On the Detection of Melodic Pitch in a Percussive Background*

PREETI RAO, *AES Member,* **AND SAURABH SHANDILYA**

Department of Electrical Engineering, Indian Institute of Technology, Bombay, India

The extraction of pitch (or fundamental frequency) information from polyphonic audio signals remains a challenging problem. The specific case of detecting the pitch of a melodic instrument playing in a percussive background is presented. Time-domain pitch detection algorithms based on a temporal autocorrelation model, including the Meddis–Hewitt algorithm, are considered. The temporal and spectral characteristics of percussive interference degrade the performance of the pitch detection algorithms to various extents. From an experimental study of the pitch estimation errors obtained on a set of synthetic musical signals, the effectiveness of the auditory-perception–based modules of the Meddis–Hewitt pitch detection algorithm in improving the robustness of fundamental frequency tracking in the presence of percussive interference is discussed.

0 INTRODUCTION

The problem of pitch (or fundamental frequency) extraction of periodic signals in the presence of interfering sounds and noise is an important problem in both speech and music applications. Apart from the value of pitch information per se, a knowledge of the time-varying fundamental frequency can be useful in the separation and reconstruction of a harmonic source from a sound mixture. A number of pitch detection algorithms (PDAs) have been proposed over the decades. But while each has had a measure of success in the targeted application, no single PDA is found suitable for all types of signals and conditions. This engineering report presents an investigation of the performance of some well-known PDAs in estimating the fundamental frequency of a melodic instrument playing in the presence of percussive background. This is a restricted case of the larger problem of musical pitch detection in polyphony. Nevertheless it is an important problem. For instance, classical Indian vocal and instrumental music is always accompanied by percussive instruments providing the rhythmic structure. The melody itself is strongly characterized by the presence of microtones and continuous pitch variation. Detecting the melodic pitch contour has important applications in music recognition and for generating metadata in audio content retrieval systems.

Pitch determination of speech signals has been the subject of research for decades [1]. It is marked by challenges arising from the complex temporal and spectral structure of speech as well as its nonstationary nature. While musical applications require a higher accuracy of pitch estimation than speech applications, tracking the pitch of a melodic instrument or singing voice is easier than tracking that of speech signals due to the relatively slowly changing signal characteristics. The presence of interfering sounds from percussive accompaniment, however, would be expected to adversely affect the accuracy of pitch estimates of any given PDA for musical applications. Percussive sounds are characterized by rapidly varying temporal envelopes, mixed partials-plus-broad-band noise spectra, and low values of signal-to-interference ratio in localized time intervals. The peculiar problems in pitch detection posed by such interference form the main focus of the present study. In particular, we consider the degradation caused by the presence of inharmonic interfering partials. The robustness of pitch detection methods to additive broad-band noise has been studied in various contexts in the literature (see, for example, [2]). While the motivation for the present work is the pitch tracking of the melodic instruments including the singing voice in the presence of percussion, we use in our experiments test signals from a set of MIDI instrument voices to enable controlled experiments focusing on the effects of percussion, with access to the "ground truth" pitch available.

In this study an important subclass of PDAs, namely, those based on the detection of periodicity in the time-domain signal by means of short-term correlation, is considered. Autocorrelation-based pitch determination, used

*Manuscript received 2003 July 9; revised 2004 February 9. An early version of this work was presented at the 114th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 2003 March 22–25, under the title "Pitch Detection of the Singing Voice in Musical Audio."

widely in speech analysis [1], has also been found suitable for the pitch tracking of monophonic musical signals [3]. The present study investigates the Meddis–Hewitt perceptual PDA as an example of a more sophisticated algorithm also based on the detection of periodicity via temporal autocorrelation.

The engineering report is organized as follows. Section 1 provides a brief overview of various PDAs, with an introduction to the PDAs chosen for the present study. The subsequent sections describe the implementation of the functional blocks of the PDAs and the evaluation of the PDAs by an experiment on the synthetic signal test set. The study concludes with a discussion of the observations targeted toward obtaining insights into the performance of the PDAs with respect to signal and interference characteristics.

1 PITCH DETECTION ALGORITHMS

Time-domain PDAs, the oldest pitch detection algorithms, are based on measuring the periodicity of the signal via the repetition rate of specific temporal features. Frequency-domain PDAs, on the other hand, are based on detecting the harmonic structure of the spectrum. Among the simpler time-domain PDAs is the popular autocorrelation function (ACF)–based PDA. The definition of the "biased" autocorrelation function is given by [1]

$$\mathrm{ACF}(k,\tau) = \sum_{i=0}^{N-\tau-1} y(k+i)y(k+i+\tau) \qquad (1)$$

where k and τ are position of window and correlation lag, respectively, and y is the input signal.

For a pure tone, the ACF exhibits peaks at lags corresponding to the period and its integral multiples. The peak in the ACF of Eq. (1) at the lag τ corresponding to the signal period will be higher than that at the lag values corresponding to multiples of the period. For a musical tone consisting of the fundamental frequency component and several harmonics, one of the peaks due to each of the higher harmonics occurs at the same lag position as that corresponding to the fundamental, in addition to several other integer multiples of the period (subharmonics) of each harmonic. Thus a large peak corresponding to the sum contribution of all spectral components occurs at the period of the fundamental (and higher integral multiples of the period of the fundamental). This property of the ACF makes it very suitable for the pitch tracking of monophonic musical signals. The ACF PDA chooses as the pitch period the lag corresponding to the highest peak within a range of lags.

