Contributed article

Modeling the precedence effect for speech using the gamma filter

Odelia Schwartz\textsuperscript{a}, John G. Harris\textsuperscript{b,*}, Jose C. Principe\textsuperscript{b}

\textsuperscript{a}University of Florida, Computer Science and Engineering, Gainesville, FL 32611-6300, USA
\textsuperscript{b}University of Florida, Electrical and Computer Engineering, Gainesville, FL 32611-6300, USA

Received 31 July 1996; revised 3 March 1998; accepted 3 March 1998

Abstract

The ability of the human auditory system to localize the direction of a sound source in a reverberant environment extends from simple clicks to complex speech signals. The perceived location of the auditory event is dominated by the direct sound, for small enough time intervals between the direct sound and reflection—a phenomenon known as the precedence effect. In this paper we present a computer simulation of the precedence effect for speech, implemented using Matlab and tested with speech signals from the TIMIT database. The model is based on the biological assumption that the rate of change of a signal (onset) is critical in triggering the precedence effect. The model demonstrates the precedence effect on speech signals, as opposed to prior models that have only been tested on clicks. In addition, a novel onset enhancement method is described and implemented using the gamma filter, a new class of linear systems for adaptive signal processing. This method can either use fixed coefficients and obtain similar results to prior methods of onset enhancement, or use an adaptive framework. The adaptive framework is promising in its ability to enhance onsets and to reduce the steady state portions of the signal. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Precedence effect; Reverberant; Sound localization; Speech; Complex signals; Gamma filter; Onset enhancement; Auditory modeling

1. Introduction

The human auditory system is capable of localizing a sound source, despite reflections off walls, ceilings and other interfering objects. The perceived location of the auditory event is dominated by the direct sound, for small enough time intervals between the direct sound and reflection—a phenomenon known as the precedence effect (Wallach et al., 1949; Haas, 1951; Blauert, 1983; Blauert, 1996; Zurek, 1987). For time differences that exceed a so-called echo threshold, we perceive two distinct auditory events corresponding both to the direct sound and to the reflection. Psychoacoustic experiments demonstrate that the precedence effect is not limited to simple click stimuli but extends to complex signals such as music or speech (Wallach et al., 1949; Haas, 1951). Typically, most theoretical studies and implementations of the precedence effect have focused on synthetic clicks because of their relative simplicity (Lindenmann, 1986a; Lindenmann, 1986b). With more realistic signals such as speech, the direct sound and reflection overlap in time, creating a complex and variable spectrum. A precedence effect model for speech is of interest theoretically, as it could demonstrate the plausibility of certain biological theories regarding the precedence effect on complex signals. This comprises part of our more general understanding of how complex signals arriving at our ears develop into a sensible auditory scene (Bregman, 1993). The model could also be utilized for practical engineering applications, such as a preprocessor to a more robust speech recognition system.

Lindenmann (Lindenmann, 1986a; Lindenmann, 1986b) demonstrates a precedence effect model for clicks as an extension to the standard cross-correlation model for horizontal localization (Jeffress, 1947; Colburn and Durlach, 1978). The primary cross-correlation peak initiates an inhibition mechanism, that prevents subsequent peaks of the cross-correlation function for a time period that depends on the echo threshold. As such, reflections that follow shortly after the direct sound are inhibited and their directional information is disregarded. The model lacks an explicit onset detection mechanism that could provide a more robust measure of when to trigger the inhibition mechanism for complex signals.

The role of onsets as a trigger for the precedence effect is established both by psychoacoustic experiments and by signal processing theory. Already in the middle of the century, Wallach et al. (1949) distinguish between sounds of a transient nature (clicks, stringed instruments, speech) that
demonstrate the precedence effect well, and between steady tones that do not exhibit precedence. Moreover, more recent psychoacoustic experiments demonstrate that the onset of a signal is more dominant than its steady state portion, in localization in reverberant environments (Rakerd and Hartman, 1985; Rakerd and Hartman, 1986; Hartman and Rakerd, 1989). From the signal processing perspective, onsets are defined as a significant increase in signal energy. Thus, onsets enable us to isolate portions of the signal that correspond mostly to the direct sound or mostly to the reflection. This is particularly crucial for the localization of complex signals which temporally overlap.

