Audio Engineering Society

Convention Paper 5767

Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

A new approach to detecting auditory onsets within a binaural stream

Ben Supper[1], Tim Brookes[1], and Francis Rumsey[1]

[1] *Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK.*

Correspondence should be addressed to Ben Supper (b.supper@surrey.ac.uk).

ABSTRACT

The human auditory system is particularly sensitive to spatial information conveyed in the first two milliseconds of an auditory event. Therefore, in order to analyse a stream of binaural data in a perceptually relevant way, it is important to determine quickly and precisely the onset of each event within a data stream. This paper details the design of an auditory onset detector which is intended to assist in the extraction of spatial parameters from an arbitrary binaural stream. A fast predictive filter architecture is specified, and tested using binaural recordings. These items highlight the strengths, limitations, and difficulties of computerised onset detection, and of this approach in particular.

***Note:** Throughout this paper, the term 'binaural data' and its variants are used to refer to continuous data in a binaural format. No implicit distinction is intended between sound received by a listener experiencing a sound field, and the data presented to a computer as a two-channel audio file.*

1. INTRODUCTION

A listener will make sense of a continuous stream of auditory data by experiencing it as a synthesis of spatially and contextually separated events, which begin and end at certain points in time [1]. If spatial information is to be extracted from an auditory data stream, then onsets are of key importance. Many current measurement standards for spatial impression are based on the precept that the early portion of a sound conveys the most salient information about source direction and source-related spatial attributes. Any process which focuses on the spatial properties of auditory events, and fails to treat onsets and similar sudden level fluctuations specially, will furnish little insight about all but the most contrived stimuli. This is exemplified in the paper by Mason and Rumsey [2], in which the spatial characteristics of different musical instruments are found to differ within the same simulated environment.

A number of approaches to the sound segmentation exercise have been taken in recent decades, each designed to satisfy different criteria of efficiency and effectiveness. The proliferation of strategies aimed at this task highlights the fact that it is not trivial. Presence of interfering direct sounds, early reflections from room boundaries, and later diffuse reverberant energy, means that a simple analysis of signal energy fluctuating over time is easily deceived by many recorded signals. Weaker onsets will be masked incorrectly by louder sounds, and strong, isolated room reflections will frequently look like separate onsets.

Practical segmentation procedures are often tailored to a certain type of source material — most often to speech. These models must overcome the confusion caused by powerful early reflections when localising sounds. Generally, they do this by incorporating a simplified model of the precedence effect.

The precedence effect was first explored in classic papers by Wallach et al. [3] and Haas [4], but its existence as a psychoacoustic phenomenon was acknowledged far earlier [5]. It is an inhibitory process within the brain which allows listeners to locate sounds in highly-reverberant environments, without becoming confused by multiple early reflections. The human auditory system favours information from the earliest portion of an auditory onset — purely direct sound — over later arriving sound. This effect governs the spatial perception of sound similarly to the way in which psychoacoustic masking determines its audibility, but it can affect far more prominent signals than masking.

The precedence effect will frequently cause the later of two sounds to be localised and fused with the earlier sound. If the two sounds are exceptionally close together in time, the perceived image will occur at a point spatially between the two. The earlier sound is observed to dominate about one millisecond after onset of a sound, and its influence agreed to cease after 50–80ms, depending on the nature of the subject material.

Huang's model [6] and Martin's model [7] are two examples which employ different means of coping with the precedence effect. Huang's implementation is designed to assist the algorithmic separation of two spatially distinct sounds. No attempt is made to emulate the interpretation of these sounds by the human auditory system, and it would not be feasible to modify this model to do so.

The model of the precedence effect that Martin implements is similar to the one described herein. However, Martin's onset detector works by taking an input signal divided into critical bands, examining the amplitude envelope of each one, and flagging significant maxima as onsets. The detector is allowed to look ahead to ensure that it does not mark small local peaks which precede larger ones. Models which operate in this way tend to generate a very high rate of detected onsets: far too high for any spatial analysis task other than source localisation to be executed effectively.

This paper details a recently developed onset detection model. The model is intended to divide a stream of data simply into separate auditory events, so that some of their spatial attributes can be extracted and investigated. These attributes cannot be studied without the sense of context an onset detector provides, because they are sensitive to the amplitude envelope of the source material, and most particularly to the beginning of each auditory event.

1.1 Cross-correlation models for spatial analysis

It must be noted that detection models with basic onset detectors already exist for analysing the spatial properties of arbitrary binaural stimuli. The most popular starting point for spatial analysis is the running cross-correlation technique first proposed by Jeffress [8]. An international standard now exists for calculating the interaural cross-correlation function, or IACF, of a signal [9]. Calculation of the running IACF over a series of different time delays is achieved using the following formula:

$$IACF_{T_1T_2}(\tau) = \frac{\int_{T_1}^{T_2} p_l(t)p_r(t+\tau)dt}{\sqrt{\int_{T_1}^{T_2} p_l^2 dt \int_{T_1}^{T_2} p_r^2 dt}}$$

where T_1 and T_2 are beginning and end times of the measurement, p_l and p_r are sampled left and right binaural impulse responses, and τ is the interaural time difference. Calculations are made for $-1\text{ms} < \tau \leq 1\text{ms}$.

Although it is not intended to be applied to signals other than impulse responses, the IACF can be modified to analyse arbitrary signals. As it appears above, it is not very useful for this task because the denominator expression normalises the output to its input level. This means that gain applied to the input will not change the IACF. If a series of short IACF measurements is taken from a source recording and each one is normalised separately, the IACF which emerges will be equally sensitive to loud and quiet portions of the audio signal. One will not be able to distinguish between them from the data.

Jeffress's model was improved subsequently, most notably by Lindemann [10]. One of his improvements was a simulation of part of the precedence effect. However, the output which emerges from Lindemann's model still contains no data relating to changes in signal energy. Although Lindemann included mechanisms which account for auditory onsets, these are dealt with internally, employing localised filters and attenuators during the cross-correlation process. The output of the Lindemann model emerges in the same format as the data provided by the Jeffress model: a function of cross-correlation and interaural time delay against time.

It is no easier to place cross-correlated data in a context of auditory events than it is to attempt this with an unprocessed binaural stream. Therefore to make proper use of the cross-correlation data, it is necessary to consult a separate and dedicated onset detector.

The detector detailed in this paper has been designed to work in tandem with a cross-correlation mechanism. It is proposed to work on general source material, and not just speech. It will take data from an auditory filter bank, as the cross-correlation model requires data in this format.

The onset detector is intended eventually to run in real time. Since the detection process is computationally intensive, speed of execution is favoured over accuracy whenever a compromise is needed. This real-time criterion also imposes another restriction on the detector: it is not permitted to look far ahead into the audio signal.