In contrast to the simplicity of the ACF pitch detector are more recent PDAs, also based on autocorrelation, but derived more closely from the mechanism of temporal coding in the human auditory system. These PDAs can in fact be viewed in terms of preprocessing of the signal followed by autocorrelation-based detection. There are a number of variants in this class of PDAs, but they all share some important characteristics [4]. They decompose the signal into frequency bands defined by the auditory filters of the cochlea. Next, nonlinear processing corresponding to hair-cell transduction is applied and the temporal periodicity detected separately in each frequency channel by means of autocorrelation. Finally the across-channel information is combined to produce a single pitch estimate. The recent pitch perception model of Meddis and Hewitt [5] has gained much prominence due to its demonstrated ability to predict the results of certain crucial pitch perception experiments. The Meddis–Hewitt PDA is based on the functional modules of the auditory periphery with added processing stages that emulate auditory processing, which is considered to be more central. The various stages of the PDA are a bandpass filter representing the transfer function of the outer ear and middle ear canal, a bank of filters modeling the basilar membrane response, followed by a model of the inner hair cell applied to each filter channel output to simulate neural transduction, obtaining a series of firing probabilities. Next an ACF periodicity detector is applied to each of the hair-cell model outputs. Finally a summary autocorrelation function (SACF) is formed by adding the ACFs so obtained across the frequency channels. A search for the highest peak in the relevant range of SACF lags provides an estimate of the pitch period.

The added presence of noise and inharmonic partials due to an interfering signal perturbs the shape and location of the peaks contributed by the signal harmonics in the ACF. Thus the traditional ACF pitch detector applied to a musical signal with percussive accompaniment would be expected to be adversely affected by the presence of noise and inharmonic frequency components contributed by the percussion. On the other hand, in the case of perception-based PDAs, the signal is processed by a number of auditory-model–based blocks before being subjected to periodicity extraction via the ACF. In particular, a combination of linear and nonlinear filtering is applied, and the temporal periodicity information itself is computed via the ACF separately in each frequency channel. It is of interest to examine whether and to what extent these perceptually motivated enhancements improve the reliability of the pitch estimation for harmonic musical signals with conspicuous background interference. While much recent work has investigated the ability of perceptual PDAs to predict subjectively perceived pitch in psychoacoustic experiments, the present work examines the robustness of the PDA for estimating the signal fundamental frequency in the presence of interference. By means of carefully designed test signals, we study the pitch estimation errors obtained by the PDAs in the presence of percussive interference with respect to the underlying pitch of the melodic voice, and later attempt to explain these observations.

2 IMPLEMENTATION OF PDA FUNCTIONAL BLOCKS

Fig. 1 provides a modular structure for the PDA based on the Meddis–Hewitt pitch perception model [5]. The individual blocks represent various stages of the algorithm, each of which may, in principle, be implemented in multiple ways.

Block 1 represents the outer ear and middle ear (OEM) prefiltering, with the magnitude response shown in Fig. 2.

Essentially a bandpass filter with a resonance frequency near 3 kHz, this block has been implemented by the cascade of an eighth-order low-pass IIR filter and a second-order parameterized high-pass filter with a high-pass filter parameter value of 0.94 [6]. The magnitude response is similar to the inverted absolute threshold-of-hearing curve, whereas below 1 kHz it is approximated by the inverted equal-loudness contour for high loudness levels.

Block 2 represents the cochlear filter bank with filter center frequencies that are equally spaced on the ERB (equivalent rectangular bandwidth) scale with the bandwidth increasing with the center frequency. A filter bank of 27 ninth-order gammatone filters of bandwidth 1 ERB each are based on the corresponding function of the HUTear library [7]. These filters have center frequencies ranging from 123 Hz to 5.636 kHz, or 4 to 30 on the ERB scale. The output of each filter simulates the pattern of vibration at a particular location on the basilar membrane.

Next the conversion of this mechanical activity to the neural spike generation events of the auditory nerve is simulated by block 3. The implementation of this module can range from a full model of the hair cell derived from a computational analysis of actual hair-cell and auditory-nerve processes [5] to simple half-wave rectification followed by low-pass filtering [8]. In 1986 Meddis proposed a model for the hair cell which simulates, through difference equations rich in parameters, several properties of the neural transduction and auditory nerve firing [9]. Important characteristics of the hair-cell model are its nonlinearity and frequency selectivity. The present implementation is based on the hair-cell model of Meddis and coworkers [9], [10] as implemented in the auditory model library of Slaney [11].

