Adjacent-Channel Inhibition in Acoustic Onset Detection

Tareq Shahwan

Dept. of Electrical Engineering
San Jose State University
San Jose, CA 95192

Richard O. Duda

Dept. of Electrical Engineering
San Jose State University
San Jose, CA 95192

Abstract

Simple acoustic onset detectors respond to sudden changes in energy levels. Unfortunately, they respond to frequency modulation as well as to amplitude modulation, and have high false-alarm rates in the presence of siren or chirp signals. We show how adjacent-channel inhibition can be used to greatly reduce the false-alarm rate. An LMS procedure is used to obtain optimal inhibitory weights. Computer simulations reveal how the response to chirps varies with the rate of change of instantaneous frequency.

1 Introduction

Like the detection of edges in images, the detection of onsets in sound signals is a fundamental problem. The detection of common onsets is a primary cue for sound source separation, and is important for echo suppression in sound localization. However, extracting onsets reliably from acoustic waveforms is a challenge.

A standard approach is to pass the input signal through a filter bank that decomposes it into many different frequency bands or channels. The outputs of each channel are then processed independently, typically by rectifying the output and applying a smoothed derivative operator to detect sudden changes in energy [4, 2]. While this approach can readily detect the sudden onset of a fixed-frequency signal, it does not perform well with the wide range of signal amplitudes and attack rates encountered in practice unless automatic gain control and multi-scale derivatives, or similar enhancements, are used. Equally important, simple onset detectors suffer from false alarms when either short-time tone bursts or frequency-swept signals such as chirps or sirens are encountered.

The reason for this sensitivity to frequency-swept signals is easy to understand qualitatively. Consider a rising-frequency chirp whose instantaneous frequency passes through the center frequency f_c of a particular channel. If the sweep rate is moderate, the output of the channel will rise steadily, reach a peak, and then fall off as the chirp frequency passes out of the channel. From the standpoint of an observer monitoring the energy in that channel, the response is basically indistinguishable from that of a signal of constant frequency f_c having a similar attack and decay time. The resulting false alarms have been noted both by Mellinger [4] and Brown [2] as a factor that complicates the use of onsets in auditory scene analysis.

Neurophysiologists studying the peripheral auditory system of the cat have identified neurons in the cochlear nucleus that exhibit a sharp peak in activity at the beginning of a tone burst, and little activity thereafter. It is particularly interesting that these cells either give an inhibitory response or no response to swept-frequency signals. This has been attributed to a form of lateral inhibition, since these same cells also receive inhibitory inputs from cells whose characteristic frequencies are adjacent to their own characteristic frequency [6].

Mellinger and Brown both describe how such inhibitory connections might be employed in a filter-bank model of the auditory periphery. They suggest essentially the same solution, which is basically to subtract the delayed responses of adjacent channels from the input to a smoothed derivative operation. However, neither Mellinger nor Brown implemented this proposed solution, and we found that it was more difficult than might be expected to suppress the chirp response without also missing true onsets. This paper presents an alternative approach to implementing adjacent-channel inhibition and provides quantitative results showing that is effective in reducing the response to frequency-swept signals without seriously degrading the response to true onsets.

1058-6393/95 $4.00 © 1995 IEEE

2 Method of Approach

The onset-detection procedure we developed does not use smoothed derivatives. Instead, it is motivated by statistical principles of change detection, which have been effective in a similar problem of waveform segmentation[1]. The basic procedure for single-channel detection is as follows:

1. Rectify the filter output and track its envelope $x(t)$.
2. Estimate the time-varying mean of $x(t)$ by a low-pass moving average $m(t)$.
3. Estimate the time-varying variance by $s^2(t)$, a low-pass moving average of $(x(t) - m(t))^2$.
4. Form the standardized error $e(t)$, where $e(t) = (x(t) - m(t))/s(t)$.
5. Use a peak detector to locate peaks in $e(t)$. Let t_k be the time of the kth peak.
6. For each peak, find the time interval T_k over which $e(t)$ exceeds some fixed fraction α_1 of the peak value $e(t_k)$.

Of course, one could detect changes in the channel output merely by choosing a threshold θ and declaring a change whenever the standardized error $e(t)$ exceeds θ. However, not every change in a channel output corresponds to a tonal onset. For example, Fig. 1 shows the response of the band-pass filter to four different stimuli — a click, a very short tone burst, a chirp, and a step tone with a 10-ms rise time. Note that while all of these cases represent significant changes, we want the onset-detector to respond only to the step tone. Clearly, the primary feature that discriminates between these signals is the duration of the response. Thus, we use the time duration T_k to distinguish long tone bursts from other changes, and use the peak value of the standardized error $e(t)$ as a measure of confidence that an onset has occurred. We also found it necessary to delay the standard deviation estimate $s(t)$ by an amount τ_s. Otherwise, the sudden increase in $x(t) - m(t)$ that occurs at an onset would immediately produce a large and essentially meaningless value for $s(t)$ that in turn would greatly suppress the standardized error.[1]

