# A cascade autocorrelation model of pitch perception

Emili Balaguer-Ballester[a] and Susan L. Denham[b]
*Centre for Theoretical and Computational Neuroscience, University of Plymouth,*
*Devon PL4 8AA, United Kingdom*

Ray Meddis[c]
*Department of Psychology, Essex University, Colchester CO4 3SQ, United Kingdom*

Autocorrelation algorithms, in combination with computational models of the auditory periphery, have been successfully used to predict the pitch of a wide range of complex stimuli. However, new stimuli are frequently offered as counterexamples to the viability of this approach. This study addresses the issue of whether in the light of these challenges the predictive power of autocorrelation can be preserved by changes to the peripheral model and the computational algorithm. An existing model is extended by the addition of a low-pass filter of the summary integration of the individual within-channel autocorrelations. Other recent developments are also incorporated, including nonlinear processing on the basilar membrane and the use of integration time constants that are proportional to the autocorrelation lags. The modified and extended model predicts with reasonable success the pitches of a range of stimuli that have proved problematic for earlier implementations of the autocorrelation principle. The evaluation stimuli include short tone sequences, click trains consisting of alternating interclick intervals, click trains consisting of mixtures of regular and irregular intervals, shuffled click trains, and transposed tones.
© *2008 Acoustical Society of America.* [DOI: 10.1121/1.2967829]

## I. INTRODUCTION

Autocorrelation algorithms have given a useful account of a wide range of auditory pitch phenomena (e.g., Licklider, 1951; Slaney and Lyon, 1990; Meddis and Hewitt, 1991; Cariani and Delgutte, 1996a, 1996b). Their success has generated a large number of studies designed to test the limits of these theories. Some studies have identified weaknesses in the detail of early formulations but have also suggested simple remedies that leave the basic periodicity principles intact (e.g., Wiegrebe, 2001; Pressnitzer *et al.*, 2001; Bernstein and Oxenham, 2005) while other studies, to be discussed below, remain as challenges to the underlying idea of the use of autocorrelation. This report will argue that autocorrelation can continue to provide a satisfactory account if an existing model (Meddis and O'Mard, 1997) is changed to incorporate a more sophisticated model of the auditory periphery and is expanded by adding a stage with a longer integration time constant than used hitherto. This model combines Licklider's (1951) original idea of an autocorrelation function (ACF) of auditory nerve (AN) activity with the idea of generating a summary autocorrelation function (SACF) based on an aggregation of the individual ACFs across all fibers.

Licklider (1951) originally suggested using a running autocorrelation with a narrow temporal window shifted along the time axis. This window is represented by a low-pass filter having a time constant of only 2 or 3 ms. Most explorations have used time constants in this region. However, Wiegrebe (2001) provided evidence that multiple time constants are required, each proportional to the autocorrelation lag being evaluated. A value of twice the lag appeared to give a satisfactory account of the audibility of the oscillations in pitch strength in response to repeated-interval noises. According to this rule, time constants between 40 and 4 ms would be appropriate for pitches in the range 50–500 Hz. This suggestion will be adopted along with his other suggestion that some stimuli require even longer integration times. The proposed solution is to extend the existing model by passing the running SACF through a low-pass filter to produce a new function, which will be given the acronym LP-SACF.

The necessity for longer time scales has been evident since the publication of a demonstration by Hall and Peters (1981) that harmonically related pure tones can generate a pitch percept even when presented nonsimultaneously. They presented three harmonically related pure tones in rapid succession and obtained a pitch sensation consistent with that obtained when the three were presented simultaneously, but only when noise was present. The offset of the first and the onset of the third tone were separated by 60 ms and yet pitch integration still occurred. This demonstration will be used in the first evaluation of the extended model.

A new class of pitch stimuli has recently emerged, which appears to offer a more fundamental challenge to autocorrelation as a general explanation. These stimuli are high-pass filtered irregular click trains in which the distortion products have been masked with background noise; nevertheless, they can be ordered on a pitch-height scale (Kaernbach and Demany, 1998; Kaernbach and Bering, 2001). The responses of listeners to these stimuli cannot be predicted by

---

[a] Electronic mail: emili.balaguer-ballester@plymouth.ac.uk
[b] Electronic mail: s.denham@plymouth.ac.uk
[c] Electronic mail: rmeddis@essex.ac.uk

any current autocorrelation algorithms and this has resulted in calls to search beyond autocorrelation for alternative theories. However, it will be shown that a radical change of direction is not required and that the extended model can be used to predict these data.

The use of short time constants to compute the running ACFs has led to an acknowledged weakness when the stimulus contains fluctuations whose durations are longer than the time constant used in the calculations. The work-around used in earlier studies to deal with this problem involved inspecting the SACF at the end of a complete pitch period (e.g., Meddis and Hewitt, 1991). While this worked with periodic stimuli, it is clearly only a short-term expedient and not an elegant long-term solution to the problem. The challenge of aperiodic click stimuli makes the problem more pressing because it is not possible to specify an ideal time at which to sample the SACF. The new extended model removes the problem by using a longer integration window, thus stabilizing the output function.

## II. THE MODEL

The original model has been fully described and extensively evaluated in previous publications (Meddis and Hewitt, 1991; Meddis and O'Mard, 1997). It consists of three stages: (1) an auditory model to simulate AN spiking probabilities, (2) an autocorrelation algorithm, and (3) an algorithm for making predictions of the pitches heard by listeners and their pitch discrimination abilities.

### A. Auditory model

An auditory model simulates the generation of spikes in AN fibers by simulating the processing stages between stimulus reception and the inner hair cell (IHC) AN synapse. Successive processing stages are (1) stimulus input, (2) stapes response, (3) multichannel basilar membrane (BM) response, (4) IHC receptor potential, (5) IHC/AN synapse transmitter vesicle release rates, and (6) AN spiking probabilities. Computational models of the auditory periphery continue to evolve but the version used here has been fully described in a recent study of AN first spike latencies (Meddis, 2006). All formulas and parameters are given in the Appendix to that report; and used but unchanged here; except the transmitter release permeability in the IHC presynapse, which takes the value indicated in the work of Sumner et al. (2002). The input to the model is the stimulus used in the corresponding psychophysical experiment. The output is a stream of AN spiking probabilities. The implementation used here has 60 channels with best frequencies (BFs) ranging from 100 to 10 000 Hz along a logarithmic scale. The auditory model was evaluated at an integration period of $1/44 100$ s, except for Evaluation 4, which used a period one-quarter of that duration.