Block 4 calculates the ACF [Eq. (1)] of the signal input to this block. Keeping in mind the range of 150–800 Hz for the expected fundamental frequency, we have used a 40-ms window with 50% overlap (frame space 20 ms) for the computation of the ACF. Once the ACFs are obtained for all the channels [by implementing Eq. (1) on the hair-cell model output of each channel], block 5 performs the task of combining them. Combining can occur in the form of either simple or weighted addition. We use simple sum-

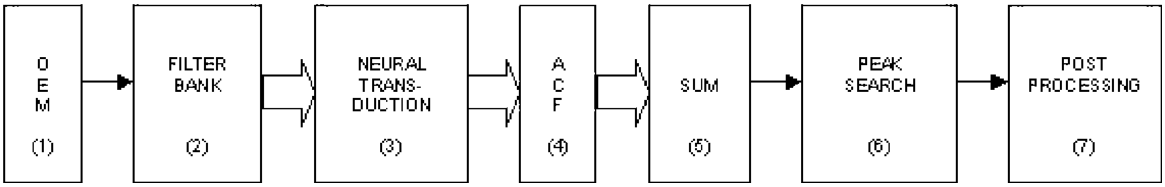


Fig. 1. Block diagram of functional blocks of Meddis–Hewitt PDA with postprocessing added.

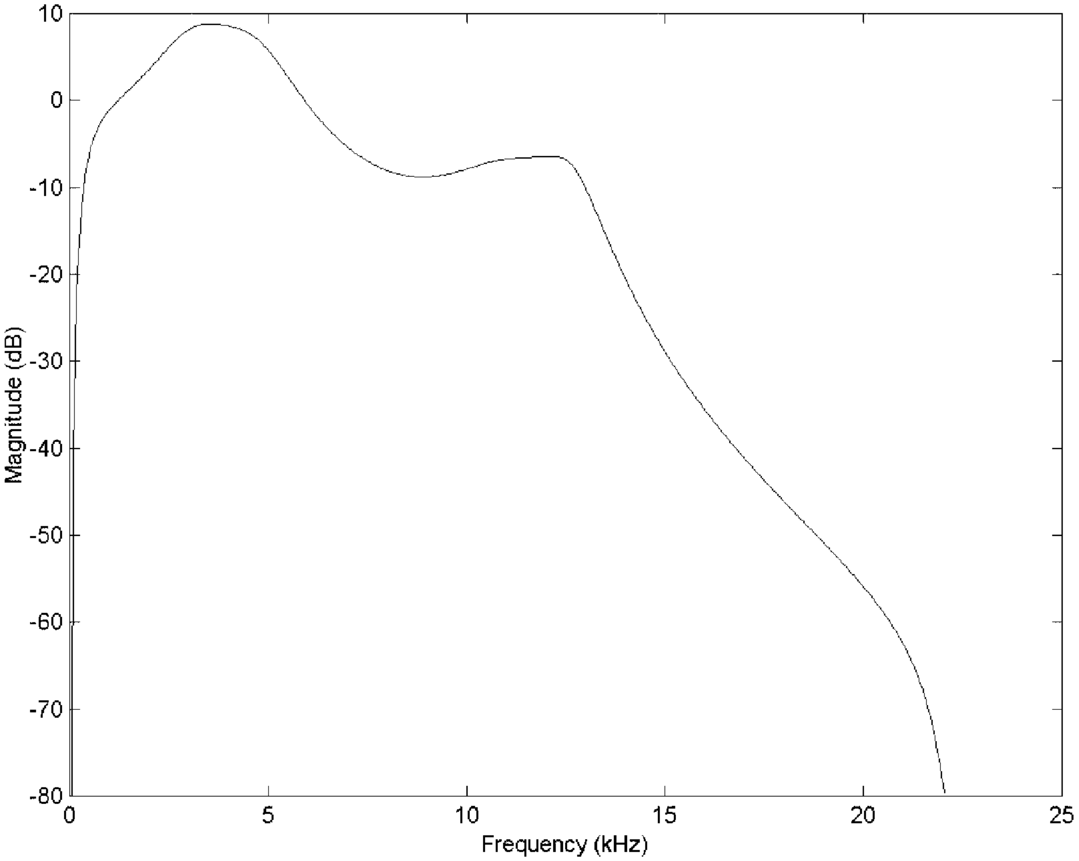


Fig. 2. Magnitude spectrum of outer-ear–middle-ear (OEM) filter.

ming. The combined ACFs are known as summary ACF (SACF). In block 6 the SACF is searched for the highest peak within a prespecified range (corresponding to the expected fundamental frequency range of 150–800 Hz). The lag value corresponding to the highest peak is accepted as the estimated pitch period. Block 7 is a postprocessing block, which smoothens out local variations in the pitch estimates across frames using a simple three-point median filter. The combination of all seven blocks constitutes the Meddis–Hewitt PDA with added postprocessing. In the next section we describe a procedure for evaluating the contribution of the various functional blocks of the Meddis–Hewitt PDA to improving robustness over direct ACF peak-based pitch detection.

3 EXPERIMENTAL EVALUATION

In order to investigate the performance of the pitch detection algorithms for the pitch estimation of a harmonic signal in a percussive background, a set of test signals was designed from available MIDI songs. Apart from the access to the "ground truth" pitch, the use of MIDI files provides great flexibility by allowing the inclusion or elimination of individual monophonic instrument channels, modifying the relative strengths of the component sounds, pitch transformations, and choice of the instrument playing the melody line as well as the percussive instrument.