2. SPECIFICATION

Fig. 1 shows an overview of the detection procedure detailed in this paper. The binaural source is first divided into critical bands. Each of these signals is then passed through a separate predictive filter. The disparity between the predicted data and the actual data is measured to yield an error function. This function is correlated both to signal level and to the predictability of the input signal.

The error function is presented to an onset calculator. This is an arrangement of scaling, comparison, and filter operations, with carefully-chosen time constants. Finally, onset data from each critical band is compared and summed into one homogeneous signal.

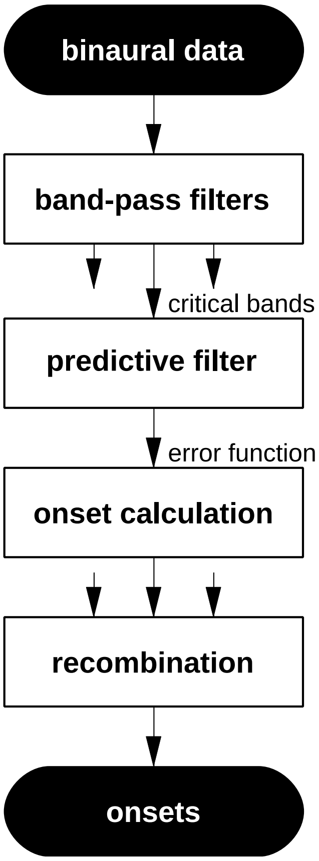


Fig. 1: An overview of the onset detection procedure. In practice, the band-pass filters will divide audio into more than the three bands shown — the current prototype uses sixteen — but it is rarely necessary to analyse all of these in practice.

2.1 Band-pass filters

The bank of band-pass filters divides the binaural data into a number of high-Q frequency bands. Slaney’s gammatone filter algorithm is used [11], with modified frequency and bandwidth parameters.

Currently, sixteen bands are produced, covering the range from 20Hz to 3.2kHz. This emulates the frequency division operation of the inner ear, and covers the lower seven octaves of the audio spectrum. Inside this range, the hearing system is particularly sensitive to interaural time differences. Outside this range, it favours interaural level differences. Therefore the cross-correlation model which this detector supports is not perceptually relevant above 3.2kHz.

Many more elaborate cochlear models take this frequency division operation a stage further, half-wave rectifying and then low-pass filtering each critical band to simulate the transduction mechanism of the inner ear. This process captures the amplitude envelope of higher-frequency bands and the rectified fine detail of lower-frequency ones. These operations have not been applied here, because distortion caused by rectification reduces the effectiveness of the predictive filter.

Although the current filter algorithm produces sixteen frequency bands, little is gained under most circumstances by analysing all of them: this tends only to duplicate information. The examples in section three of this paper show that useful information can be obtained from a smaller number of isolated critical bands.

2.2 The predictive filter

The use of a predictive filter for onset detection is not entirely novel. The novelty and effectiveness of this detection system resides in this implementation of its filter, and in the interpretation of its output.

A predictive filter is a class of adaptive filter. The input signal is passed through this standard digital filter, and the filter’s coefficients are modified continually, according to properties of the input signal.

This particular predictive filter is presented with audio data from the recent past, and is made to run past this data for several samples with its input forced to zero. Its output during this period is treated as a prediction of the

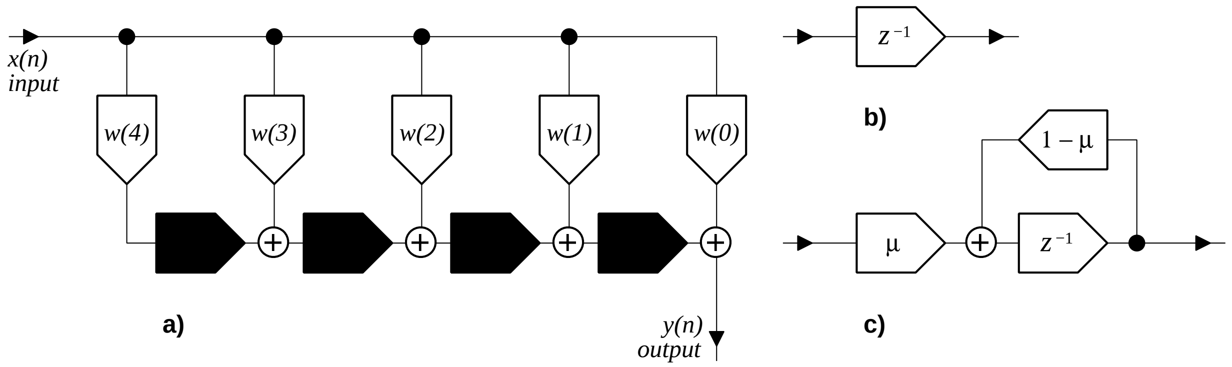


Fig. 2: FIR and gamma filter architectures compared. Part a) represents a generic fourth-order filter structure, in which *w*(0) to *w*(4) are weighting coefficients. The processes represented by black boxes in this figure are equivalent to the simple unit delay b) in a FIR filter, and to c) in a gamma filter. In gamma filters, extra flexibility is provided by the feedback variable μ, which may be altered in the same manner as the weighting coefficients. The same energy is presented to the system by processes b) and c), as long as 0 < μ ≤ 1.

next few audio samples.

The accuracy of this prediction is judged using a weighted squared difference function to compare it with the actual audio data. This is called the *error function* *e*(*N*), and may be represented algebraically (after Clarkson [12]):

$$e(N) = \sum_{k=0}^{K-1} w(k)\big(y(N+k) - d(N+k)\big)^2$$

where:

N = the current audio sample;

K = the prediction length in samples;

w(*k*) = a weighting function which gives priority to the earliest predicted values, and;

y(*n*) = data predicted by the filter;

d(*n*) = audio data.

This error function specifies the closeness of the prediction to the data. It is calculated many times during each operation with a number of experimental coefficients. Minimising the error function guides the prediction.

For the sake of simplicity, an exponential weighting function is used in the formula above:

$$w(k) = \alpha^k$$

α is the weighting coefficient; 0 < α < 1.

2.2.1 Filter architecture

A gamma filter [13] is employed as the predictive filter. This class of filter has been chosen because it is easy to control, stable, and enables long impulse responses to be expressed using a small number of filter coefficients. Low-frequency resolution in a gamma filter is therefore superior to that of a FIR filter of similar complexity.

A gamma filter can be seen as an extension of a FIR filter, where the unit delay operation is replaced with a dispersive delay. This stores a proportion of the input signal and releases it slowly, with an exponentially decaying characteristic. Fig. 2 shows this in more detail.

The weighting coefficients of this filter can be changed to alter its response. Additionally, the feedback constants within the delays may also be changed. This makes the system quite flexible. (The gamma filter architecture is exploited similarly in the onset detector by Schwartz et al. [14]).