Figure 1(a) shows the response of the auditory model to a 100 ms harmonic complex consisting of the third to sixth harmonics of 100 Hz. The lower panel in Fig. 1(a) shows the AN spiking probabilities, $p(t,k)$, arranged as a channel $x$ time matrix. This is the input to the next stage, the autocorrelation algorithm.
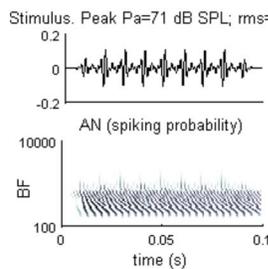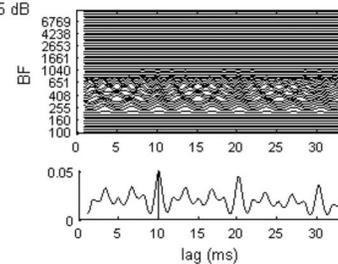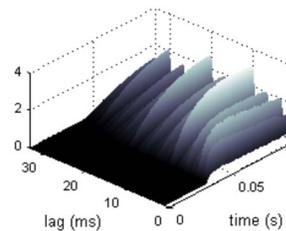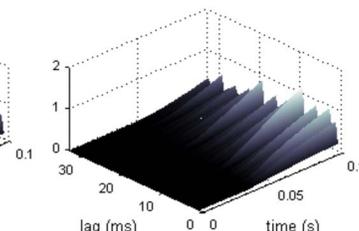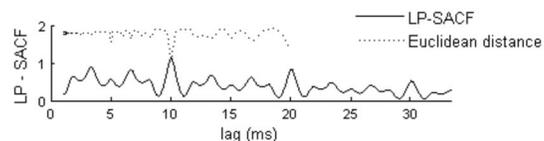


FIG. 1. (Color online) Response of the model to a 100 ms complex stimulus consisting of harmonics 3, 4, 5, and 6 of a 100 Hz fundamental. (A) Stimulus waveform (upper panel) and AN spiking probabilities generated by an auditory model (lower panel). (B) Multichannel ACFs (upper panel) and SACF (lower panel) after 100 ms. (C) Evolution of the SACF over time. (D) Evolution of the LP-SACF over time. (E) LP-SACF at the end of the stimulus (continuous line). The Euclidean-distance function is shown as a dotted line. The minimum of this function (10 ms) predicts the pitch.

### B. Autocorrelation algorithm

The algorithm for computing the SACF function is implemented as described in the work of Meddis and O'Mard (1997). The computation of the individual running autocorrelations in each channel, $h(t,l,k)$, is based on the spike probabilities, $p(t,k)$,

$$h(t,l,k) = p(t,k) \cdot p(t-l,k) \cdot \frac{\Delta t}{\tau(l)} + h(t-\Delta t,l,k) \cdot e^{-\Delta t/\tau(l)},$$

(1)

where $t$ is time, $l$ is the autocorrelation lag, $\Delta t$ is the sampling interval, and $k$ is the channel number. $\tau(l)$ is a time constant specific to a given lag and is set to $2l$ (Wiegrebe, 2001). This equation is the same as Eq. (1) in the work of Meddis and O'Mard (1997) but expressed as a recursive function. 191 lags were used in this study. They were linearly spaced between $1/30$ and $1/1000$ s and were the same for all BFs.

The ACFs in Eq. (1) are summed across channels to create the SACF

$$S(t,l) = \sum_{k=1}^{N} h(t,l,k).$$

(2)

Figure [1(b)] shows the running ACFs and the running SACF as they appear at the end of the stimulus presentation while Fig. [1(c)] shows how the SACF changes over time.

At this point, a new stage is added. The running SACF is passed through a low-pass filter implemented as an exponentially decaying average,

$$P(t,l) = S(t,l) + P(t - \Delta t, l) \cdot e^{-\Delta t/\lambda}, \qquad (3)$$

where $\lambda$ is the time constant of the filter. $P(t,l)$ is the LP-SACF. Figure [1(d)] shows how the LP-SACF changes over time. $\lambda$ was set to 120 ms throughout the study except for evaluation 4 which was an order of magnitude larger. A detailed study of the optimum value was not attempted. This value was chosen simply because it was the minimum time constant large enough to yield acceptable results for the data of Hall and Peters (1981) described in the first evaluation below. It is consistent with the minimum estimate of integration period of 210 ms proposed by Grose *et al.* (2002) given that an exponential decay function with a time constant of 120 ms will have decayed to 17% of its starting value after 210 ms.

Since the original model was published, a number of detailed improvements have been suggested by different authors. We have adopted Wiegrebe's suggestion that the time constant of integration of the SACF should be linked to the individual lag. On the other hand, Pressnitzer *et al.* (2001) have suggested a weighting function that reduces the magnitude of the SACF as a function of lag and Bernstein and Oxenham (2005) and Denham (2005) have suggested schemes for omitting some lags from the individual channel ACFs. Notwithstanding the agreed merits of these two suggestions, they have not been adopted here in order to simplify the discussion of the contribution of the other changes introduced in this study and because they do not alter the conclusions to be drawn in this report.

### C. Pitch predictions

The simplest method for predicting pitch uses the reciprocal of the lag associated with the highest peak in the SACF, and the same argument applies to the new LP-SACF. This often leads to unambiguous and accurate predictions. However, there are some situations when it cannot be used; these include small pitch shifts caused by mistuned harmonics and predictions of the ability to discriminate the pitch of two stimuli (see Meddis and Hewitt, 1991 and Meddis and O'Mard, 1997 for details). On these occasions, it is better to follow the method used in the experimental procedure. This typically involves searching for a best pitch match by adjusting the pitch of a second tone until it matches with the pitch of a reference tone. This is the approach adopted here.

To implement the matching procedure, the LP-SACF is generated for a range of periodic comparison stimuli. These are then compared, one by one, with the LP-SACF of the test stimulus used in the study to be simulated. This is achieved by computing the Euclidean distance between the LP-SACF of the comparison tone and the LP-SACF of the test stimulus. The fundamental frequency of the comparison stimulus
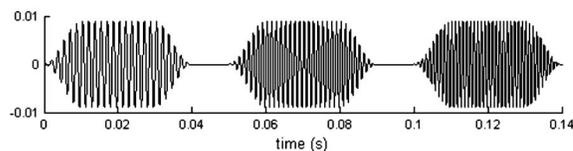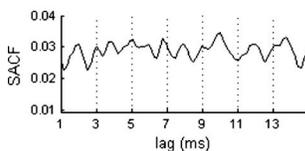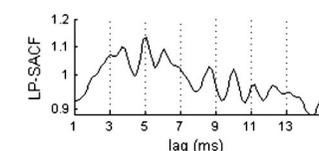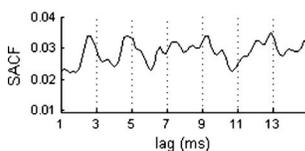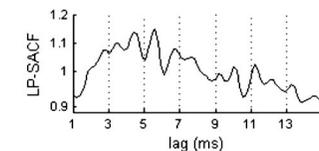


FIG. 2. Predicted virtual pitch evoked by successive tones. (A) The stimulus waveform for a sequence of 40 ms pure tones of 600, 800, and 1000 Hz separated by a silent period of 10 ms (Hall and Peters, 1981), before the addition of noise. (B) The average final SACF in response to five presentations of the 600–800–1000 Hz tone sequence in the presence of noise. (C) The average final LP-SACF in response to five presentations of the 600–800–1000 Hz tone sequence in the presence of noise. (D) The average final SACF in response to five presentations of a 720–900–1080 Hz tone sequence in the presence of noise. (E) The average final LP-SACF in response to five presentations of a 720–900–1080 Hz tone sequence in the presence of noise.

with the smallest associated Euclidean distance is then chosen as the predicted "best-match" pitch [Figure [1(e)].[1]

## III. EVALUATIONS

### A. Evaluation 1

The need for a window of pitch integration longer than 3 ms is clear from a demonstration by Hall and Peters (1981). They showed that nonsimultaneous tones presented in background noise can combine to create a virtual pitch that can be matched to other pitch-evoking stimuli. Other more recent studies using nonsimultaneous components such as temporal fringes (Carlyon, 1996; Micheyl and Carlyon, 1998), mistuned delayed harmonics (Ciocca and Darwin, 1999; Gockel *et al.*, 2005), and stimuli with interpolated silences (Plack and White, 2000) have provided further evidence that substantial temporal integration is required to explain how pitch perception aggregates stimulus information across silent intervals.