3.1 Test Signal Set

A MIDI song of length 8 seconds was selected. It had a single harmonic instrument (alto sax) playing the melody accompanied by several percussive (nonpitched) voices, namely, hi hat, kick drum, and low agogo channels. The pitch range of the melody was 350–620 Hz and consisted of four similar phrases, with each phrase comprising five notes of various durations. [The pitch contour of the melody can be seen as the solid line in Fig. 5(b).] Further, in order to create a number of different test conditions, pitch-shifted versions of the melody were created as follows: high (up by 4 semitones to the range of 440–787 Hz) or low (down by 12 semitones to the range of 174.5–311 Hz). It may be remarked that the pitch transformations are achieved via "instrumental" pitch shifting, which implies that the relative amplitudes of the harmonics remain unchanged across fundamental frequency changes, in contrast to formant-corrected pitch shifting.

The set of percussive instruments represents a range of signal characteristics, as illustrated in the spectrograms of Fig. 3. The kick drum is a relatively fast decaying signal with predominantly low-frequency content. The hi hat is characterized by a slow time decay and a broadly spread spectral mixture of moderately strong partials and noise. Low agogo has low-noise content and strong partials all the way from 1.1 to 10 kHz, with a moderate rate of decay.

To obtain a variety of combinations of target and interference timbres, the song was transformed by changing the target instrument and then selecting only one of the interfering (percussive) instruments at a time. The selected target instrument voices were of different timbres, as shown by the magnitude spectra of a fixed note in Fig. 4. For the middle- and high-frequency ranges, baritone sax (prominent high harmonics), flute (weak high harmonics), and oboe (harmonics spread in frequency) were used. For the low-frequency range the flute was replaced by alto sax to incorporate a more natural sound.

The JAZZ MIDI sequencer, available as shareware,[1] was used to achieve the needed transformations. The melody line was recorded, switching off all the other channels, in each of the three pitch ranges with each of three target instruments. Thus a set of nine files containing pure melody was obtained. For each of these melody files we created three "corrupted" versions, each with only a single percussion channel turned on. The synchronization be-

[1] www.jazzware.com.

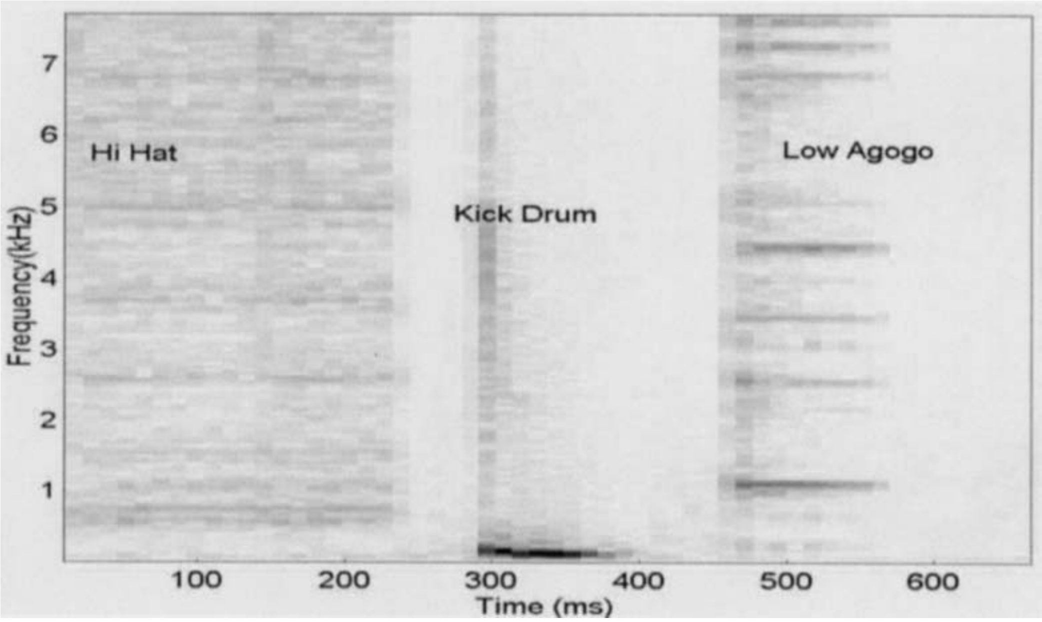


Fig. 3. Spectrogram illustrating time–frequency behavior of three percussion instruments.

tween the melody and each of the percussion tracks was such that a minimum number of percussion strikes fell in the silence region between target instrument notes. This led to percussion onsets being located at a variety of positions with respect to target note onset, steady state, and decay. Then relative amplitudes of the target and interfering signals were set so that the ratio of signal power to interference power (each of the powers being computed as the corresponding average over the nonsilent regions of the musical piece) remained at a fixed predefined value (equal to 2.0) for a set of test signals across various target instruments, percussive instruments, and pitch ranges. The signal-to-interference ratio, however, is only an average value with local values deviating greatly, depending on the position of the percussion strike with respect to the target instrument note onset. We thus obtained a total of 27 test signals, all sampled at 44.1 kHz.

3.2 Experiment

The PDAs were run on the pure test signals, with the same postprocessing applied to each PDA estimate to ensure a fair comparison.