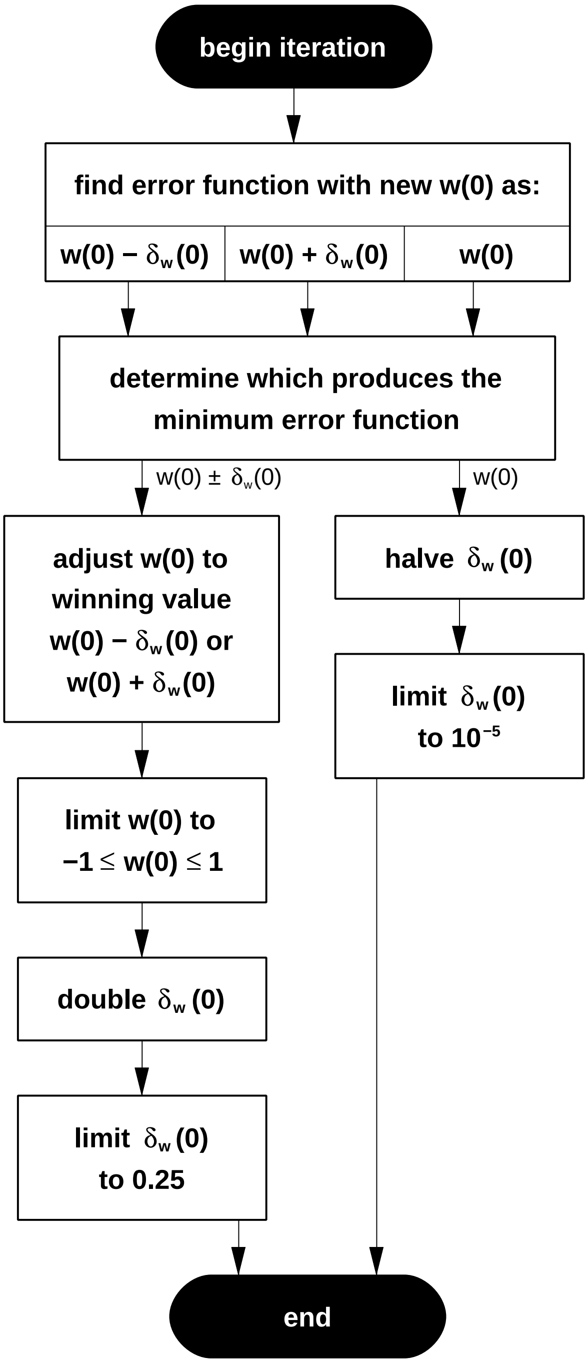

begin iteration
find error function with new w(0) as:
w(0) − δw(0)
w(0) + δw(0)
w(0)
determine which produces the minimum error function
w(0) ± δw(0)
w(0)
adjust w(0) to winning value w(0) − δw(0) or w(0) + δw(0)
halve δw (0)
limit δw (0) to 10⁻⁵
limit w(0) to −1 ≤ w(0) ≤ 1
double δw (0)
limit δw (0) to 0.25
end

Fig. 3: Flow chart of the adaption procedure for *w*(0). This procedure is performed twice before moving on to the next constant, μ(1), and then on to *w*(1). When the whole sequence is completed as far as w(*P*), the routine moves to the next sample.

2.2.2 Adaption

This section describes in detail the way in which filter coefficients are changed so that the error value approaches a minimum. The surface of $e(N)$ has more than one minimum. Finding the optimal solution — the lowest minimum of $e(N)$ — is comparable to trying to locate, at every instant, the lowest ripple on a windy lake. To keep this model fast, it is designed to find only a nearby minimum. This following procedure has been tested for input data with a sampling frequency of 11.025kHz, which is output from the filter bank. Slight changes would be necessary to control the rate of change of coefficients for other sampling frequencies.

The filter weighting coefficients $w(0\ldots P)$ are initialised to zero, and the feedback constants $\mu(1\ldots P)$ to 1. Two more arrays, $\delta_w(0\ldots P)$ and $\delta_\mu(1\ldots P)$, determine the search interval. These are initialised, arbitrarily, to 0.1.

Fig. 3 charts the adaption process for one coefficient. As a new audio sample enters the system, a prediction error is calculated using the current coefficients. It is then calculated again, changing $w(p)$ to trial values of $w(p)-\delta_w(p)$ and $w(p)+\delta_w(p)$. If either of the error values thus produced is lower than the original value, $w(p)$ is changed permanently to the trial value, and the search interval $\delta_w(p)$ is doubled. If the original value is found to be the lowest of the three, $\delta_w(p)$ is halved. A second pass is made of this procedure before moving on to the next coefficient. All of the coefficients are adjusted in the order $w(0)$, $\mu(1) \ldots \mu(P-1)$, $w(P)$.

Finally, a value of the error function, $e(N)$, is derived from the most recent coefficients. Before moving on to the next sample, the values contained within the filter's delay taps are stored so that it can be run quickly and accurately again: this speeds up subsequent calculations considerably.

All parameters are limited throughout this operation so that the filter remains stable. The following conditions are forcibly maintained:

$$-1 \leq w(p) \leq 1; \qquad 0 < \mu(p) \leq 1.$$

The first boundary condition prevents the filter attempting to react to a sudden transient in the input signal, after a prolonged period of low input level, by multiplying the small numbers stored within filter taps by enormous weighting coefficients. This behaviour was observed in early experiments. It greatly increased the adaption time of the filter, and huge, momentary spikes were produced in the error function as the filter took time to recover.

The second boundary condition is a requirement of gamma filters. All other values of $\mu(p)$ produce results which are unstable ($\mu(p) > 1$), zero ($\mu(p) = 0$), or oscillate powerfully at half the sampling frequency ($\mu(p) < 0$).

In addition to these, two other boundary conditions determine the accuracy and the speed of the search for an approximation:

$$10^{-5} \le \delta_w(p) \le 0.25; \quad 10^{-5} \le \delta_\mu(p) \le 0.125.$$

A fourth-order gamma filter, with a prediction length of twelve samples, has been used to create the examples presented throughout this paper.

2.3 Onset calculation

The error function possesses two important characteristics. Firstly, noise-based and suddenly-changing parts of the signal are emphasised relative to steady-state and periodic parts. This is a particular advantage when processing sounds produced by percussive and plucked instruments, since these tend to have sudden, noisy attacks and relatively slow, periodic decays. Since the the predictive filter is a linear and deterministic process, the error function will also vary in proportion to signal level.

The onset calculator must be sensitive enough to detect small onsets, and robust enough for onsets not to be detected where they do not exist. An entirely data-driven approach must be taken at this stage, inspecting the error function whilst auditioning the audio. It is possible to formulate two axioms by inspecting the output data:

1. When an onset occurs, the error function becomes substantially higher than its predecessors. Therefore both its absolute value and its gradient function become high.