In this paper, we demonstrate a precedence effect model for speech. The computer simulation attempts to follow the signal processing of the peripheral auditory system and includes some of the following subsystems: a cochlear model, an onset enhancer, a binaural cross-correlation model for horizontal localization and an adaptive inhibition mechanism. As part of the complete model, we describe two onset enhancement mechanisms—the method of Smith (1994) and a novel approach motivated by adaptive signal processing. Finally, we introduce the gamma model as a unifying framework for the above two approaches. In the following sections we discuss the two stages of the precedence model, namely the preprocessing of the input to enhance onsets followed by the cross-correlation inhibition.

2. Preprocessing of signal to enhance onsets

The preprocessing consists of the following stages (as depicted in Fig. 1):

1. Delay and add—we assume that the raw input signal to each ear consists of the summation of the direct signal and a delayed version of the direct signal (corresponding to the reflection). The horizontal locations of the direct sound and reflection are simulated by introducing interaural time differences.

2. Cochlear model—the raw input signal is entered into a cochlear model consisting of a bandpass filter bank (16 bands) and full wave rectification. We use Malcolm Slaney’s implementation of the Patterson and Holdsworth auditory filter bank to model the basilar membrane (Slaney, 1993). This is followed by rectification, since membrane responses can assume both positive and negative values whereas firing rates are positive. The rectification stage is a common technique for modeling the physiology of a population of inner hair cells. Note that more comprehensive models of inner hair cells that include adaptation could be employed (Meddis and Hewitt, 1991). Adaptation reduces the cellular response following the onset of a signal. This could improve the performance of our onset enhancement mechanism in the precedence model, but should not change the nature of our results.

3. Onset enhancement—enhances the onsets for each frequency band produced by the cochlea. The details of the onset operator are discussed in the next subsections.

4. Peaks of slope—considers only the areas in which the onset operator changes maximally. This is in line with psychoacoustic studies that demonstrate that the rate of onset is important in triggering the precedence effect (Rakerd and Hartman, 1986).

In the following subsections, we will analyze the onset operator with respect to a single frequency band. We present two methods for onset enhancement—Smith’s method and an adaptive filter approach. We then describe the gamma filter as a unifying perspective for these two methods.

The equations throughout the paper will usually be described in discrete time, to match our digital implementation. However, we assume that the mechanisms that occur in the brain are continuous. For readability, we will sometimes also point out the continuous model.

2.1. Onset enhancement based on Smith’s method

The onset operator is computed as the difference of two low pass filters on the input signal $x(n)$. The first filter is an
average over recent time, and the second filter an average over a longer span of recent time (Smith, 1994):

$$O(n; \tau_1, \tau_2) = \sum_{k=0}^{n} [\tau_1^{-n} - \tau_2^{-n}]x(k-n)$$  

(1)

where $n$ is the current time and $\tau_1, \tau_2$ are time constants, such that $\tau_1 < \tau_2$.

The equivalent model in continuous time is:

$$O(t; \tau_1, \tau_2) = \left[ \frac{1}{\tau_1} - \frac{1}{\tau_2} \right] x(t)$$  

(2)

For simplicity, we will consider families of such onset operators, in which $\tau_1 = \tau$ and $\tau_2 = r\tau$ and $r > 1$. Biologically, the onset operator is based on the assumption that onset neurons receive the same driving input twice, once excitatorily and the other inhibitorily (Pickles, 1988; Smith, 1994). In this regard, the filter with the shorter time constant $\tau_1$ represents the excitatory input, and the filter with the longer time constant $\tau_2$ represents the inhibitory input. Thus, Smith's model can be viewed as an abstraction of onset cells in the cochlear nucleus.

For a constant input $x(n) = C$ over a long time, the difference filter has the appropriate property of outputing 0 (Smith, 1994; Schwartz, 1996). For a rapid change in the input, the small time constant responds more quickly. This in turn increases the difference of the two low pass filters and thus increases the value of the onset operator. In the frequency domain, the difference of two low pass filters is a bandpass filter. The center frequency and the bandwidth are roughly determined by $\tau$ and $r$, respectively (Schwartz, 1996).

2.2. Onset enhancement using the adaptive filter approach

Traditionally, onset enhancement strategies compute the energy function using fixed low pass filters (Smith, 1994). The adaptive filter approach does not compute energy per se, but rather offers an alternative definition to onsets as changes in the input signal the system cannot predict.