Fig. 5 shows a sample of the experimental results obtained. The selected test signal is the combination of baritone sax and low agogo percussion. Fig. 5(a) shows the spectrogram of the test signal. The relatively continuous dark lines correspond to the harmonic partials of the baritone sax, and the very short dark segments that occur during the first (2 strikes), third, and fourth notes of each phrase correspond to the partials of the low agogo. Fig. 5(b) compares the true pitch track with the pitch track estimated by the traditional ACF pitch detector followed by a three-point median filter. We see that there are large pitch estimation errors, which coincide with the occurrence of the percussion and last over several frames. The simple three-point median postfilter only corrects isolated pitch errors. The instances of percussion where pitch errors do not occur seem to be characterized by overlapping partials between target and percussion. Fig. 5(c) illustrates the pitch contour obtained by using the Meddis–Hewitt PDA, also followed by the postfilter, on the same test signal. Indeed the performance has improved on using this algorithm.

A controlled-parameter experiment was carried out to study the behavior of the PDAs on the underlying signal characteristics, and to obtain an understanding of the role of each of the functional modules of the Meddis–Hewitt algorithm in influencing the pitch detection. Specifically, four PDAs were evaluated on the test data set. All four algorithms are derived from the generic block diagram of Fig. 1 by choosing different combinations of subblocks and/or different realizations of a specific subblock. The postprocessor of block 7 is a three-point median filter that is included in all four algorithms. The details of the four PDAs are as follows.

1) AC1: This algorithm incorporates blocks 4, 6, and 7 only. It corresponds to the traditional ACF pitch detector. Block 4, that is, the ACF calculation block, uses a rectangular window and the biased ACF computation of Eq. (1).

2) AC2: This algorithm incorporates block 1 with blocks 4, 6, and 7. This again is a traditional ACF pitch detector, but with outer ear/middle ear filtering included as a preprocessing function.

3) AC3: This algorithm comprises blocks 1, 3, 4, 6, and 7. This algorithm is an extension of AC2, where the neural transduction block (based on the hair-cell model of [9]) has been introduced. Unlike the Meddis–Hewitt PDA, the signal is not decomposed into separate frequency channels. Rather, the hair-cell nonlinearity is applied to the full band signal followed by ACF pitch detection to result in a single estimate of temporal periodicity per frame.

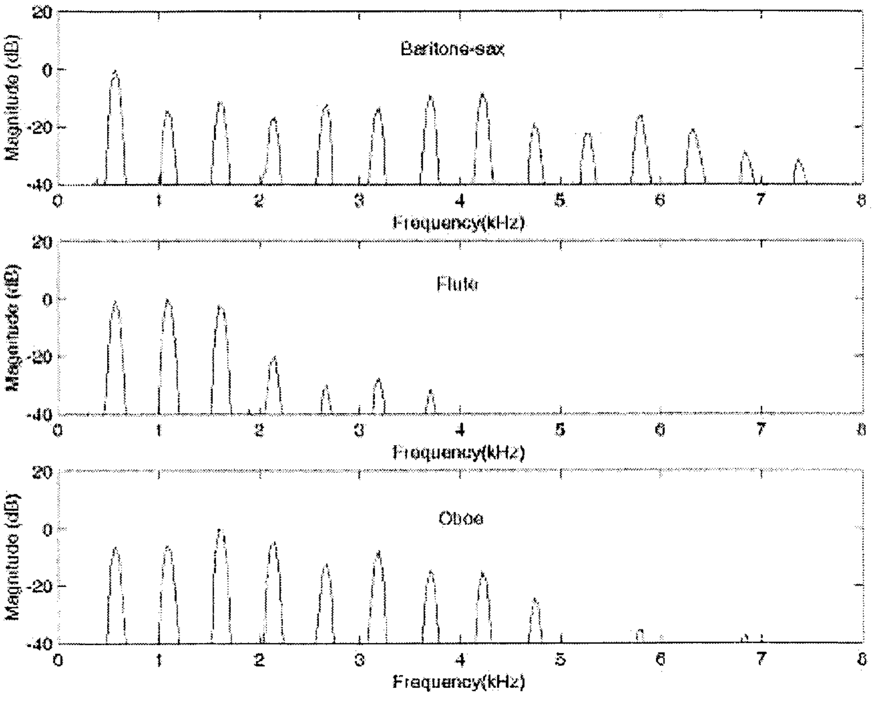


Fig. 4. Magnitude spectra of melodic instruments used in experimental study.

4) MH1: All blocks 1 to 7 are included, making this a complete implementation of the Meddis–Hewitt algorithm with postprocessing added. For the ACF, a Hamming-windowed biased ACF is computed for each channel using an efficient fast Fourier transform implementation.

3.3 Observations

The bar charts of Figs. 6 to 9 display the results of the experiment in terms of a count of the pitch errors with respect to the known reference pitch contour, arranged by the PDA configuration used. A pitch estimate is obtained for every analysis frame only in regions where the target instrument is playing. At a frame spacing of 20 ms, this comprises a total of 285 frames. A pitch error is defined to occur whenever the detected pitch deviates from the reference pitch by 3% or more (about half a semitone) of the reference pitch frequency. Of the detected pitch errors, those of magnitude less than 6% are labeled "fine" errors, whereas those of higher magnitude are labeled "gross." The gross errors are found typically to be pitch octave errors. It may be noted that in the absence of percussion, no pitch errors were observed in any of the PDA configurations.