2. Sometimes, a sharp discontinuity will occur in the error function, even during steady-state signals. This is caused by a lack of stored energy within the filter. The predicted signal will be severely limited in amplitude, and will not be able to track a higher-energy signal effectively. These discontinuities may be distinguished from genuine onsets because the error function decays more rapidly. Genuine onsets can be seen from inspection to possess a noisy exponential decay, which takes far longer to fade away.

An effective solution which satisfies both axioms entails the construction of a tracking function. This is designed to follow the error function closely. In other onset detection models, onsets are marked either when a tracking function exceeds a certain value (e.g. [6], [15]), or when the error function exceeds a tracking function by a certain amount (e.g. [14]). However, there is no fundamental methodological difference between these two approaches.

In the model used here, an onset will be detected at any instant where the error function exceeds the tracking function. The basis of this tracking function is an amplified and low-pass filtered version of the error function, so that an onset will be detected only if the error function increases quickly enough.

This means that the choice of rise time for the low-pass filter is critical: it must be fast enough for discontinuities to be rejected, but not so fast that the error function is tracked too closely and onsets are not discovered, or are discovered too late.

This model uses a variable rise time to solve the problem. The rise time is very short to begin with: an initial rise of 10dB/ms has been found to work satisfactorily. As consecutive increases are required in the tracking function, this rise time is quickly reduced. At any instant when the tracking function is decreased, the rise time is reset to its initial value.

A different time constant is used to reduce the level of the tracking function. This is considerably slower, and models the precedence effect. Barron [16] and Haas [4] plotted a series of level-time characteristics, determining the threshold of independent perception for a reflection in the presence of the direct sound. Both suggest a slope of 0.3dB/ms, and this value has been used in the onset calculator. Whenever a fall in the tracking function is required, it is pulled to zero using this time constant. Otherwise a dual slope will be traced: one caused by the decay of the error function, and one caused by the fall time of the tracking function. Pulling the tracking function to zero keeps the system responsive to quick changes.

Haas observed that an interfering speech reflection, delayed between 5ms and 25ms relative to the direct sound, must be up to 10dB louder than the direct sound to be judged as equally loud. This relationship can be extrapolated to join Barron's plot of the threshold at which an early reflection is judged to be 'disturbing'. An simplification of this relationship has been incorporated. Whenever an onset is detected, the tracking function is raised 10dB higher than the error function, and is held there for 25ms before it is allowed to decay.

Before comparing the error and tracking functions for onset detection, a proportion of the long-term average of the error function is added to the tracking function. This raises it further above the noise floor of the signal, so that the detector is less susceptible to very small fluctuations during quiet sections of audio.

2.4 Recombination

At present, a simple, empirically devised system is used to combine onsets detected in each critical band. At each sample point on every critical band, a value is calculated depending on the time that has elapsed since the last onset. These *onset potentials* are combined additively across every critical band to produce an output.

The potential of each band is calculated using a simple exponential function of time:

$$V(t) = \beta^{t-t_m}$$

where t_m is the time of the previous marked onset. β is chosen so that when $t-t_m = 1 / f_c$, where f_c is the centre frequency of the critical band, the potential is $1 / \sqrt{2}$.

When the potential equals or exceeds a certain threshold, an overall onset is flagged. No more onsets may be flagged within 50ms, or until the potential falls below this level again. This threshold is currently 1.12 (= 1dB above 1.0), regardless of the number of critical bands being combined. Otherwise, narrow-band signal changes are lost as further critical bands are combined.

3. EXAMPLE SIGNALS

To illustrate the principles of operation of the onset detector, two recorded signals are now presented and analysed. These examples were acquired in a classical recording studio with a floor area of 250m² and a reverberation time of around 1.5s, using a Neutrik dummy head. The stimuli have been chosen because they pose specific challenges with which the onset detector must cope.

3.1 Speech extract

The first example is an adult male speaking the sentence 'How many cellists does it take to change a light bulb?' This was recorded five metres in front of the dummy head. The recording head was placed with its centre about 15cm from a boundary wall. No special attempt was made by the speaker to enunciate clearly. Some of the less important words therefore run into one another, or are too quiet to hear easily.

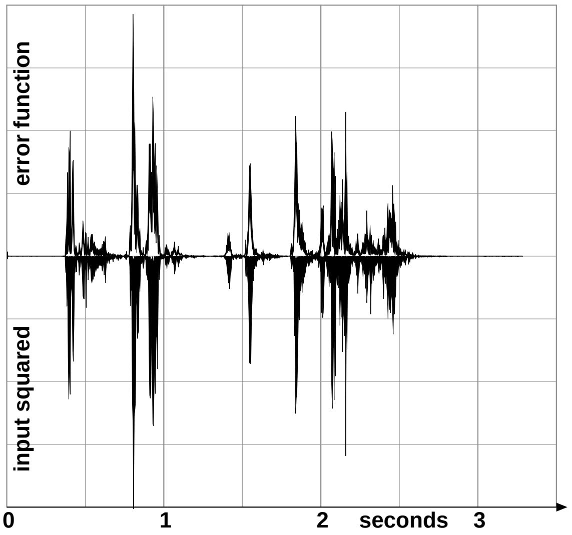


Fig. 4: The predictive filter's error function compared with the square of the input signal [linear graph scale]. Entire speech extract. 920–1080Hz frequency band. To facilitate comparison, the signal-squared function has been inverted, slightly scaled, and placed underneath the error function.

In Fig. 4, the error function of the predictive filter is shown against the square of its input signal. Both are taken from the 920–1080Hz critical band. The signals appear superficially to be very similar, and most of the syllables show very clearly. However, the difficulty in this kind of onset detection is that onsets must be noted as they occur, with very minimal knowledge of the forthcoming signal.

For the purposes of this demonstration, it is sufficient only to show the left channel being treated. This treatment will illustrate each part of the onset detection process, and will demonstrate that the predictive filter makes onset detection more precise by sharpening and exaggerating onsets, and by smoothing and lowering steady-state portions.

Fig. 5 shows part of the speech example where an onset is detected. The differences between the input signal and the error function are much clearer at this scale, and the effect of onset enhancement is easy to notice.