This basic procedure can be effective at separating tonal onsets from clicks and short tone bursts over a wide range of signal amplitudes, and it can reject either very slowly changing or very rapidly changing chirps. However, there is a non-negligible intermediate range of chirps for which the frequency changes fast enough to produce a significant value for $e(t)$ and also stays in the channel passband long enough to produce a long interval T_k. The responses of adjacent channels provide the information needed to suppress these unwanted responses. Neglecting the slightly different delay times for different channels, we can say that the responses to a step tone occur essentially simultaneously in all adjacent channels, while the responses to a frequency-swept tone will be displaced. While the amount of displacement is hard to measure for either slowly swept tones or very rapidly swept tones (which are like clicks), such signals do not evoke onset responses. Fortunately, the differences in onset times Δt_k are measurable for roughly the range of sweep rates that would cause a false response for a decision based on the time interval T_k alone.

A block diagram for one channel of our system is shown in Fig. 2. We extract three onset-related features from the signal envelope: the time duration T_k, the peak value $e(t_k)$ of the standardized error, and the onset time difference Δt_k. A classical LMS procedure is used to combine the logarithms of these three features linearly. The final result is an onset measure y in the range $[0, 1]$, where $y = 0$ for no onset and $y = 1$ for an ideal onset. Adjacent-channel inhibition comes from the negative weight that becomes attached to the Δt_k input.

The onset measure can either be used in an onset map, as was done by Mellinger and Brown, or can be thresholded to yield an onset decision. In the remainder of this paper, we present the results of experiments showing the value of the adjacent-channel inhibition in reducing the false alarm response to frequency-swept tones.

3 Onset Detector Parameters

There are eight main parameters in the system: (1) the order of the band-pass filter, (2) the channel center frequency f_c, (3) the channel bandwidth Δf, (4) the spacing between adjacent channels, (5) the time

[1] An additional problem arises when the SNR is very high. In the noise-free case (rarely encountered in practice but frequently encountered in computer simulations), any slight departure of the envelope from its mean is reported as an onset. As Fig. 2 illustrates, we worked around this problem by substituting $\alpha_2 m(t)$ for the delayed $s(t)$ whenever the former was larger. Thus, in the infinite SNR case, the standardized error $e(t)$ becomes $(x/m - 1)/\alpha_2$, which depends on the input waveform, but not on the input amplitude. The value of α_2 was arbitrarily set at 0.1.

constant τ for the mean and standard-deviation estimators, (6) the additional delay τ_s for the standard deviation estimator, (7) the fraction α_1 of the peak standardized error used to define the onset interval T_k, and (8) the weights in the linear combiner.

We used Slaney's implementation of Patterson's fourth-order gammatone auditory model for the bandpass filter [8, 5]. While a complete model would include channels that span the audible range of frequencies, we simulated only one channel, arbitrarily choosing $f_c = 1000$ Hz as a representative channel frequency. We used the Glasberg-Moore relationship between center frequency and bandwidth ($\Delta f = 24.7(.00437 f_c + 1)$, [3]) to obtain $\Delta f = 133$ Hz. We spaced the two adjacent channels half the bandwidth below and above, with resulting center frequencies of 934 Hz and 1066 Hz, respectively.

The time constants τ used for low-pass estimates of the mean and standard deviation was set at 30 cycles of the center frequency, i.e., 30 ms. With this value, 8-cycle 1-kHz tone bursts with a 20-dB SNR produced a standardized error about 80% that of long-duration tone bursts. The additional delay τ_s was set at 10 cycles of the center frequency. Finally, we experimentally found $\alpha_1 = 0.1$ to be a reasonable choice for defining an onset interval that seemed to discriminate well between short and long tone bursts. While all of these parameters could be further optimized, these choices seemed to yield generally good performance.

4 Test Results

A set of training examples and a classical LMS procedure were used to determine the values of the weights in the linear combiner. To obtain a unique solution, the training data included examples of signals with and without onsets that were difficult to separate:

1. 48 1-kHz tones, rise times from 0 to 100 ms, SNR's from -6 dB to 60 dB
2. 48 long 1-kHz tone bursts, durations from 5 to 20 cycles, SNR's from -6 dB to 60 dB
3. 24 short 1-kHz tone bursts, durations from 0.5 to 2 cycles, SNR's from -6 dB to 60 dB
4. 42 white noise signals
5. 54 rising and falling chirps, 40-dB SNR, moving from 500 Hz to 2 kHz (or vice versa) at 27 rates from 10 to 1000 octaves per second.