Hall and Peters' (1981) stimulus is illustrated in Fig. [2(a)]. It consists of three tones played successively against a white noise background. Each 50 dB sound pressure level (SPL) tone lasts 40 ms and is separated from the following tone by a gap of 10 ms. In this particular example, the frequencies used are 600, 800, and 1000 Hz. The noise was white noise at a level of 65 dB SPL rms. The authors specify that the tones were 6 dB *above* masked threshold. The individual tones of the sequence were still audible, but listeners

Balaguer-Ballester *et al.*: Cascade autocorrelation model

## A. Final LP-SACF 600-800-1000 Hz in quiet
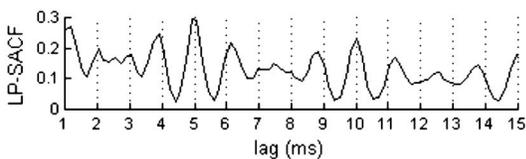


## B. Final LP-SACF 720-900-1080 Hz in quiet
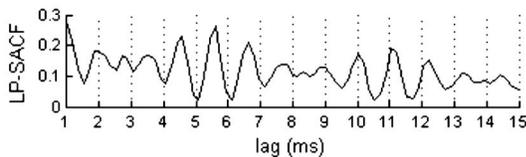


FIG. 3. LP-SACF response to short tone sequences in quiet. (A) Final LP-SACF for the 600–800–1000 Hz tone sequence in quiet. (B) Final LP-SACF for the 720–900–1080 Hz tone sequence in quiet.

were instructed to attend to the lowest of the perceived pitches. It was found that listeners matched the lowest perceived pitch to a 200 Hz pure tone. To further illustrate the effect, Hall and Peters (1981) compared this stimulus with another version where the component frequencies were 720, 900, and 1080 Hz. This stimulus was matched by their subjects to a pure tone with a lower pitch (180 Hz) even though the component frequencies were higher in the second stimulus.

The waveform in Fig. 2(a) shows the stimulus before the noise was added. The SACF for the stimuli in the presence of the noise [Figs. 2(b) and 2(d)] is unable to predict the low virtual pitch. Figures 2(c) and 2(e) show the LP-SACFs for the two stimuli in the presence of the noise. These are both based on an average of five trials to reduce variability created by the noise background. Due to the loudness of the added noise, we based the predictions in this evaluation on the highest LP-SACF peak rather than on the Euclidean metrics. The LP-SACF for the 600–800–100 Hz stimulus shows a broad peak around a lag of 5 ms corresponding to a pitch in the region of 200 Hz. An examination of the fine structure of the LP-SACFs reveals a shift to the right (longer lags) in the case of the 720–900–1080 Hz stimulus, indicating a lower predicted pitch around 5.6 ms (180 Hz).

The response of the computer model shows that low-pass-filtering the SACF to form the LP-SACF was successful in integrating periodicity information across the whole stimulus. This was not achieved with the SACF alone because of its short time constants. For example, in the model the time constant associated with the 5 ms lag (corresponding to a periodicity of 200 Hz) is only 10 ms, which is considerably shorter than the time interval between the first and the last tones.

Hall and Peters (1981) found that the low virtual pitch was not heard when the tones were presented in quiet. It is therefore of some interest to examine the model response to the sequence of tones without the background noise. Figure 3 shows the LP-SACF output for the two stimuli presented in quiet. Major peaks are present at lags of 5 and 5.6 ms, respectively. This would appear to predict corresponding pitches of 200 and 180 Hz, pitches that were not reported by

subjects in the experiment. However, in quiet, the LP-SACF contains a series of peaks, potentially indicating a much higher pitch. The listeners in Hall and Peters' (1981) experiment never matched the stimulus to a single pure tone with low pitch but always preferred to match these stimuli to a higher pitch in the region of individual component tones. This suggests caution against the common tendency to identify a single large peak in an ACF with the predicted pitch and indicates the need to take the whole of the function into account.

In pitch experiments it is not unusual to find that subjects listen analytically and base their judgements on individual tone components rather than the total stimulus complex. Normally, subjects are discouraged from matching the pitch of one of the stimulus components. A comparison between the model responses in Figs. 2 and 3 suggests that noise has the effect of reducing the relative size of the more rapid oscillations in the LP-SACF corresponding to the component tones and makes it more likely that a judgement of the lowest perceived pitch will be based on the peaks in the range of the virtual pitch. Hall and Peters (1981) also suggested that the transient amplitude changes (audible when the tones are presented in a quiet background) may trigger some other mechanism that emphasizes the separate nature of the tones and tips the perceptual balance in favor of analytic hearing. Unfortunately, the computer model presented here does not contain any mechanism at this level of sophistication.

In conclusion, the introduction of a low-pass filter to further process the SACF into the LP-SACF was successful in integrating tone sequences presented successively in noise to generate a pitch prediction that agreed with the lowest perceived pitch reported by listeners. However, the demonstration indicated an unsolved problem concerning how the balance of synthetic and analytic listening is established in general for sounds presented in quiet.

### B. Evaluation 2

When a regular click train is played, subjects hear a pitch whose frequency is the reciprocal of the time interval between successive clicks. Kaernbach and Demany (1998) modified this stimulus by placing a single additional click at a random time between each regular click. The stimuli were high-pass filtered (in this case at 6 kHz) to remove all resolved harmonics. The modified stimulus with the interpolated random clicks has only a weak pitch compared to the regular interval train. Indeed, their subjects found it difficult to discriminate between it and a completely random click train with the same overall click rate. The interpolation of random clicks had greatly reduced the sensation of pitch even though regular intervals were present between every second click.

This result is not expected if a prediction is based on an autocorrelation of the click train itself. A simple autocorrelation of the stimulus waveform shows a clear peak at the regular interval for both the simple click train and the one
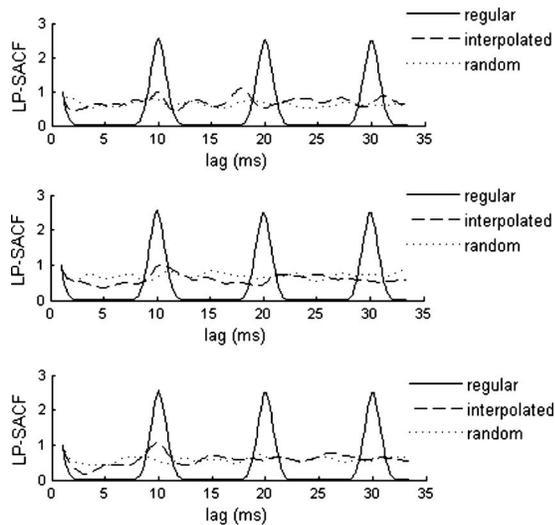
FIG. 4. The effect of the LP-SACF on adding randomly interpolated clicks (dashed line) to a regular click train (solid line). The results can be compared with those for a completely random click train (dotted line). The click trains have been made as explained in the text. The LP-SACFs are shown for three different stimulus samples.

with the interpolated clicks. This approach to autocorrelation would lead to the prediction that both stimuli should have a similarly clear pitch.