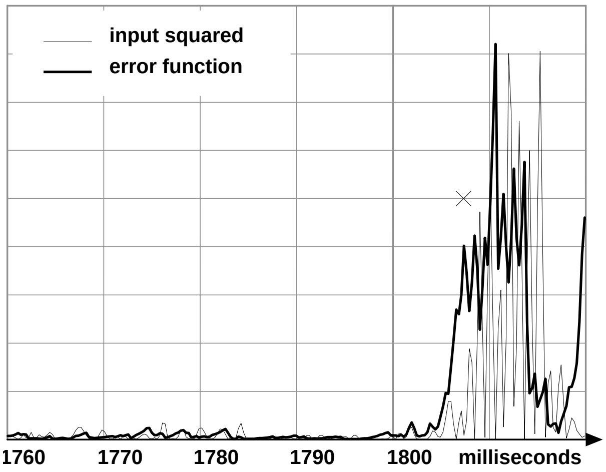


Fig. 5: Magnified portion of Fig. 4 — the interval between the words ‘to change’. Comparison of the square of the predictive filter’s input and its error function. 920–1080Hz frequency band. The functions have been scaled mutually for clarity. The cross marks the time of detected onset.

The early portion of this detail shows mostly reverberant decay. Here, the error function remains consistently lower than the input, and its gradient is flatter. When the characteristic of the signal changes suddenly, in a manner which the filter is unable to predict, the error function’s instantaneous value and its gradient greatly exceed those of the input.

This phenomenon cannot be attributed to the filter looking ahead and its error function being influenced by later, higher values of input function. In this example, the prediction length is twelve samples, or 2.7ms at this sampling frequency. The exponential weighting constant α is set to 0.765. This value was chosen so that the twelfth sample has 4% of the weight of the first sample. The onset enhancement and steady-state suppression seen in this extract is therefore entirely an artefact of the predictive filter.

Fig. 6 shows the relationship between the error function and the onset calculator’s tracking function. The different attack and decay time constants of this function are clear. The tracking function can be seen here to increase in level by 10dB and freeze for 25ms after the onset event. What is most interesting about this figure is the earliness of the onset detection, and the level of the error function throughout. Its level at onset is at least 15dB lower than its value only 20ms later, yet no additional onsets are detected and the mechanism is not confused by the noise-like signal.

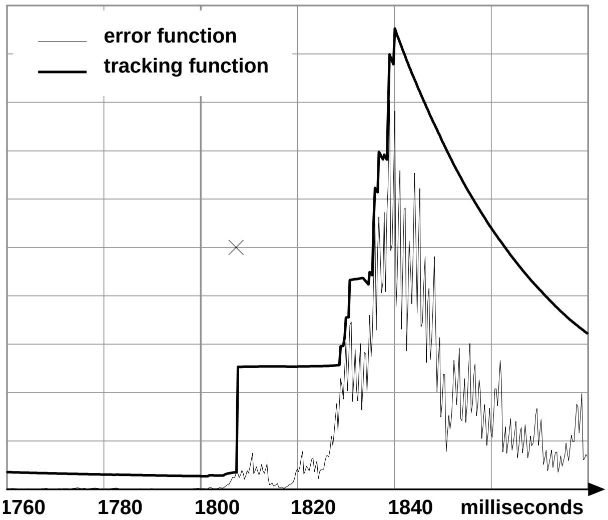


Fig. 6: Zooming out from Fig. 5 to compare the tracking function and the error function. In this figure, the signals are mutually to scale.

Five critical bands from this signal have been run through the onset detection mechanism, and are recombined to produce the trace in Fig. 7. After recombination, ten onset decisions have been detected in this thirteen-syllable extract.

Detected onsets were auditioned in context by inserting a short click into the audio signal where each one was found, and then playing back the audio at half speed.

Each onset is placed at or near the start of a syllable. During the word ‘cellists’, for example, it is not the initial fricative which is marked as an onset, but the vowel after it.

The correlation between onsets and syllables is not surprising, since the start of a syllable represents a sudden change both in signal level and predictability. It is rather more surprising that there are no onsets detected during smaller transitions within the syllables — for example, between the vowel and fricatives at the end of the word ‘cellists’ — and therefore that the onset detector works remarkably well as a syllable detector. (Incidentally, the three syllables which the detector did not locate are the words ‘does’, ‘it’ and ‘a’; these were barely articulated).

However, this process is not intended as a syllable detector, and its performance as such is only a curiosity. Informal examination of the cross-correlation of this speech signal revealed that the correlogram pattern does not change significantly throughout. This is for two reasons. Firstly, the information rate of speech is high: ten onsets are detected in this example in two seconds, and many changes which were only slightly less significant were not marked. Secondly, the head and speaker were oriented in such a way that reflections from the wall did not disrupt the interaural time differences caused by the direct sound: instead, they reinforced them.

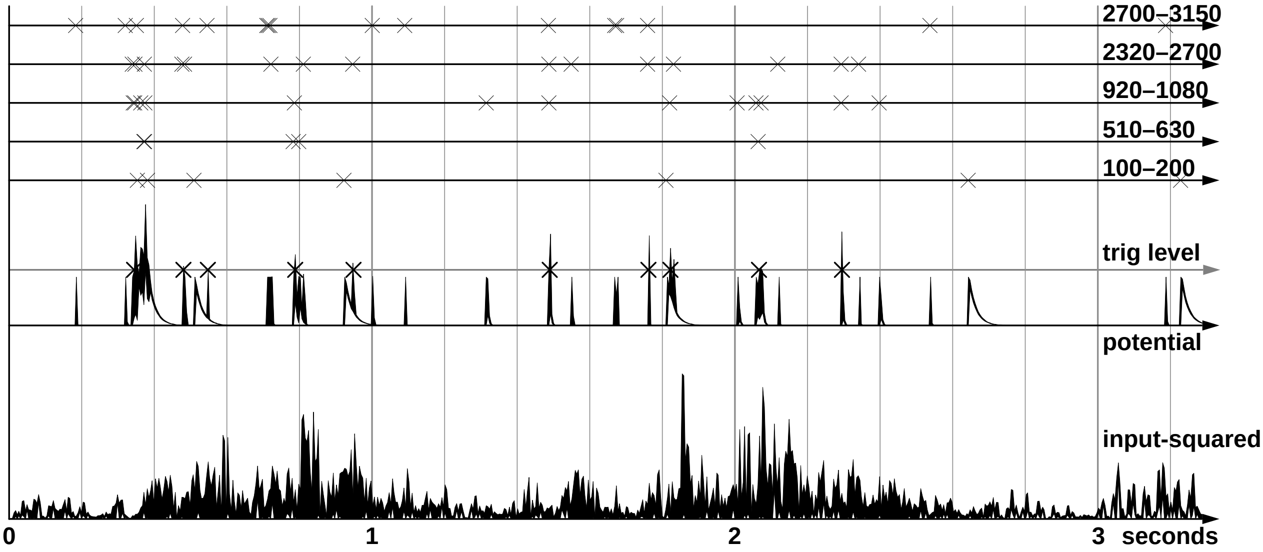
To examine the interaction of the onset detector with a simple cross-correlation function, a more critical, non-speech extract will need to be examined.

3.2 Three Instruments

This extract features three musical instruments, with different methods of producing sounds. They were chosen for their different spectral, attack and decay characteristics. These instruments were recorded at different times in the studio, keeping the dummy head in the same position. Afterwards, they were mixed to form the three-second except.