A one-step adaptive predictor attempts to predict the current value of the input, based on its previous values. The desired signal $d(n)$ is equal to the input signal, $x(n)$. A delayed version of the input signal is entered into an adaptive filter, yielding $y(n)$. The error of the prediction $\epsilon(n)$, is defined as (Widrow and Stearns, 1985):

$$\epsilon(n) = d(n) - y(n)$$  

(3)

The adaptive filter is designed as an FIR filter with adjustable weights. The number of taps $L$ corresponds to the number of past values of the signal, used to predict the current value. The output $y(n)$ of the filter is computed as (Widrow and Stearns, 1985):

$$y(n) = W^TX(n-1)$$  

(4)

where $W$ is a weight vector of size $L$ and $X(n-1)$ is a vector that contains the past $L$ values of the input signal. Therefore,

$$\epsilon(n) = d(n) - W^TX(n-1)$$  

(5)

The weights are adjusted, such that the squared error is minimized. According to the exponential weighted recursive least squares (EWRLS) algorithm, we define the squared error at time $n$ over an exponential window (Clarkson, 1993):

$$J(n) = \alpha^n \sum_{l=-\infty}^{n} (\epsilon(l))^2$$  

(6)

where $0 < \alpha < 1$. The goal then is to adjust the weights such that $J(n)$ is minimized:

$$\nabla_w J(n) = 0$$  

(7)

For purposes of simulation, the weights of the filter are initialized to a zero vector of length $L$, and are updated for each value of $n$ (i.e. for every new speech sample). Following the initialization, the weights are constantly adapting as the stimulus is changing.

We interpret high error readings as an onset, since we cannot predict the current input based on previous input values. Also, higher error levels can only be detected if there is a high signal energy. Therefore, this definition of onset is in effect a combination of signal energy and predictability. The two parameters of interest in the adaptive filter approach are the number of taps and $\alpha$. Increasing the number of taps improves performance but also increases the computational complexity (e.g. the computation of the weights for each time sample requires $O(L^2)$ multiplications). As $\alpha$ increases, the algorithm is more stable but the filter becomes less sensitive to changes in the error. A more detailed account about EWRLS and its computational requirements can be found in Clarkson (1993).

2.3. The gamma filter for onset enhancement

The gamma filter has been recently introduced as a new class of linear systems for adaptive signal processing (Principe et al., 1993). The key idea is to substitute the ideal delay operator in discrete time (or the integrator operator in continuous time) by a dispersive delay element (or a leaky integrator in continuous time). The impulse response of the discrete gamma filter with $K$ stages is given by (Principe et al., 1993):

$$h(n) = \sum_{k=1}^{K} w_k g_k(n)$$  

(8)

where

$$g_k(n) = \binom{n-1}{k-1} \mu^k (1-\mu)^{n-k}$$  

(9)

$k = 1, \ldots, K (1 > \mu > 0)$ and $n - k > 0$. Note that the
of the problem. In the adaptive approach, the filter coefficients are always changing to best predict the signal, and those parts of the signal that are not well predicted are classified as onsets.

2.4. Results: preprocessing of inputs to enhance onsets using the gamma filter

We demonstrate the gamma filter for onset enhancement using: (1) the fixed coefficients; and (2) the adaptive framework. The input to the simulation is the waveform for the word ‘teeth’, taken from the TIMIT database. The fixed coefficients obtain results that are very similar to the method of Smith (1994) the difference of two low pass filters [see corresponding results in (Schwartz, 1996)].

For simplicity, we do not introduce echoes to the speech signals in these examples. For each input signal, the figures depict the time domain representation, a single cochlear band, the fixed coefficient method and the adaptive filter method. The number of taps and $\alpha$ for the EWRLS are taken as 40 and 0.993, respectively. The value of $\mu$ is chosen as 0.005 in the fixed coefficient method, to provide smoothing of the signal and thus obtain results that are similar to Smith’s method. In the adaptive approach, $\mu$ is chosen as 0.8, to optimize the prediction ability of the model rather than provide smoothing. For more details on how to choose these parameters one can refer to (Schwartz, 1996).

Fig. 3 contains the second band of the word ‘teeth’ (4000–6000 Hz) as an input to the onset enhancer. The change corresponding to ‘t’ at 40.5 ms is the abrupt onset that we need to detect, since the other changes in the signal are too gradual to be considered an onset. The adaptive filter method is more effective in enhancing onsets and eliminating other less significant changes in the signals. In this example, the ratio between the onset and the less significant change at 220 ms is 20 in the adaptive filter model, which is approximately four times higher than the corresponding ratio using the fixed coefficients.