From an inspection of the bar charts we note that the extent of pitch errors depends not only on the PDA but also on the percussion instrument, the pitch range of the target instrument, and the target instrument itself. Of the latter three factors, the most marked is the dependence of the PDA performance on the percussion characteristics. The ACF pitch detector AC1 makes a large number of errors for all three percussions. The target instrument pitch range that is most affected is seen to depend on the spectral characteristics of the interference. The kick drum with its low-frequency support affects the lowest pitch range the most, whereas the low agogo with its broad spectral mixture of partials and noise impacts all three pitch ranges. The introduction of OEM filtering in AC2 has the effect of an overall lowering of the extent of pitch errors. A strong exception to this happens to be the hi hat in the high-pitch range (note the changed scale of the error axis). The hair-cell model followed by autocorrelation (AC3) serves to reduce all errors further, with the only significant errors remaining in the low agogo signals. Finally, the full Meddis–Hewitt PDA MH1 reduces the errors in the low agogo signals of the low- and middle-pitch ranges, but worsens the performance slightly in the high-pitch range.

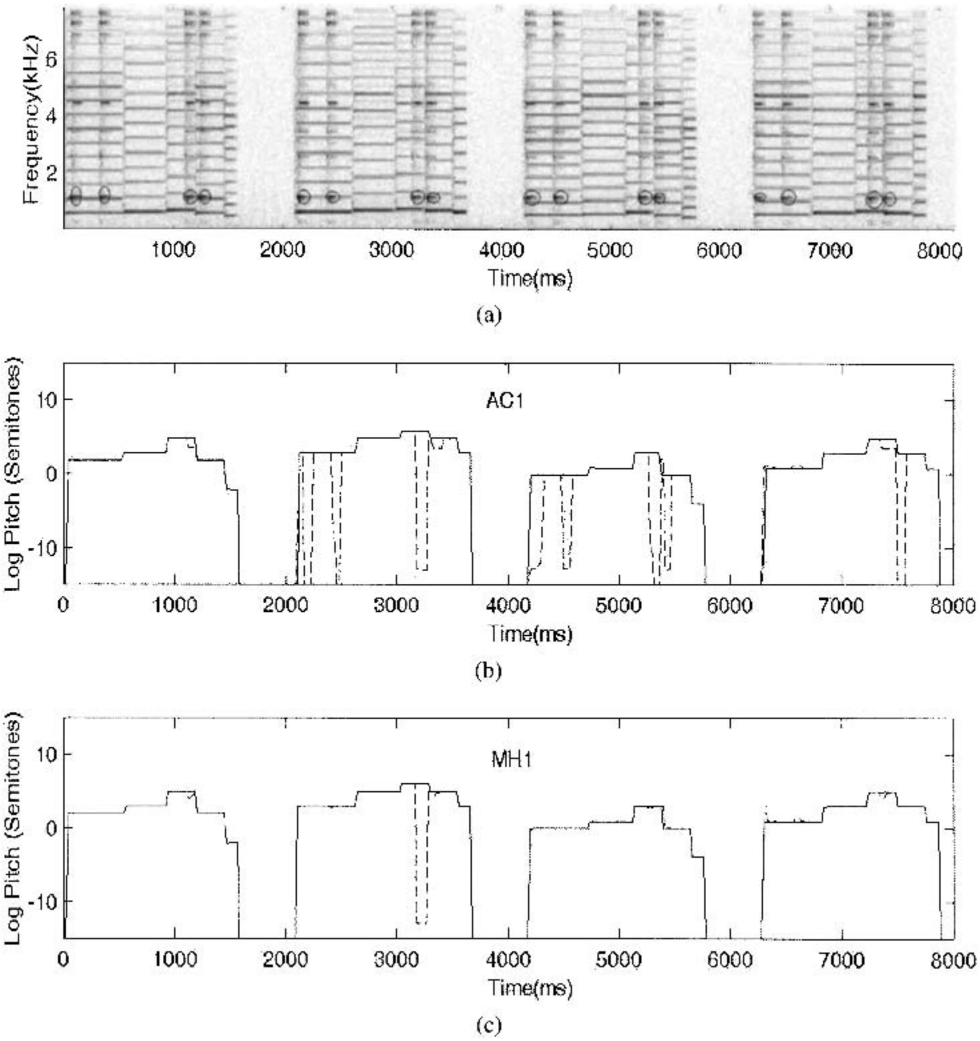
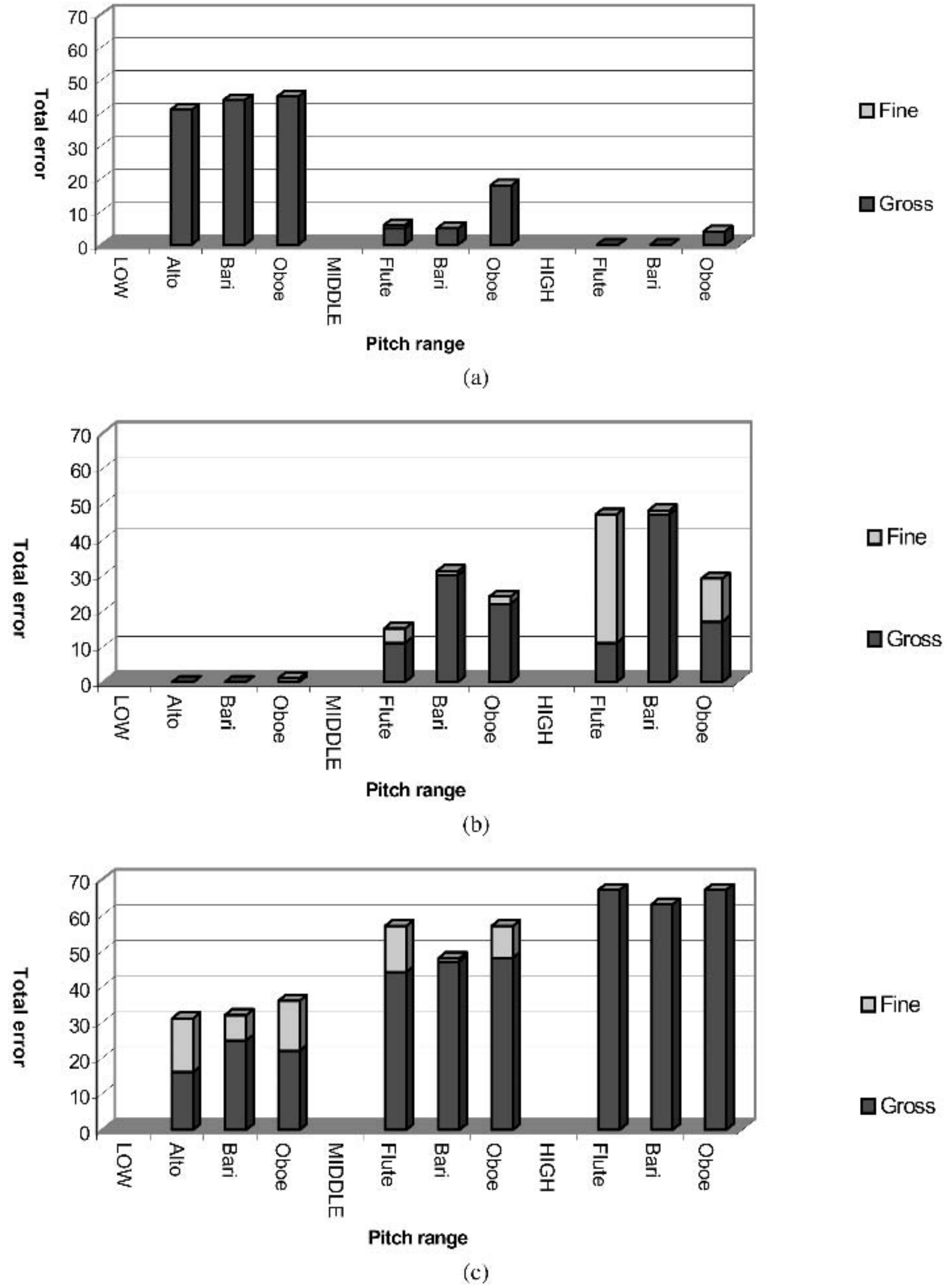