However, different predictions are made when the autocorrelation is based on the output of the auditory model. To demonstrate this, the model was evaluated using three different stimuli: regular, interpolated, and random. The click rate in the regular click train was 100 clicks/s. The random click train had a mean rate of 200 clicks/s and was constrained so that any run of three consecutive clicks with above average intervals was deleted and replaced. The interpolated click train contained repeated *ABX* sequences containing a fixed interval ($A+B=10$ ms) in which an interpolated click was randomly placed between each pair of regular clicks. The *AB* pair was then followed by another random interval ($X$) with an average duration of 10 ms. The overall click rate of this stimulus was 150 clicks/s. All clicks were bandpass filtered between 6 and 10 kHz using a cascade of six first-order Butterworth filters. They were presented to the model at a spectrum level of 55 dB SPL in a low-pass filtered white noise background at a spectrum level of 30 dB SPL. The duration of all stimuli was 500 ms.

The responses of the model to these stimuli are shown in Fig. 4. Because these stochastic stimuli change from trial to trial, three sets of results are given. The click train with random intervals produces a LP-SACF (dotted line) with no reliable structure. The regular click train, on the other hand, gives a regular multipeak LP-SACF (continuous line) with the first major peak at 10 ms, as expected. When the interpolated clicks are added, the LP-SACF (dashed line) shows a reliable but small peak at 10 ms but regularly spaced peaks at 20 and 30 ms are typically not present. On visual inspection, the LP-SACF for this interpolated click stimulus looks much more like the LP-SACF for the random click stimulus than that for the regular click train. In summary, the model predicts an unambiguous pitch for the regular click train but

a relatively weak pitch for the interpolated click train. It also suggests that the interpolated click train will be discriminated from the random click train only with difficulty. This prediction is consistent with the reported observation that the interpolated sequence does not sound regular and is perceived by listeners as similar to a random click train. Nevertheless, the interpolated click train does produce a reliable small peak at 10 ms and this is consistent with the observation that, in the long run, listeners are better than chance at identifying it as more tonal than the completely random stimulus (Kaernbach and Bering, 2001).

In summary, the use of an auditory model as the input to the autocorrelation algorithm gives a good account of the main finding that the addition of randomly interpolated clicks degrades the tonality of a regular click train. This contrasts with the predictions of Kaernbach and Demany (1998) based on the autocorrelation of the stimulus waveform where little degradation is expected. Clearly, autocorrelation of the *stimulus waveform* is not a viable predictor of pitch perception in all cases but its shortcomings do not apply when the output of a model of the auditory periphery is used as the input to the autocorrelation calculations.

## C. Evaluation 3

The success of Evaluation 2 raises the question of why the auditory model succeeds when an autocorrelation of the stimulus waveform fails. Pressnitzer *et al.* (2002, 2004) drew attention to the important changes that can take place in the representation of the stimulus as a consequence of nonlinearities inherent in auditory peripheral processing. Specifically, they suggested that the half-wave rectification that takes place in the electrical response of the IHC cell can introduce previously absent spectral components if the stimulus frequency components are not resolved by the peripheral auditory system. To demonstrate this, they applied autocorrelation at different stages in the peripheral processing and showed that changes took place following the half-wave rectification that were critical for distinguishing the pitch characteristics of two carefully chosen click-train stimuli (described below). These two click trains had the same average click rate but different pitches. They were able to show that autocorrelation of the output following half-wave rectification resulted in profiles that could be related to the pitches heard by listeners.

The peripheral model used in the work of Pressnitzer *et al.* (2002) was highly schematic, used long-term autocorrelation, and employed only linear peripheral filters. The model evaluated here uses nonlinear peripheral filtering to represent the response of the BM and has a considerably more complex representation of the generation of the IHC response. It also features short-term autocorrelation computations followed by a low-pass filtering stage and this contrasts with their approach using long-term autocorrelation. It will be useful to demonstrate that the findings of Pressnitzer *et al.* (2002) remain valid with the new model, because their insight is fundamental to understanding the success of the autocorrelation model in explaining the perceived pitch of many high-pass click-train stimuli.

Balaguer-Ballester *et al.*: Cascade autocorrelation model
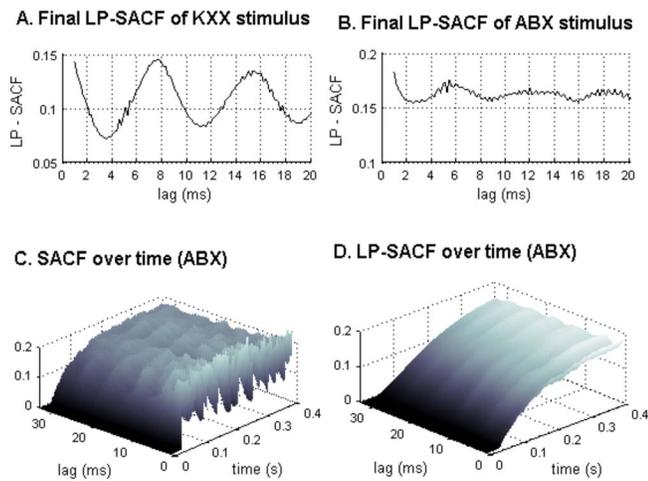
FIG. 5. (Color online) Model responses for the *KXX* and *ABX* click trains, with *K*=5 ms (see text). (A) Final LP-SACF for the *KXX* stimulus. (B) Final LP-SACF for the *ABX* stimulus. (C) Evolution of the SACF across time for the *ABX* stimulus. (D) Evolution of the LP-SACF across time for the same *ABX* stimulus.

The two click trains used by Pressnitzer *et al.* (2002, 2004) had mixtures of regular and random interclick intervals. The first stimulus (*KXX*) contained a single interval of fixed duration (*K*) followed by two intervals of random duration (mean duration, *K*/2). The second stimulus (*ABX*) consisted of trains of three random intervals with the constraint that the duration of the first pair of intervals summed to *K* and the third interval had a mean duration of *K*. Both click trains have the same average click rate of $3/(2K)$ clicks/s. Also, both click trains contain an interval of duration *K*. For the *KXX* stimulus, this interval occurs between the first and second clicks while for the *ABX* stimulus, the interval occurs between the first and the third clicks. Despite having identical mean click rates, the stimuli are reliably heard to have different pitches; *ABX* has a higher pitch than *KXX*.