The instruments, their positions, and their musical motives are as follows:

Fig. 7: **Onset decisions for the speech extract. *Top axes:* Onset decisions for five critical bands. The numbers refer to their frequency ranges. *Middle axis:* Potential function *V*(*n*) for the combination of onset decisions. The onsets are marked with crosses on the ‘trig level’ line. *Bottom axis:* The square of the audio input signal.**



- Grand piano, 40° right. Four staccato octaves in the bass.
- Clarinet, 40° left, three detached notes in the upper register.
- Snare drum, 80° right, ratamacue (six notes).

All instruments were played five metres from the dummy head. Fig. 8 shows the performance of the onset detector across the first seven critical bands for the left and right ear signals.

It is clear that the signal is of considerable complexity: this is particularly noticeable in the first 0.5s of the waveform, where the instruments start almost simultaneously. The contribution that each instrument makes to the waveform in Fig. 8 is clear: the clarinet notes are sustained and decay suddenly, whilst the piano notes are almost triangular in profile on the logarithmic scale. The snare drum rudiment adds a number of spikes across the first second of the signal. Spatial displacement of the clarinet (left) and the piano (right) are also quite clear if the levels of the left and right waveforms are compared. Clearly, interaural level differences feature as a cue even around 700Hz.

A unified onset decision was made from the two ear signals by merging the overall onset decisions, and then removing the later of any set of two onsets which appeared within five milliseconds of one another. Of the twenty-three onsets on this figure, only one was removed by this process. The algorithm thus found twenty-two onsets in the extract. This compares with the thirteen notes which can be counted in the example.

The large number of onsets may be attributed to the nature of noise from the snare, which has undoubtedly created many extra onsets within the first second of the extract. In the following section, these onsets are used to drive a cross-correlation function.

3.2.1 Cross-correlation analysis

Fig. 9 is a correlogram taken around the 810ms onset. Interestingly, some of the largest values of interaural cross-correlation occur just before the onset is detected. This reinforces one of the main arguments behind this paper: that without a sense of context in which to interpret the events within it, a raw correlogram is of little use. However, it is clear that one of the most precise results in the correlogram occurs in the first five milliseconds after the attack is detected. A peak is seen to move between about 40° and 50° right. This is reasonable, as a two metre grand piano can be shown to subtend twenty degrees of a listener's field at around five metres.

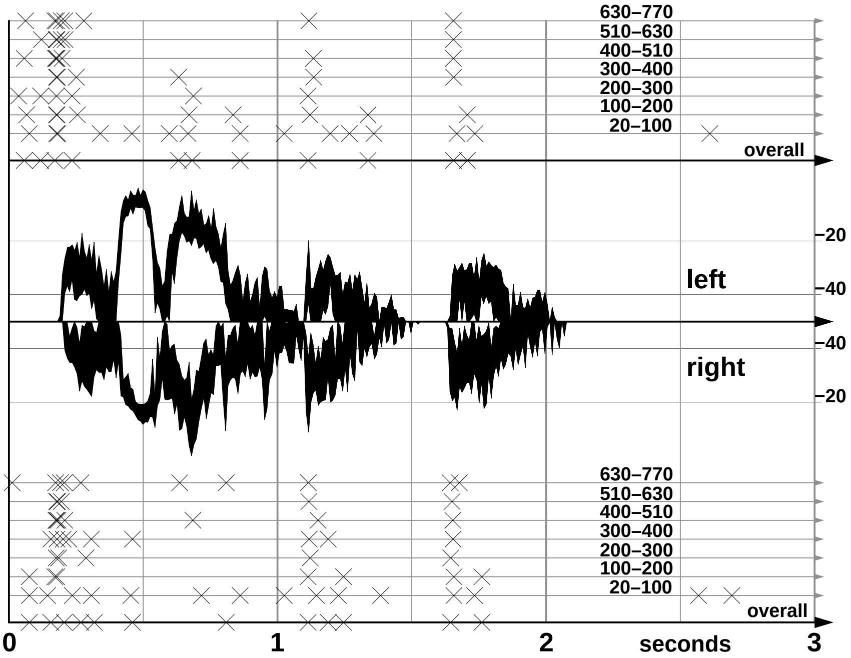


Fig. 8: Onset decisions for the three-instrument mixture; left and right ears.

Onsets are plotted as in Fig. 7. The entire top half of this figure was calculated from the left ear signal. The bottom half was calculated from the right.

The central waveform, included for reference, is the error function from the 630–770Hz critical band.

As the error function is always positive, the signal from the right ear signal has been inverted and placed underneath the left ear signal.

The waveforms are plotted in dB relative to the common peak level.

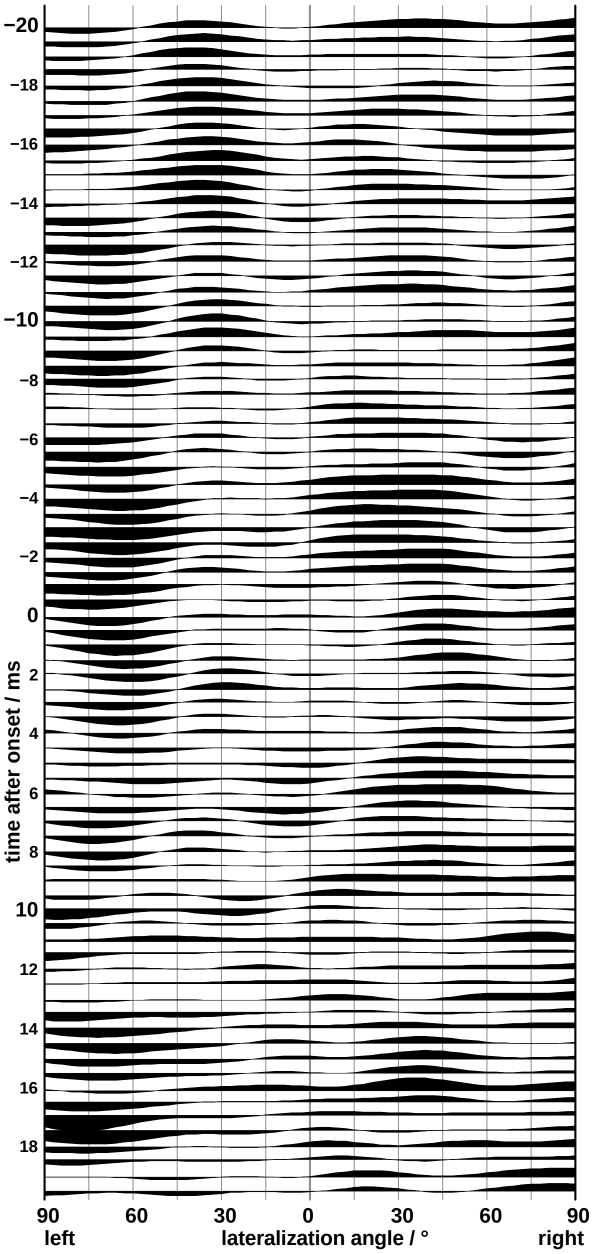


Fig. 9: Correlogram. This run was started 20ms before the right-ear onset at 810ms, and finishes 20ms after it. Time increases down the y-axis. Each segment is the average of 500µs of individually normalised readings, taken across every critical band for every audio sample using the international standard formula [9]. The x-axis has been warped to show the lateralization angle. The peak absolute value of this set of data is about 0.4.