Fig. 4 contains the sixth band of the word ‘teeth’ (1500–3000 Hz) as input. In this example, the adaptive filter emphasizes the onsets corresponding to ‘t’ and ‘th’ and reduces the changes that are repetitive and rather...
Fig. 3. Onset enhancement with the second band of the word ‘teeth’ (4000–6000 Hz) used as input to the analysis: (a) waveform of the word ‘teeth’; (b) rectified band: the word ‘teeth’ is passed through a bandpass filter bank, and the absolute values of a single band (4000–6000 Hz) are displayed; (c) Smith’s (Smith, 1994) onset operator on the rectified band, using a gamma filter with fixed coefficients; (d) adaptive onset operator on the rectified band, obtained by adjusting the coefficients of the gamma filter at every time sample to minimize the prediction error.

predictable in nature. In such situations, methods that are not adaptive perform poorly and might emphasize insignificant onsets. We conclude the the adaptive method is promising in its ability to enhance onsets and reduce the steady state portions of the signal. The onset operator for the complete precedence model can be computed using either fixed coefficients [similar to the method of Smith (1994)] or using the adaptive approach. For precedence, we are primarily interested in obtaining localization information from the abrupt onsets and reducing less significant changes in the signal. Although both onset methods can be utilized, the adaptive filter approach seems more appropriate for this particular task.

Fig. 5 depicts the four stages of preprocessing of the model for a single ear, with the word ‘teeth’. It includes: (a) the delay and add to produce 10 ms echo; (b) a single rectified band produced by the cochlea; (c) the adaptive gamma filter onset operator; and (d) the peaks of the slope of the onset operator. Note that it is difficult to differentiate between portions of the signal that belong to the direct sound and portions that belong to the 10 ms echo. Nevertheless, we expect that the initial peaks in (d) belong to the direct sound. The peaks thereafter might belong to the echo but they will be inhibited. This process is described in detail in the following section.

Fig. 4. Onset enhancement with the sixth band of the word ‘teeth’ (1500–3000 Hz) used as input to the analysis: (a) waveform of the word ‘teeth’; (b) rectified band: the word ‘teeth’ is passed through a bandpass filter bank, and the absolute values of a single band (1500–3000 Hz) are displayed; (c) Smith’s (Smith, 1994) onset operator on the rectified band, using a gamma filter with fixed coefficients; (d) adaptive onset operator on the rectified band, obtained by adjusting the coefficients of the gamma filter at every time sample to minimize the prediction error.

Fig. 6. Cross-correlation model with adaptive inhibition.
3. Cross-correlation model with adaptive inhibition

The precedence model is presented in Fig. 6. The preprocessed inputs, \( l(n) \) and \( r(n) \), are fed into two delay lines. The onset operator is computed using the adaptive filter approach. The cross-correlation value at each tap is computed as the minimum of the values coming in from the left and right ears. This is to ensure that an onset will be detected only if the signal values at both ears are sufficiently high. If the value of the winning tap, \( W_i \), is high enough to signify an onset, it will encode the interaural delay of the sound source. This interaural delay value is used to indicate the horizontal location of the stimulus, such that a value of zero represents the midline and large values indicate sources toward one side.

The adaptive inhibition mechanism sets a threshold as to when a cross-correlation peak should be considered an onset. The inhibition function \( I \) is defined as follows:

\[
I(n) = kV_{\text{ons}} \tau_f ^{- (n-n_0)} + V_{\text{ons}} \tau_s ^{- (n-n_0)} + V_{\text{thr}}
\]  

(15)

where \( V_{\text{ons}} \) is the most current onset value, \( n_0 \) is the time that \( V_{\text{ons}} \) occurred and \( n \) is the current time. The equivalent definition in continuous time is:

\[
I(t) = kV_{\text{ons}} e^{- (t-t_0)/\tau_f} + V_{\text{ons}} e^{- (t-t_0)/\tau_s} + V_{\text{thr}}
\]  

(16)

The value of \( I \) is initially set to some minimal threshold, \( V_{\text{thr}} \). If an onset is detected, \( I \) rises by a constant \( k + 1 \) times \( V_{\text{ons}} \) and then decreases exponentially based on a fast and a slow time constant, \( \tau_f \) and \( \tau_s \), respectively. The fast time constant models the precedence effect. It is set such that after echo threshold time the function returns to the value of the onset. The slow time constant causes the function to return to the default threshold if no onset is detected for a long time.

3.1. Results: cross-correlation model with adaptive inhibition

The model consists of 16 delays in the delay lines, representing 17 horizontal locations of the sound source. The time delay between each tap is .0625 milliseconds, based on the observation that the total time for sound to travel from one ear to the other is at most 1 ms (Jeffress, 1947). The fast time constant, \( \tau_f \), is set to 50 ms corresponding to the echo threshold for speech (Jeffress, 1947). The slow time constant, \( \tau_s \), is set to 250 ms. The value of \( k \) is chosen as 20 and \( V_{\text{thr}} \) as 500 (equivalent to about 4% of the highest onset values that were experienced in this study). In the following examples the direct sound is presented to the right ear 0.5 ms after to the left, corresponding to a correlation at tap 13. The echo is presented simultaneously to both ears, corresponding to a correlation at tap 9.