These stimuli, *KXX* and *ABX* (*K*=5 ms), were presented to the model at a spectrum level of 60 dB SPL for a duration of 400 ms with onset and offset ramps of 5 ms. Click trains were high-pass filtered at 3 kHz. The response of the model is shown in Fig. 5 where it can be seen that the LP-SACFs for the two stimuli are clearly different. The LP-SACF of the *KXX* stimulus has a major peak at around 7.5 ms repeating at multiples of 7.5 ms indicating a predicted pitch of 133 Hz. The same prediction was made using the Euclidean-distance metric (not shown). In contrast, the *ABX* stimulus has its repeating peak starting at around 5.5 ms (predicted pitch 182 Hz). These results are consistent with the psychophysical observation of a lower pitch for the *KXX* stimuli. This result replicates the findings of Pressnitzer *et al.* (2004) and confirms that the operation of the present model is consistent with their principles.

The lower panels in Fig. 5 show the evolution of the SACF and the LP-SACF during one of the two stimuli (*ABX*). The development of the major autocorrelation peaks can be seen in the LP-SACF. However, the peaks are only intermittently represented in the SACF. The LP-SACF by virtue of its longer time window is able to average these

peaks and produce a more stable representation. Hence, a long-term autocorrelation as used by Pressnitzer *et al.* (2002) is not necessary.

## D. Evaluation 4

Carlyon *et al.* (2002) studied the pitch of a click train with alternating intervals of 4 and 6 ms that was bandpass filtered between 3.9 and 5.4 kHz to remove any resolved frequency components. They found that this stimulus generated pitch matches in the region of 4.5–7 ms when presented at a level of 54 dB SPL against a background of pink noise. The geometric mean of all matches was 5.7 ms. The regular intervals between the clicks of 4 and 6 ms might have been expected to give rise to pitch matches corresponding to either or both of these values. The second-order interval of 10 ms might also be a candidate. However, none of these were regularly reported by listeners. The authors performed a SACF analysis based on the method used by Meddis and O'Mard (1997). Their computations indicated likely pitch matches at 4, 6, and 10 ms. The authors concluded that autocorrelation analysis was unable to explain the results.

Carlyon *et al.* (2008) demonstrated that the combined AN responses to the 6 ms interval, measured as compound action potentials (CAPs), were stronger than for the 4 ms interval. Therefore, they suggested that a population of more central neurons, which respond only when their inputs exceed a fixed threshold value, would respond preferentially to the 6 ms intervals, resulting in a total average response closer to this interval, which would explain the listeners' preference for matching a pitch very close to the longer first-order interclick interval.

This raises the question of whether the AN spiking probabilities generated by the auditory peripheral model could reflect in some indirect way the above findings. If the responses were slightly stronger for the 6 ms intervals than for the 4 ms intervals, then a long-term periodicity analysis might also illustrate the dominance of the 6 ms period (on average within frequency channels). We repeated the autocorrelation analysis using the current model and obtained a more favorable outcome. Particular care was taken to reproduce the experimental stimulus exactly by filtering the clicks appropriately and adding background pink noise. Care was also taken to use the same pitch-match comparison stimuli as used in the original psychophysical experiment. The experiment was simulated using click trains with alternating intervals of 6 and 4 ms. The clicks were presented at an overall level of 78 dB SPL for 400 ms. They were bandpass filtered with cut-off frequencies of 3900 and 5300 Hz, generated using an eight-order Butterworth filter yielding an attenuation of 24 dB at half an octave above and below the cut-off frequencies. The background pink noise had the characteristics indicated in the psychophysical experiment. The stimulus was gated on and off with 50 ms raised-cosine ramps.

Comparison stimuli were generated according to the author's description: 29 isochronous pulse trains, with periods ranging from 2 to 14 ms in steps of 7%, with the period rounded to the nearest 0.1 ms. These were filtered in the same way as the test stimuli and LP-SACFs were generated

## A. Stimulus (400 ms duration)
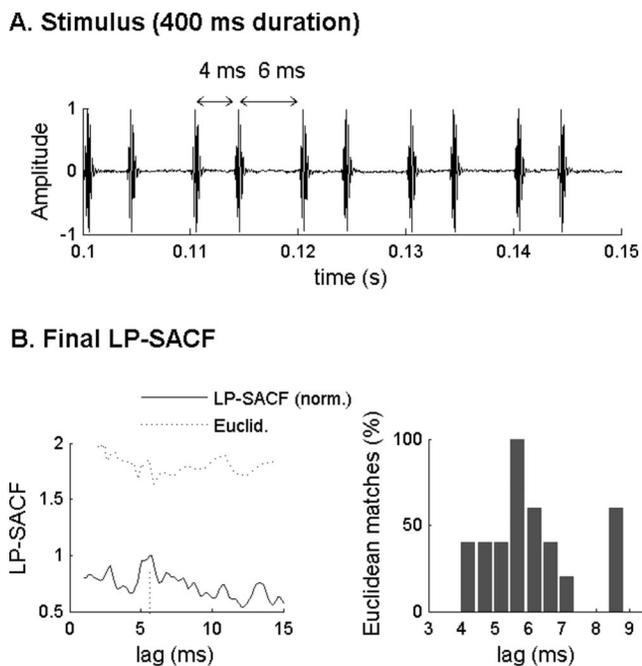


## B. Final LP-SACF



FIG. 6. Model response to the 6–4 ms alternating click train (Carlyon *et al.*, 2002) with bandpass filtering and pink noise included (see text). (A) Stimulus waveform (amplified for better visualization). (B) The left plot shows the final LP-SACF normalized response (solid line) and the corresponding Euclidean distance (dotted line). The maximum peak occurs at 5.76 ms and the minimum of the Euclidean distance occurs at 5.91 ms. The right plot shows a histogram of the Euclidean-distance matching for 20 different realizations of the background noise.

to act as comparison templates. The best-match comparison pulse train was chosen on the basis of the smallest Euclidean distance between the test stimulus LP-SACF and that of the comparison stimulus.

The LP-SACF derived from the model response is shown in Fig. 6(b). It has a maximum peak at 5.76 ms, in agreement with the perceptual data (Carlyon *et al.*, 2002, 2008). The dotted line above the LP-SACF shows the Euclidean distances between the range of templates and the test stimulus. For the Euclidean-distance measure, the *x*-axis represents the interclick interval so that the best pitch match (smallest Euclidean distance) is at 5.91 ms. This contradicts our intuition that matches would be most likely to occur at both 4 and 6 ms, the intervals between the successive clicks. In this respect, the result agrees with the main findings of the original experiment. The distribution of matches for several realizations of the background noise [right plot in Fig. 6(b)] is not clearly unimodal, in contrast with the results obtained Carlyon *et al.* (2002, 2008). Nevertheless, the mean value of the predictions is robust for large numbers of stimulus realizations. This result further supports the claim that the *current* model is qualitatively consistent with the experimental results in that it systematically predicts pitch matches closer to the longest first-order interclick interval. As a result we argue that an autocorrelation account of pitch perception is not contradicted in a fundamental way, by recent results (Carlyon *et al.*, 2008). However, more research is needed to accurately model CAP responses of the AN to this stimulus (Carlyon *et al.*, 2008).
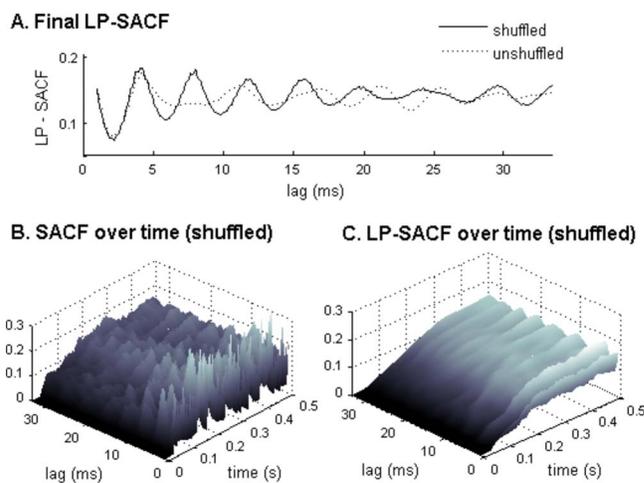