The target value for the LMS procedure was $y = +1$ for the first two classes and $y = 0$ for the last three. Input signals were classified as having tonal onsets if and only if $y > 0.5$. Because the total number of signals (226) was so much larger than the number of LMS weights (4), generalization was not an issue, and the same data were used for both training and testing. Since the training examples were intentionally selected to present a challenging problem, the resulting error rates were high. However, they provide a clear and useful measure of relative performance.

The results of four experiments using these signals are summarized in the table below. In the first three experiments, the adjacent-channel input Δt_k was not used. In the first experiment, with no chirp signals, the error rate was 14.8%. Most of these errors were due either to missing weak onsets or committing false alarms on strong, short-duration tone bursts. In the second experiment, the chirps were added without changing the weights. Since no signals with onsets were added, the miss rate stayed the same. However, all 54 chirps triggered false alarms, and the overall error rate more than doubled. In the third experiment, the weights were adapted in an attempt to reject the chirps. The false alarm rate did fall from 57.5% to 23.3%; however, the miss rate rose to an unacceptable 43.8%. In the fourth experiment, the adjacent-channel input was provided. Although the weaker onset signals were still missed, the false-alarm rate was dramatically decreased, and the overall 18.5% error rate was close to the original 14.8% error rate.

Expt.	Description	Miss	FA	Error
1	No chirps	9.4	22.7	14.8
2	Chirps added	9.4	57.5	36.1
3	Adapt	43.8	23.3	32.4
4	Adapt, inhibit	20.8	16.7	18.5

Additional insight can be obtained by seeing how the onset response y varies with the chirp rate. The upper pair of curves in Fig. 3 shows the response to chirps in the first experiment, where there was no inhibition and no chirps in the training data. In general, the response decreases monotonically with chirp rate over the 10 to 1000 octave-per-second range investigated, but even the smallest responses are above the 0.5 threshold. The middle pair of curves shows that this response can be reduced by training, but that about half of the chirps are still above threshold. The lower curves show that when adjacent-channel inhibition is included, the response for some rapidly falling chirps actually increases; however, the response to the more troublesome and more numerous slowly-changing chirps are very effectively suppressed.

5 Conclusions and Comments

We have shown that adjacent-channel inhibition can be quite effective in reducing the response of a tonal onset detector to frequency-swept tones, particularly for relatively slowly changing chirps. Since the differences in onset times Δt_k are small for rapidly changing chirps but increase with increasing channel separation, even better performance could no doubt be obtained by including inputs from more distant channels as well. Our onset-detector architecture exploited normalization to cope with dynamic range problems and logarithmic features to allow a linear combiner to behave conjunctively. Both of these techniques encounter difficulties at low signal levels. It would be interesting to see if alternative approaches, such as the use of AGC or more complicated neural networks, could provide further improvement.

Acknowledgements

This work was supported by the National Science Foundation under NSF Grant No. IRI-9214233, and is an extension of the M.S. project by the first author [7]. We also appreciate the interest and feedback we have received from Richard F. Lyon at Apple Computer, Inc., and Malcolm Slaney at Interval Research Corp.

References

[1] Basséville, M., and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application.* (Prentice Hall, Englewood Cliffs, NJ, 1993).

[2] Brown, G. J., "Computational auditory scene analysis: A representational approach," PhD dissertation, Department of Computer Science, University of Sheffield, Sheffield, England, UK (September, 1992).

[3] Glasberg, B. R., and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research,* vol. 47, pp. 103-138 (1990).

[4] Mellinger, D. K., "Event formation and separation of musical sound," Report No. STAN-M-77, Center for Computer Research in Music and Acoustics, Department of Music, Stanford University, Stanford, CA (December 1991).

[5] Patterson, R. D., K. Kobinson, J. Holdsworth, D. McKeown, C. Zhang and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*(Y. Cazals, L. Demany and K. Horner, Eds), pp. 429-446 (Pergammon Press, Oxford, England, 1992).

[6] Rhode, W. S., and P. H. Smith, "Physiological studies on neurons in the dorsal cochlear nucleus of cat," *Journal of Neurophysiology,* vol. 56, pp. 287-307 (August, 1986).

[7] Shahwan, T. W., "An adaptive procedure for the optimization of an acoustic onset detector," Technical Report No. 8, NSF Grant No. IRI-9214233, Department of Electrical Engineering, San Jose State University, San Jose, CA (May, 1994).

[8] Slaney, M., "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Technical Report No. 35, Advanced Technology Group, Apple Computer, Inc., Cupertino, CA (1993).