Excepting the clarinet tone which appears at 40° left at the very beginning of this correlogram, the only other correct angular information displayed happens 14ms after attack, where the piano is again correctly localised. All other usable information in Fig. 9 appears to be disturbed by interfering sounds.

Fig. 10 shows a close-up of the 810ms onset in the time domain. Intriguingly, it appears not to be an onset at all, but a sudden offset. It is fortunate in this circumstance that the decay characteristic of a piano allows the instrument to sound clearly above other sounds, including its reflections.

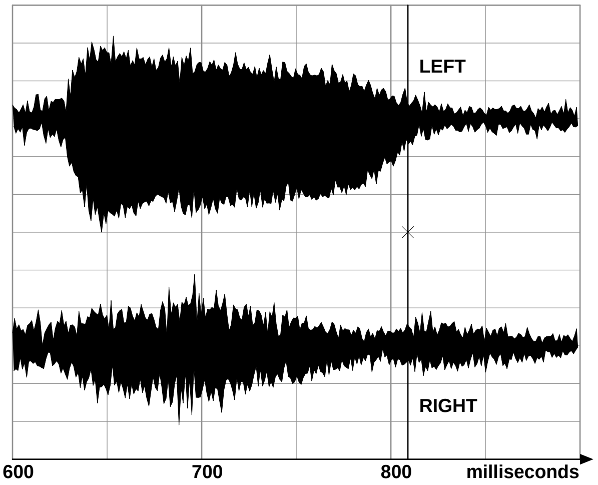


Fig. 10: A magnification of the source waveform, showing the 810ms onset in detail. Although there is some sharp detail in the right-hand waveform shortly after the onset is marked, it seems more likely that the detector has marked a sudden offset.

Fig. 11a depicts a number of correlograms taken at each onset point, and then held until the next onset. The onset detector thus feeds a basic scene analysis task, which locates the position of each onset.

The holding operation emulates the way in which the human auditory system is presumed to work for the first fifty milliseconds after onset. Under these circumstances, the leading millisecond of the new auditory event determines its position. Further information, excepting any overriding onset, contributes to spatial attributes of the located source. This assertion seems to fit comfortably with studies into the precedence effect conducted by Wallach et al. [3], Haas [4], Barron [16], and Barron and Marshall [17]. Although Wallach et al. suggest that early reflections can influence the perceived location of the fused sound, this phenomenon occurs only under certain circumstances. Furthermore, Wallach et al. maintain that the reflection may influence the fused image to a maximum of only seven degrees.

The cross-correlation shows a strong a peak at about twenty degrees right, even though there is no instrument placed there. This could be attributed to a strong early reflection from the back wall.

It is clear that the onset-based results of Fig. 11a show the positions of the instruments better than the control data in Fig. 11b. In Fig. 11b, the steady-state clarinet tones and the slowly-decaying piano tones are portrayed quite clearly at 40° left and right, but the sharply-attacking snare drum at 80° right is almost entirely lost.

However, neither result is definitive. Although the spatial rendering in Fig. 11a is reasonably accurate, it is not possible to see the beginnings of notes, to count the number which occur, or to determine what is playing when. For example, one cannot see that the clarinet notes are detached from one another, that they are evenly spaced, or even that there are three of them. In fact, there is very little precise detail in the time domain.

4. DISCUSSION

It has been shown that the onset detection model is able to handle recorded speech precisely, and can interpret a complex musical stimulus to a lesser degree. There are plenty of ways to set about improving this model to improve its performance with complicated source material.

Much of the system which this paper describes has been guided by informed trial and error, or based on other researchers' work which has been formulated in a similar way. The process of continually adapting, testing, and refining such a model is inevitable: there is no other way of emulating a complicated neural process. Developing the tracking function in particular required a synthesis of extant psychoacoustics knowledge (for example, in the implementation of the hold and fall time constants), and informed guessing and refinement to

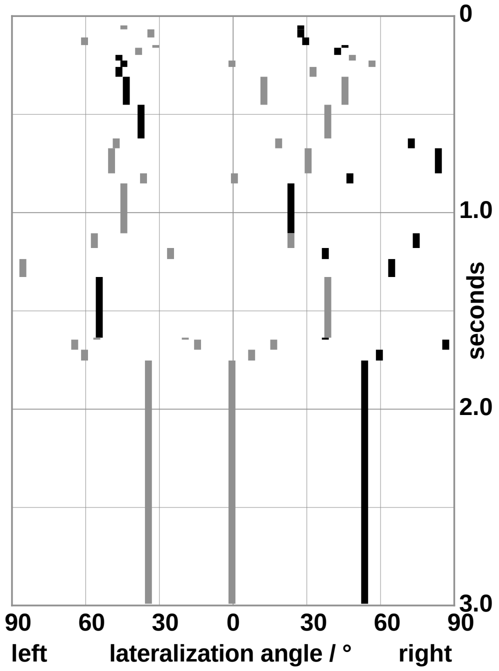
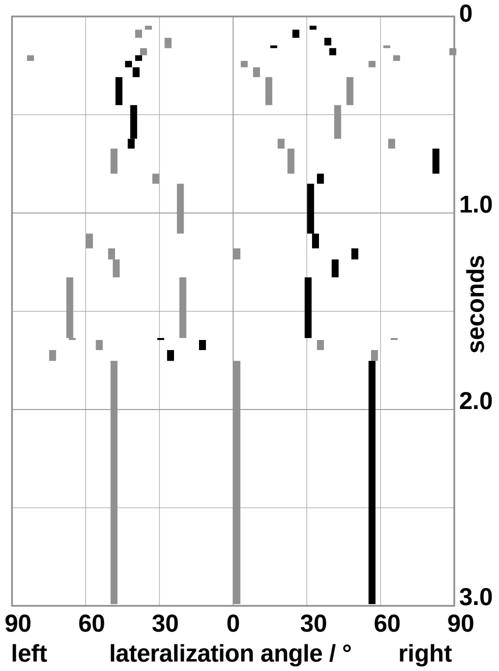


Fig. 11a (above): Onset-triggered cross-correlation results. Each black vertical line represents the maximum correlation in each time frame; the bold grey lines represent less significant maxima.