FIG. 7. (Color online) Response of the model to the *KX* click trains (*K* =4 ms). (A) Final LP-SACF for the shuffled (solid line) and unshuffled (dotted line) click trains. (B) Evolution of the SACF over time for the shuffled click train. (C) Evolution of the LP-SACF for the same shuffled stimulus.

### E. Evaluation 5

Yost *et al.* (2005) experimented with *KX* click trains where a regular interval (duration *K* ms) was alternated with a random interval (*X*). The duration of the random interval was uniformly distributed between 0 and 2*K* ms. This stimulus was then used to generate a second stimulus by randomly reordering all the interclick intervals. The "shuffled" click train contained exactly the same interclick intervals as the first "unshuffled" click train; only the sequence of the intervals was different. Surprisingly, the randomly shuffled click trains were typically heard to have a *greater* pitch strength than the unshuffled click trains, even though they were less regular as a result of the shuffling. When Yost *et al.* (2005) computed the SACF using an earlier version of the model (Meddis and Hewitt, 1991), there was little to indicate that the shuffled click train would be judged to have a greater pitch strength. The same was true of the autocorrelation of the stimulus waveform. They concluded that current autocorrelation models based on the long-term ACFs *cannot* account for the data of this study. Here, it will be shown that their results are indeed consistent with an autocorrelation analysis if an appropriate low-pass filtering is applied to the SACF.

The authors did note that the shuffled click trains contain longer runs of consecutive regular intervals than the unshuffled click trains. This is because an unshuffled click train, by definition, can never have two consecutive *K* intervals. In their view, the longer consecutive runs of the fixed interval are the key to understanding the phenomenon. This suggestion led us to expect that the new ACF with the low-pass filtering would reflect this long-term property of the stimulus

Figure 7 shows the LP-SACF of two (*KX*) click trains (*K*=4 ms), one unshuffled and the other shuffled. The random nature of these click trains means that these patterns will change from stimulus to stimulus but the examples given are typical. Unshuffled *KX* stimuli always have a single strong peak at *K* ms. The absence of secondary strong

peaks at multiples of $K$ ms reflects the fact that such intervals occur only rarely in the stimulus. On the other hand, the LP-SACFs of the shuffled click trains have a number of peaks at multiples of $K$ ms. The regularly spaced multiple peaks are caused by runs of consecutive regular intervals that occur by chance. Figure 7(b) compares the SACF with the LP-SACF over time. The obvious repeating peaks are not easily visualized in the SACF but are clearly present in the LP-SACF

The height of the first peak is approximately the same for both LP-SACF functions. This is not surprising as both shuffled and unshuffled stimuli have the same number of fixed-duration intervals. The height of the first peak of the SACF has often been taken to predict the strength of the pitch percept (Yost, 1996; Patterson *et al.*, 1996). In this case, it does not appear to be a useful guide for predicting which of these two stimuli will be perceived as more tonal.

The main difference between the two functions is the presence of a repeating series of equally spaced LP-SACF peaks in the case of the shuffled click train that are absent for the unshuffled click train. This repetition of equally spaced peaks is characteristic of the LP-SACFs of stimuli with a generally acknowledged clear pitch, such as harmonic tone complexes. If we accept the reasonable proposition that the presence of these additional regularly spaced peaks contributes to the overall tonality of click trains, we can conclude that the result is consistent with an autocorrelation approach to pitch perception.

## F. Evaluation 6

Oxenham *et al.* (2004) used "transposed stimuli" to further explore the arguments surrounding periodicity theories of pitch. These transposed stimuli are high-frequency carrier tones multiplied by a half-wave rectified low-frequency sinusoid. Essentially, these are pulses of high-frequency tones. The simplest example is a 4 kHz carrier tone pulsed at 100 Hz [see Fig. 8(a), upper panel]. The stimulus was presented at a level of 77 phons in a white noise background, low-pass filtered at 600 Hz and at a level of 27 dB below the overall level of the tones.

This stimulus gives rise to a weak pitch sensation [see Fig. 2(a) in the work of Oxenham *et al.*, (2004)]. Figure 8(a) (lower plot) illustrates the LP-SACF for this stimulus (solid line). It has broad peaks around 10, 20, and 30 ms. However, these are clearly less prominent than those in the LP-SACF of a similar stimulus but in which the carrier frequency is 100 Hz (dashed line), which has a clear pitch. This result suggests that the pitch sensation of the transposed tone is weak, which is consistent with the findings of Oxenham *et al.* (2004).

A more complex stimulus is the combination of three carrier tones with frequencies of 4, 6.35, and 10.08 kHz modulated at 300, 400, and 500 Hz, respectively.[2] This stimulus was presented at an overall level of 65 dB SPL, in a background pink noise bandpass filtered (31.2–1000 Hz) at a similar level to that in the previous experiment.

If the auditory system aggregates periodicity information, one might expect that this would also be heard as a
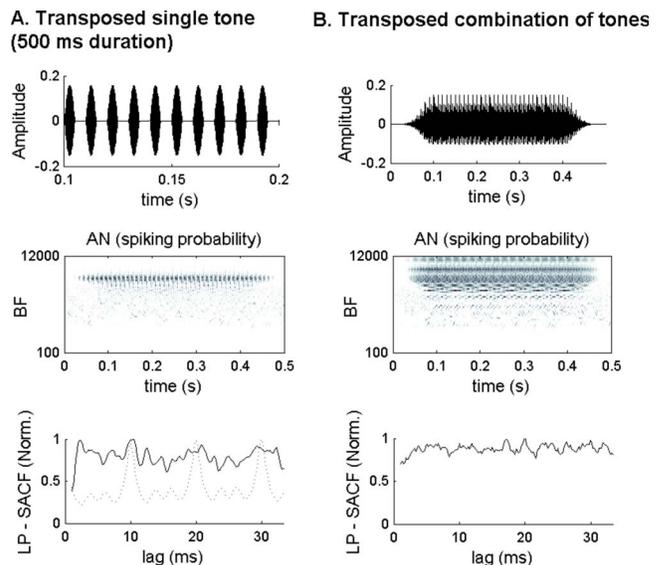