Fig. 5. Pitch estimates for test signal "middle pitch, baritone sax with low agogo." (a) Input signal spectrogram (prominent low-frequency partials of percussion encircled). (b) Actual pitch (——) and pitch estimated from AC1 PDA (- - -). (c) Actual pitch (——) and pitch estimated from MH1 PDA (- - -).

4 DISCUSSION

In each of the PDA configurations used in the experiment, the final pitch estimate (prior to postprocessing) is obtained by searching the ACF (or SACF) for the lag value corresponding to the highest peak within the expected range of lags. The ACF is computed either directly on the input signal or after nonlinear processing by the hair-cell model. In the case of the Meddis–Hewitt PDA, the hair-cell nonlinearity followed by the ACF is computed separately in frequency bands as determined by the gammatone filter bank, and then combined linearly to obtain the SACF. To understand the behavior of ACF peak-based pitch detection better it is useful to think of the ACF of a signal comprising several components (harmonic and inharmonic) as the inverse Fourier transform of the power spectrum of the signal [1]. The signal power spectrum is insensitive to the relative phases of the components, to the extent that the window is long enough that there is no significant leakage of the frequency components. Due to the linearity of the Fourier transform, the ACF of the signal is the summation of the ACFs of the individual components in the signal power spectrum, and is therefore insensitive to the phase relations between components. Based on this interpretation of the ACF, the observations of the previous section are discussed and justified via simulations using simplified implementations of channel separation (ideal bandpass filters) and hair-cell nonlinearity (a half-wave rectifier followed by a low-pass filter given by first-order Butterworth with 1-kHz cutoff frequency).