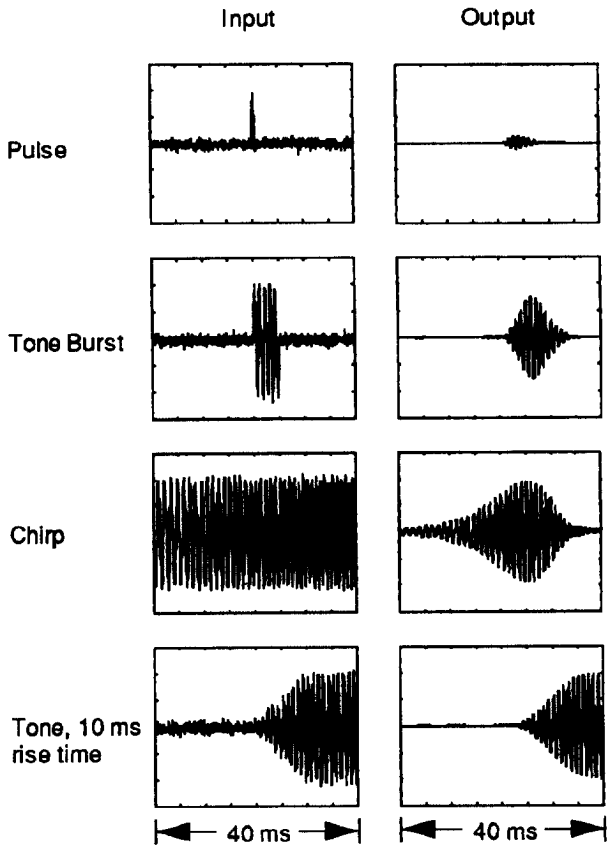

Fig. 1 Band-pass filter responses to three signals without and one signal with a tonal onset.

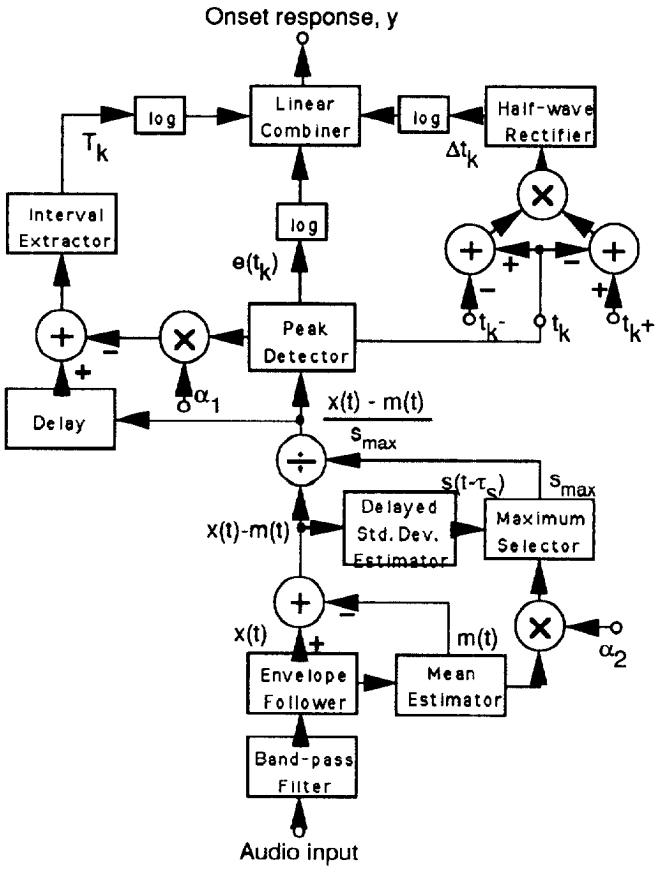

Fig. 2 Block diagram of the acoustic onset detector

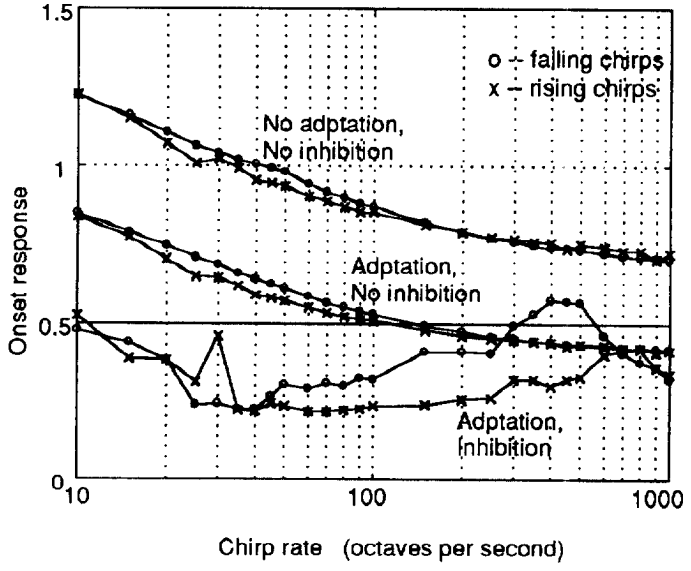

Fig. 3 Response of three different onset detectors to chirp inputs. The desired response is a value less than 0.5.