FIG. 8. (Color online) Autocorrelation analysis for 500 ms transposed stimuli (Oxenham *et al.*, 2004). (A) The upper panel shows the waveform of a single transposed pure tone of 4 kHz modulated at 100 Hz (amplified for better visualization). The middle panel shows the corresponding AN spiking probabilities (before the addition of noise). The lower panel shows the final LP-SACF. (B) Waveform of the sum of three transposed pure tones of 4, 6.35, and 10.08 kHz modulated at 300, 400, and 500 Hz, respectively. The middle panel shows the corresponding AN spiking probabilities (before the addition of noise). The lower panel shows the final LP-SACF.

weak 100 Hz pitch by analogy with a stimulus consisting of three pure tones at 300, 400, and 500 Hz. For this stimulus, however, the LP-SACF has an almost totally random structure [Fig. 8(b), lower plot]. Small peaks are present at 20 and 30 ms but the pattern is much less strongly modulated than the LP-SACF for the 100 Hz modulated carrier described above. This result agrees with the observations of Oxenham *et al.* (2004) who found that subjects performed even more poorly with multiple transposed tones than with a single transposed tone in a pitch discrimination task. Only one out of four of their subjects was able to discriminate the virtual pitch for these stimuli and that subject was not able to make pitch matches to the missing $F0$.

The experimenters analyzed their stimuli using an autocorrelation model very similar to the model studied in this report and obtained a different result from ours; the pure and transposed three-tone harmonic complexes produced very similar SACFs in their analysis. Both stimuli showed a distinct peak at a time interval corresponding to the reciprocal of the $F0$. Thus, the model correctly predicted that the $F0$ would be perceived in the case of the pure tones, but incorrectly predicted a similar pitch percept in the case of the transposed tones.

One of the several possible factors that could explain the difference between their analysis and that given in Fig. 8(b) are the changes to the auditory model at the level of the BM that have taken place since they performed their study. Oxenham *et al.* (2004) used a bank of linear gammatone bandpass filters to simulate cochlear filtering. The current model uses nonlinear filters to simulate the compression that takes place on the BM and this has consequences for the shape of the filters as a function of signal level.

However, the nonlinear response of the BM is such that the width of the filters increases at higher signal levels. At high signal levels, the frequency components will not necessarily be resolved as a result of the wider filters. Figure 8(b) (middle row) shows the pattern of AN responses across channels predicted by the model (before the addition of noise). The narrow bands implied by the term "resolved" are not clearly visible. It may be true that the BFs corresponding to the carrier frequencies contain only single-frequency activity but the intermediate channels are being excited by more than one frequency. The computation of the SACF involves a mandatory summation across *all* channels and therefore the intermediate channels will be well represented. Dreyer and Delgutte (2006) examined the AN response in cats to transposed tones and found that phase locking to transposed tones degraded substantially as signal levels were raised above threshold. SACFs resulting from stimuli containing unresolved components are typically flatter because the envelope of the stimulus has a greater influence than the individual sinusoidal components. This could be the case here. However, from our simulations, it is not conclusive that the within-channel interactions of the filters are responsible for the inaudibility of the pitch. Other possible factors include level-dependent compression, and possible saturation effects at signal levels such as the ones used in this experiment.

Nevertheless, at very low signal levels, our model replicates the results of the linear model used in the work of Oxenham *et al.* (2004) (not shown). Therefore, it is a prediction of the model that the pitch of the transposed complex should be audible at very low levels where the auditory system is functioning linearly and the auditory filters are sharply tuned, as assumed in the analysis of Oxenham *et al.* (2004). Such conditions may exist near threshold or in certain individuals for whatever reason. We note that one of the four subjects in the experimental study (their Fig. 3 subject S7) was able to make successful pitch discriminations at $F0$ for the transposed complex. A speculative explanation for this might be reduced compression in this subject. We do not, however, underestimate the difficulties of carrying out such a test given the need to demonstrate linear responses, narrow filters, and audibility of all components when close to threshold.

In summary, the results of Oxenham *et al.* (2004) are mirrored qualitatively in the response of the model to transposed stimuli and their data do not contradict the autocorrelation approach to modeling pitch perception. The difference in the modeling results could be a consequence of the nonlinear characteristics of the BM response in the new model. However, our conclusion should be qualified by the uncertainty that surrounds any model that purports to represent the exact pattern of action potentials in the human AN. We have no way of checking this and the evidence for such models is always indirect.

## IV. DISCUSSION

The question is whether the autocorrelation approach should be completely rejected on the basis of recently published psychophysical studies. The results of experiments using click trains consisting of a mixture of regular and irregular interclick intervals (e.g., Kaernbach and Demany, 1998; Yost 1996) are certainly inconsistent with an approach based on the application of a long-term autocorrelation analysis to the acoustic *waveform* of the stimulus. However, the main tradition of the application of autocorrelation that began with Licklider (1951) has stressed the application of the analysis to the activity of the AN. Such analysis depends on a combination of two separate theories. The first is a theory of how the AN activity is generated; the second concerns the most appropriate method for analyzing this activity. If either theory has shortcomings, this will be reflected in a failure of the model to cope with some of the data.

In this report we have revisited a number of published psychophysical studies whose data had been argued to be inconsistent with an existing autocorrelation model of pitch. Here it has been shown that these difficulties can be reduced if modifications are made both to the peripheral model and to the method of analysis. Modifications to the peripheral model include the addition of nonlinearities in the BM filtering and an improved model of the generation of the IHC receptor potential. Modifications to the analysis method consisted of the introduction of lag-dependent time constants used in computing the ACFs, and an additional stage that integrated the output of the SACF over a longer time window. Together they give a more useful account of the data.

The 2 or 3 ms short time constant originally suggested by Licklider (1951) has proved in the past to be successful for many stimuli, but the pitch characteristics of irregular click-train stimuli cannot be so easily accommodated because the regularities in the stimuli can only be assessed over a longer time period. The solution to the problem can be found in Wiegrebe's (2001) study of the pitch of repeated pulses of noise. He suggested a multistage approach employing a second wider temporal integration window. The present cascade approach is the practical application of this idea. The longer time constant is the major contributor to the ability of the revised model to explain the pitch properties of the irregular click trains described above.

The shuffled click-train data of Yost *et al.* (2005) present another problem concerning the interpretation of the autocorrelation analysis. Their shuffled click trains were judged to be more "tonal" than the unshuffled click trains. This could not be predicted on the basis of the height of the highest peak in the LP-SACF because both shuffled and unshuffled click trains generated peaks of (on average) the same height. The key difference between the two functions was to be found in the pattern of minor peaks. A shuffled click train can be distinguished from the unshuffled version by the presence of a repeating series of equally spaced minor peaks. These peaks are commonly observed in the SACF (and in the LP-SACF) of harmonic tone complexes but often ignored by researchers as redundant. However, these new stimuli indicate that the repeating peaks should contribute to our predictions of the salience of the perceived pitch.

