are the starts of most voiced sequences. Investigation continues into a more intelligent way of grouping relatively slow onsets. The next challenge is to correlate the pattern of onset cell firing with the individual phonemes and phoneme types that cause them which would lead to a reduction in false positives. In [21] such a system is detailed for the separation of stop consonants which shares much in common with the generation of AN-like spikes presented here.

The current system is non-adaptive. This has the advantage of simplicity, but the disadvantage that it implies human tuning of all the parameters. In fact, the parameters are not critical: however, if the system were adaptive, the tuning would be simpler. In particular, there is an interaction between the precise characteristics of the flow rates between the reservoirs in the depressing synapses, the degree of interconnection (or spread) between the spike generation stage and the onset neurons, and the weight at the depressing synapses. Incorrect parameter settings show themselves by either the onset neurons firing insufficiently, or too often. Providing adaptation to maintain a degree of homeostasis would make the system easier to use. It is also highly biologically plausible.

The implementation used here has been entirely software based. The simplicity of the processing makes hardware implementation a practical possibility. Bandpass filtering is a straightforward technique in many forms of implementation. The deterministic spike generation technique here is simple to implement in VLSI, whether digital or analogue. Depressing synapses may present more of a problem. A different type of depressing synapse (essentially a two reservoir model) has recently been implemented in sub-threshold analogue VLSI [34]: the three reservoir model used here does matter because it allows us to adjust both the degree of depression and the rate of recovery independently. Several models of integrate-and-fire neurons have been implemented, both in analogue VLSI [35], and in FPGA digital technology [36]. Work is already under way (in conjunction with Oxford University) on hardware implementation of the

system.

ACKNOWLEDGMENT

This work was partially funded under the EPSRC contract number GR/R74574. I thank the anonymous referees for useful suggestions.

REFERENCES

[1] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *International conference on acoustics, speech and signal processing*, 1999, pp. 3089–3092.

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

[3] I. Winter, A. Palmer, L. Wiegrebe, and R. Patterson, "Temporal coding of the pitch of complex sounds by presumed multipolar cells in the ventral cochlear nucleus," *Speech Communication*, vol. 41, pp. 135–149, 2003.

[4] J. Blauert, *Spatial Hearing*, revised ed. MIT Press, 1996.

[5] L. Smith, "Using depressing synapses for phase locked auditory onset detection," in *Artificial Neural Networks: ICANN 2001*, ser. LNCS, G. Dorffner, H. Bischof, and K. Hornik, Eds., vol. 2130. Springer, 2001, pp. 1103–1108.

[6] ——, "Phase-locked onset detectors for monaural sound grouping and binaural direction finding," *Journal of the Acoustical Society of America*, vol. 111, no. 5, p. 2467, 2002.

[7] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," NIST/LDC, 1993.

[8] J. Bilmes, "Timing is of the essence," Master's thesis, Massachussetts Institute of Technology, 1993.

[9] L. Smith, "Onset-based sound segmentation," in *Advances in Neural Information Processing Systems 8*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. MIT Press, 1996, pp. 729–735.

[10] C. Tait, "Wavelet analysis for onset detection," Ph.D. dissertation, Department of Computing Science, University of Glasgow, 1997.

[11] A. Bregman, *Auditory scene analysis*. MIT Press, 1990.

[12] G. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[13] R. McNab, L. Smith, D. Bainbridge, and I. Witten. (1997, May) The New Zealand Digital Library MELody inDEX, http://www.dlib.org/dlib/may97/meldex/05witten.html. Department of Computer Science, University of Waikato. [Online]. Available: http://www.dlib.org/dlib/may97/meldex/05witten.html

[14] J. Moorer, "On the transcription of musical sound by computer," *Computer Music Journal*, vol. 1, pp. 32–38, 1977.

[15] M. Goto and M. Muraoka, "A real time beat tracking systems for audio signals," in *Proceedings of the 1995 international computer music conference*, 1995, pp. 171–174.

[16] L. Clarisse, J. Martens, M. Lesaffre, B. Baets, H. Meyer, and M. Leman, "An auditory model based transcriber of singing sequences," in *Proceedings of ISMIR*, 2002, pp. 171–174.

[17] L. Smith, "Sound segmentation using onsets and offsets," *Journal of New Music research*, vol. 23, no. 1, pp. 11–23, 1994.

[18] M. Marolt, A. Kavcic, and M. Privosnik, "Neural networks for note onset detection in piano music," in *Proceedings of ICMC 2002*, 2002.

[19] M. Pont and R. Damper, "A computational model of afferent neural activity from the cochlea to the dorsal acoustic stria," *Journal of the Acoustical Society of America*, vol. 89, pp. 1213–1228, 1991.

[20] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Oshnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robotics and Autonomous Systems*, pp. 199–209, 1999.

[21] G. Hu and D. Wang, "Separation of stop consonants," in *Proceedings IEEE Intl. Conf. on Acoust., Speech and Signal Proc.*, 2003, pp. 749–752.

[22] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and Language*, vol. 1, pp. 109–130, 1986.

[23] M. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," *Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 904–917, 1991.

[24] E. Rouiller, "Functional organization of the auditory pathways," in *The Central Auditory System*, G. Ehret and R. Romand, Eds. Oxford, 1997.

[25] M. Tsodyks and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," *Proc. Nat Acad Sciences*, vol. 94, pp. 719–723, 1997.

[26] M. Giugliano, M. Bove, and M. Grattarola, "Fast calculation of short-term depressing synaptic conductances," *Neural Computation*, vol. 11, pp. 1413–1426, 1999.

[27] R. Bertram, "Differential filtering of two presynaptic depression mechanisms," *Neural Computation*, vol. 13, pp. 69–85, 2000.

[28] R. Patterson, M. Allerhand, and C. Giguere, "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," *Journal of the Acoustical Society of America*, vol. 98, pp. 1890–1894, 1995.

[29] M. Cooke, *Modelling Auditory Processing and Organisation*, ser.

Distinguished Dissertations in Computer Science. Cambridge University Press, 1993.

[30] C. Koch, *Biophysics of Computation*. Oxford, 1999.

[31] B. Moore and B. Glasberg, "A revision of Zwicker's loudness model," *ACTA Acustica*, vol. 82, pp. 335–345, 1996.

[32] C. Shera, J. Jr., and A. Oxenham, "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 842–846, 2002.

[33] J. Huang, N. Oshnishi, and N. Sugie, "Sound localization in reverberant environment based on a model of the precedence effect," *IEEE Trans. Instrum. Meas.*, vol. 46, pp. 842–846, 1997.

[34] C. Rasche and R. Hahnloser, "Silicon synaptic depression," *Biological Cybernetics*, vol. 84, pp. 57–62, 2001.

[35] M. Glover, A. Hamilton, and L. Smith, "Analogue VLSI leaky integrated-and-fire neurons and their use in a sound analysis system," *Analog Integrated Circuits and Signal Processing*, vol. 30, no. 2, pp. 91–100, 2002.

[36] S. Lim, A. Temple, S. Jones, and R. Meddis, "Digital hardware implementation of a neuromorphic pitch extraction system," in *Neuromorphic Systems: Engineering Silicon from Neurobiology*, L. Smith and A. Hamilton, Eds. World Scientific, 1998.

PLACE PHOTO HERE

Leslie Smith Biography text here.

PLACE PHOTO HERE

Dagmar Fraser Biography text here.