![Graph](image)

Fig. 7. Precedence effect on the second band of the word 'teeth' (4000–6000 Hz) with 10 ms echo. A single auditory event is perceived, in the direction of the direct sound at 42.4 and 47.4 ms.
Figs. 7–9 demonstrate the results of the cross-correlation inhibition model. Fig. 7 depicts the winners and the inhibition function for the second band of the word ‘teeth’ (4000–6000 Hz) with a 10 ms echo, i.e. below the echo threshold. The winning tap at each time step, is the tap in which the minimum of the onset information from the two ears is highest. This winning tap is considered an onset, only if the value of the tap is higher than the inhibition function. The dotted line represents the inhibition function. In this example, a single auditory event is perceived, corresponding to the direct sound. Since the echo delay is within the echo threshold, the perception of the direct sound and the suppression of the directional information thereafter is in line with our expectations from psychoacoustics. The directional information is first sampled at 42.5 milliseconds, with the maximum at tap number 12 (i.e. one tap off). However, a second higher peak of the direct sound is able to push through the inhibition at 47.4 ms. The maximum this time corresponds to tap number 13 (i.e. the correct tap). As such, if the original default threshold is too low, the system is able to adjust itself and correct this localization error.

Fig. 8 contains the second band of ‘teeth’ (4000–6000 Hz) with a 100 ms echo, i.e. above the echo threshold. This time, two auditory events are perceived. The first event occurs at 42.4 and 47.4 ms, and corresponds to the direct sound at tap number 13. The second event occurs at 145.6 ms and corresponds to the echo at tap number 8 (i.e. the localization is off by one tap). Since the echo delay is above the echo threshold, the onset of the echo is not suppressed and two distinct events are determined—as expected by psychoacoustic theory.

Fig. 9 contains the sixth band of the word ‘teeth’ (1500–3000 Hz) with a 10 ms echo, i.e. below the echo threshold. Note that in this band there are two peak areas of the input. If we detect these two onsets early enough, we should be able to extract the directional information of the direct sound from both onsets. The first onset is localized at 42.1 ms corresponding to tap number 13 and the second at 217.6 ms corresponding to tap number 14. Thus, both onsets localize with respect to the direct sound and the echo is suppressed.

The model has been tested on a variety of words by different speakers. A more comprehensive review of the model and its results is provided in (Schwartz, 1996).

4. Conclusions

The results signify that the model is able to localize a
Fig. 9. Precedence effect on the sixth band of the word ‘teeth’ (1500–3000 Hz) with 10 ms echo. Two auditory events are detected at 42.4 and 217.6250 ms. Both locations correspond to the direct sound and not to the echo.

band of speech containing abrupt onsets quite accurately. Moreover, the model is in line with our expectations from psychoacoustics. For small time differences between the direct sound and echo, a single auditory event is perceived in the direction of the direct sound. For time differences that exceed the echo threshold, two distinct auditory events are perceived. Therefore, this model demonstrates the plausibility of onsets as a low-level cue for the precedence effect.

The gamma filter model offers a unifying approach to onset enhancement, that can either use fixed coefficients and obtain similar results to Smith’s method or use an adaptive framework. We demonstrate that the adaptive framework for onset enhancement is more effective than the fixed coefficient method in reducing the steady state portions of the signal. This offers an advantage of the adaptive method over traditional methods, to applications such as the precedence effect. The adaptive method will emphasize abrupt onsets that provide more precise timing information and will eliminate less significant onsets. It is plausible that the brain performs some form of prediction for onset enhancement. However, we do not know the actual circuitry to this prediction process. In particular, our implementation of the prediction using EWRLS was chosen based on efficiency and not on biological plausibility.

This work is a first step in deciphering the relation between precedence, speech and onset enhancement. The precedence model can serve as a tool to understand the precedence effect for complex signals and should be advantageous to engineering applications such as preprocessors for speech recognition systems. The gamma model offers a novel approach to onset enhancement that deserves further investigation. Both lines of research provide a promising platform for the development of more sophisticated models in auditory scene analysis, that involve real speech signals.

Acknowledgements

The authors would like to acknowledge Office of Naval Research contract no. N00014-94-1-0858.

References