The autocorrelation analysis will be of limited value if it is based on the output of an inadequate peripheral model. Such models remain primitive but are subject to continuous revision. Our insights into the significance of the many

subtleties of peripheral auditory processing are developing in parallel. Of particular interest are the ideas of Pressnitzer *et al.* (2002) concerning the role of the half-wave rectification that occurs in the generation of the receptor potential in the IHC. Here it has been shown that their proposal survives translation to a more sophisticated auditory model and the new cascade method of analyzing the model output.

Another nonlinearity in peripheral processing occurs at the level of the BM and involves compression of the stimulus waveform at signal frequencies close to the filter BF. This compression is less evident at remote frequencies with the consequence that the width of the filter increases with signal level. As a result, the individual frequency components of a stimulus spread their effects more widely over the BM and the excitation pattern changes radically as the level of the stimulus is increased. This is one of the possible factors in explaining the success of our new model in the study of Oxenham *et al.* (2004). More research is needed to understand precisely which aspect of the auditory peripheral model is primarily responsible for this perceptual phenomenon and for the perception of alternating click trains (Carlyon *et al.*, 2008).

Both peripheral models and methods for analyzing their output are continuing to evolve and we must expect increasingly sophisticated accounts of pitch perception to emerge as a consequence. The novel and challenging stimuli described above have an important role to play in this evolution. Nevertheless, we conclude that, for the present, Licklider's (1951) view that pitch perception can be understood in terms of a periodicity analysis of the activity of the AN remains intact.

## ACKNOWLEDGMENTS

[1]Software used in this study is available on request from the authors.
[2]For this evaluation, the range of BF channels in the auditory model was extended to 12 kHz for all stimuli.

Bernstein, J. G. W., and Oxenham, A. J. (**2005**). "An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination," J. Acoust. Soc. Am. **117**, 3816–3831.

Cariani, P. A., and Delgutte, B. (**1996a**). "Neural correlates of the pitch of complex tones. I. Pitch and pitch salience," J. Neurophysiol. **76**, 1698–1716.

Cariani, P. A., and Delgutte, B. (**1996b**). "Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase-invariance, pitch circularity, rate-pitch, and the dominance region of pitch," J. Neurophysiol. **76**, 1717–1734.

Carlyon, R. P. (**1996**). "Encoding the fundamental frequency of a complex tone in the presence of a spectrally overlapping masker," J. Acoust. Soc. Am. **99**, 517–524.

Carlyon, R. P., Wieringen, A., Long, C. J., Deeks, J. M., and Wouters, J. (**2002**). "Temporal pitch mechanisms in acoustic and electric hearing," J. Acoust. Soc. Am. **112**, 621–633.

Carlyon, R. P., Mahendran, S., Deeks, J. M., Long, C. J., Axon, P., Baguley, D., Bleeck, S., and Winter, I. M. (**2008**). "Behavioral and physiological correlates of temporal pitch perception in electric and acoustic hearing," J. Acoust. Soc. Am. **123**, 973–985.

Ciocca, V., and Darwin, C. J. (**1999**). "The integration of nonsimultaneous frequency components into a single virtual pitch," J. Acoust. Soc. Am. **105**, 2421–2430.

Denham, S. L. (**2005**). "Dynamic iterated ripple noise: Further evidence for the importance of temporal processing in auditory perception," BioSystems **79**, 199–206.

Dreyer, A., and Delgutte, B. (**2006**). "Phase locking of auditory-nerve fibers to the envelopes of high-frequency sounds: Implications for sound localization," J. Neurophysiol. **96**, 2327–2341.

Gockel, H., Plack, C. J., and Carlyon, R. P. (**2005**). "Reduced contribution of a nonsimultaneous mistuned harmonic to residue pitch," J. Acoust. Soc. Am. **118**, 3783–3793.

Grose, J. H., Hall, J. W., and Buss, E. (**2002**). "Virtual pitch integration for asynchronous harmonics," J. Acoust. Soc. Am. **112**, 2956–2961.

Hall, J. W., and Peters, R. W. (**1981**). "Pitch for nonsimultaneous successive harmonics in quiet and noise," J. Acoust. Soc. Am. **69**, 509–513.

Kaernbach, C., and Bering, C. (**2001**). "Exploring the temporal mechanisms involved in the pitch of unresolved harmonics," J. Acoust. Soc. Am. **110**, 1039–1048.

Kaernbach, C., and Demany, L. (**1998**). "Psychophysical evidence against the autocorrelation theory of auditory temporal processing," J. Acoust. Soc. Am. **104**, 2298–2306.

Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," Experientia **7**, 128–134.

Meddis, R. (**2006**). "Auditory-nerve first-spike latency and auditory absolute threshold: A computer model," J. Acoust. Soc. Am. **119**, 406–417.

Meddis, R., and Hewitt, M. J. (**1991**). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. Pitch identification," J. Acoust. Soc. Am. **89**, 2866–2882.

Meddis, R., and O'Mard, L. (**1997**). "A unitary model of pitch perception," J. Acoust. Soc. Am. **102**, 1811–1820.

Micheyl, C., and Carlyon, R. P. (**1998**). "Effects of temporal fringes on fundamental-frequency discrimination," J. Acoust. Soc. Am. **104**, 3006–3018.

Oxenham, A. J., Bernstein, J. G. W., and Penagos, H. (**2004**). "Correct tonotopic representation is necessary for complex pitch perception," Proc. Natl. Acad. Sci. U.S.A. **101**, 1421–1425.

Patterson, R. D., Handel, S., Yost, W. A., and Datta, J. (**1996**). "The relative strength of tone and noise components in iterated rippled noise," J. Acoust. Soc. Am. **100**, 3286–3294.

Plack, C. J., and White, L. (**2000**). "Perceived continuity and pitch perception," J. Acoust. Soc. Am. **108**, 1162–1169.

Pressnitzer, D., Patterson, R. D., and Krumbholz, K. (**2001**). "The lower limit of melodic pitch," J. Acoust. Soc. Am. **109**, 2074–2084.

Pressnitzer, D., de Cheveigné, A., and Winter, I. M., (**2002**). "Perceptual pitch shift for sounds with similar waveform autocorrelation," ARLO **3**, 1–6.

Pressnitzer, D., de Cheveigné, A., and Winter, I. M., (**2004**). "Physiological correlates of the perceptual pitch shift for sounds with similar waveform autocorrelation," ARLO **5**, 1–6.

Slaney, M., and Lyon, R. F. (**1990**). "A perceptual pitch detector," International Conference on Acoustics, Speech, and Signal Processing, (ICASSP-90), Albuquerque, USA.

Sumner, C. J., O'Mard, L. P., Lopez-Poveda, E. A., and Meddis, R. (**2002**). "A revised model of the inner-hair cell and auditory nerve complex," J. Acoust. Soc. Am. **111**, 2178–2189.

Wiegrebe, L. (**2001**). "Searching for the time constant of neural pitch extraction," J. Acoust. Soc. Am. **109**, 1082–1091.

Yost, W. A. (**1996**). "Pitch of iterated rippled noise," J. Acoust. Soc. Am. **100**, 511–518.

Yost, W. A., Mapes-Riordan, D., Shofner, W., Dye, R., and Sheft, S. (**2005**). "Pitch strength of regular interval click trains with different length "runs" of regular intervals," J. Acoust. Soc. Am. **117**, 3054–3068.