Fig. 11b (below): For comparison, correlograms taken 2ms before the onsets.



make the results fit the test stimuli (for example, in the development of the rise time constant mechanism).

Unfortunately, making these approximations harms the verisimilitude of the output. The outputs from the test material used here, and some others, suggest that the performance of the detector in the lowest two critical bands is noticeably inferior to that at higher frequencies.

The onset driven correlogram in the second extract would suggest that some aspects of the detection algorithm need to be refined in order to improve the temporal information conveyed. Perhaps a method of discriminating between onsets and offsets, or of memorising the positions of auditory events, needs to be employed; perhaps a satisfactory solution can be found more simply by refining the tracking function.

There may also be ways to improve the onset potential function used in recombining the input data. Its sensitivity reduces drastically as the time between onsets across critical bands is increased even slightly. The recombination process currently sums each critical band with equal weighting; however, it is well known that some frequency ranges are more audible, and hence more salient, than others. The potential function is also causal. If it was permitted to look a few samples into the future (as the predictive filter does), onsets could be flagged earlier. However, the difference such a modification would make is hardly likely to amount to more than a few hundred microseconds, and circumstantial evidence suggests that onsets in high-frequency critical bands are marked no earlier than those in low-frequency bands, in spite of their faster rate of change (Fig. 12).

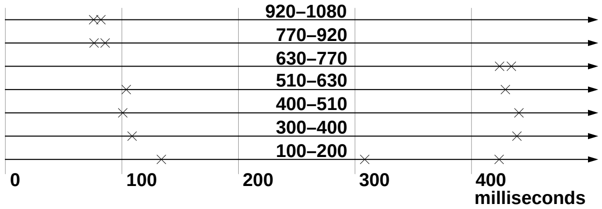


Fig. 12: A close-up of two onsets detected in an except of solo piano music. Onsets are not necessarily located in higher-frequency critical bands before lower-frequency ones, in spite of their faster theoretical response to change.

The wisdom of making confluent binary decisions about the presence or absence of onsets may be questioned. At present, an onset decision is made at every point on every critical band. Each decision is combined into a binary onset decision for each ear, and then these are consolidated into an overall onset decision. This approach is rather inelegant. It compounds approximations, and means that the precision of the error and tracking functions are discarded entirely in the process of making an onset judgement in each critical band. Better results might be produced if a fuzzy logic process were employed instead, basing its overall onset decision upon the weighted probability of an onset occurring on each critical band. The probability function could be based on the rise time of the error function, or on its relationship to a more sensitive tracking function. However, such a change would entail a considerably more sophisticated algorithm with a more advanced model of the precedence effect. It would also mean that the onset calculator and recombination processes could no longer be self-contained.

In time, it may be necessary to adjust the sensitivity of the predictive filter by altering the rate of change of its coefficients, to manipulate its parameters, or to modify the onset calculator mechanism by comparing its output with subjective data gathered from human listeners under controlled conditions.

Finally, it is important to emphasise the intended limitations of this model — specifically, the difference between this auditory onset detector and similar designs which attempt to perform note or syllable detection. This system has been designed to work alongside a spatial feature analyser so that positional information about the source, and the spatial nature of the acoustics of the environment, may be investigated. This onset detection task is simpler than note detection. For example, when a bowed instrument plays legato, the pitch of the note will change quickly while its amplitude envelope remains as continuous as any constant note. Meanwhile, early reflections will be inhibited by the direct sound. Therefore, it is doubtful that much spatial information will be conveyed during the changing notes unless the fluctuation is such that the signal crosses several critical bands. While a note detector would have to uncover these changes, it is not within the remit of this onset detector.

5. ACKNOWLEDGEMENTS

We would like to thank Pharos Communications Ltd for sponsoring this research project, and David Griesinger for giving us the code for his syllable detector.

Thanks also to Russell Mason and Miriam Iorwerth of the Institute of Sound Recording for helping to collect binaural material.

6. REFERENCES

[1] Bregman, A. “Auditory Scene Analysis: The Perceptual Organization of Sound,” Cambridge, MA: MIT Press (1993).

[2] Mason, R., Rumsey, F. “Interaural time difference fluctuations: their measurement, subjective perceptual effect, and application in sound reproduction,” *Proc. AES 19th International Conference*, pp. 252–271 (2001).

[3] Wallach, H., Newman, E. B., and Rosenzweig, M. R. “The Precedence Effect in Sound Localization,” *Am. J. Psychol.* vol. 52, pp. 315–336 (1949). Reprinted in *J. Aud. Eng. Soc*, vol. 21, pp. 817–826 (1973).

[4] Haas, H. “The influence of a single echo on the audibility of speech,” *J. Aud. Eng. Soc.*, vol. 20, pp. 146–159 (1972).

[5] Gardner, M. “Historical Background of the Haas and/or Precedence Effect,” *J. Acoust. Soc. Am.*, vol. 43, pp. 1243–1248 (1968).

[6] Huang, J. “A Biomimetic System for Localization and Separation of Multiple Sound Sources,” *IEEE Trans. Instrumentation and Measurement*, vol. 44, pp. 733–737 (1995 June).

[7] Martin, K. “A Computational Model of Spatial Hearing,” M.Sc. thesis, Machine Listening Group, MIT, Cambridge, MA. (1995).

[8] Jeffress, L. “A place theory of sound localization,” *J. Comparative and Physiological Psychology*, vol. 61, pp. 468–486 (1948).

[9] BS EN ISO 3382:2000. “Acoustics. Measurement of the reverberation time of rooms with reference to other acoustical parameters,” British Standards Publishing Limited (2000).

[10] Lindemann, W. “Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals,” *J. Acoust. Soc. Am.*, vol. 80, pp. 1608–1622 (1986 December).

[11] Slaney, M. “An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank,” *Apple Computer Technical Report #35*, (1993).

[12] Clarkson, P. “Optimal and adaptive signal processing,” Boca Raton, FL: CRC Press (1993).

[13] Principe, J., De Vries, B., and De Oliviera, P. G. “The gamma filter — a new class of adaptive IIR filters with restricted feedback,” *IEEE Trans. Signal Processing*, vol. 41, pp. 649–656 (1993).

[14] Schwartz, O., Harris, J., and Principe, J. “Modeling the precedence effect for speech using the gamma filter,” *Neural Networks*, vol. 12, pp. 409–417 (1999).

[15] Griesinger, David. Personal correspondence (2002).

[16] Barron, M. “The subjective effects of first reflections in concert halls — the need for lateral reflections,” *J. Sound and Vibration*, vol. 15, pp. 475–494 (1971).

[17] Barron, M., Marshall, A. “Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure,” *J. Sound and Vibration*, vol. 77, pp. 211–232 (1981).