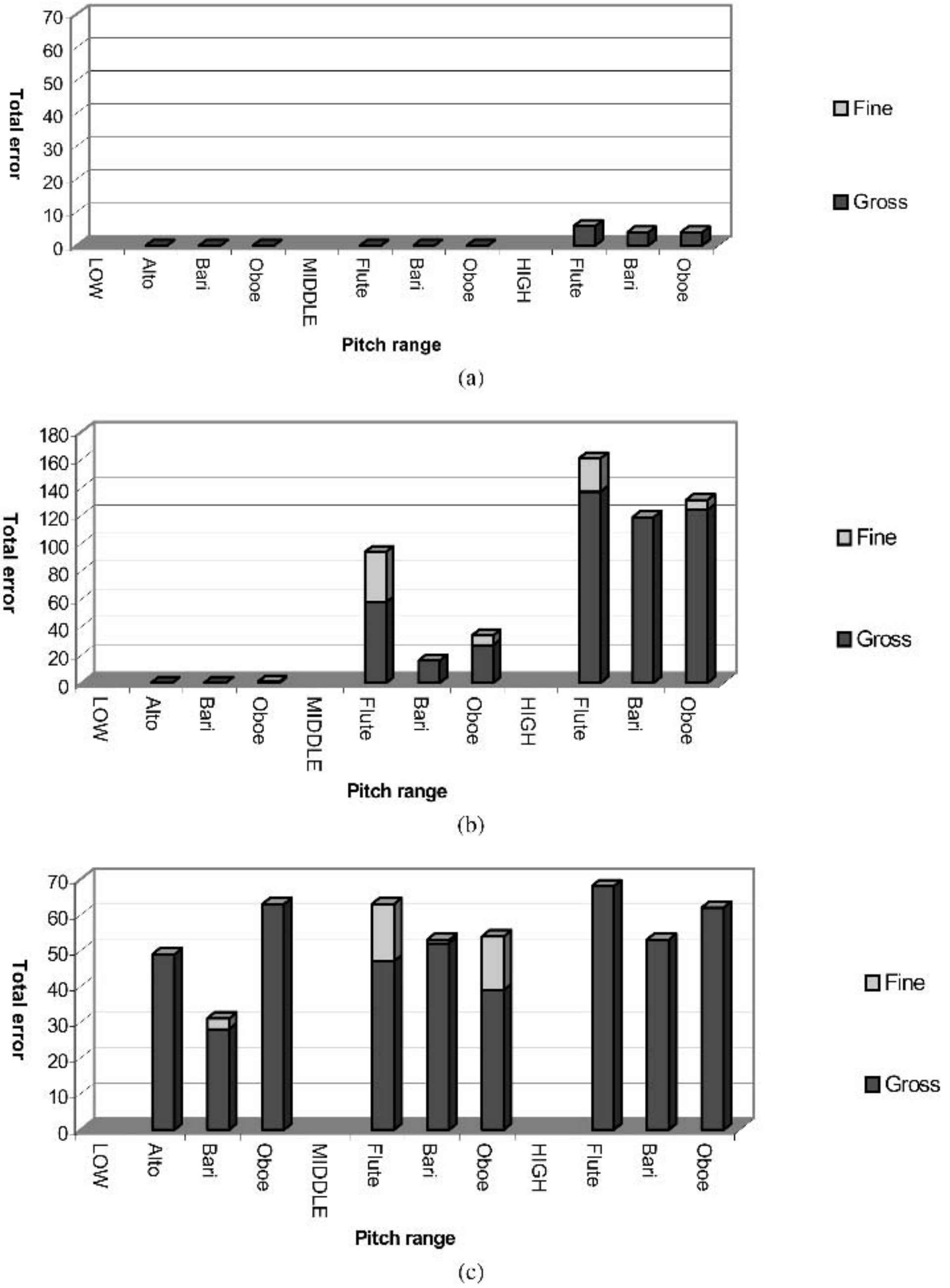
(a)

(b)

(c)

Fig. 6. Error performance of AC1 PDA for various target instruments and pitch ranges with percussion instrument in background. (a) Kick drum. (b) Hi hat. (c) Low agogo.

For the pure target signals, each of which contains a number of harmonics, including the fundamental, the windowed ACF of the input signal computed according to Eq. (1) shows peaks at lags corresponding to the pitch period and multiples of the pitch period. The highest peak corresponds to the pitch period, and there is no error in the estimated pitch. On the other hand, when the input signal to the ACF contains noise or interfering partials, there is a perturbation of the peak corresponding to the correct pitch period. The ACF of the interference partial (which can be considered to combine additively with the target ACF to form the corrupted ACF) modifies the values of the original ACF at all lags, thus modifying amplitudes at all lags to some extent. Unless the interference partial is very strong, this is not sufficient to change the locations of the prominent peaks (at pitch and pitch multiples) but affects only their relative amplitudes. As a result, the "choose the highest peak in the ACF" approach typically results in either a fine error due to a misshapen pitch peak or a gross error in the form of a pitch octave error. Fig. 10(a) shows the ACF of a periodic signal of fundamental frequency 600 Hz with the first four harmonics of amplitudes 10, 18, 14, and 12. At the sampling rate of 44.1 kHz, the signal pitch period is 73.5 samples. A single interference partial of fundamental frequency 3300 Hz and amplitude 16 (corresponding to signal-to-interference power ratio [SIR] 3.0) is added to the signal, resulting in the ACF of the noisy signal shown in Fig. 10(b). We see that the likelihood of an octave error in the ACF of the noisy signal is highest when, as depicted in Fig. 10, a valley of the ACF of the



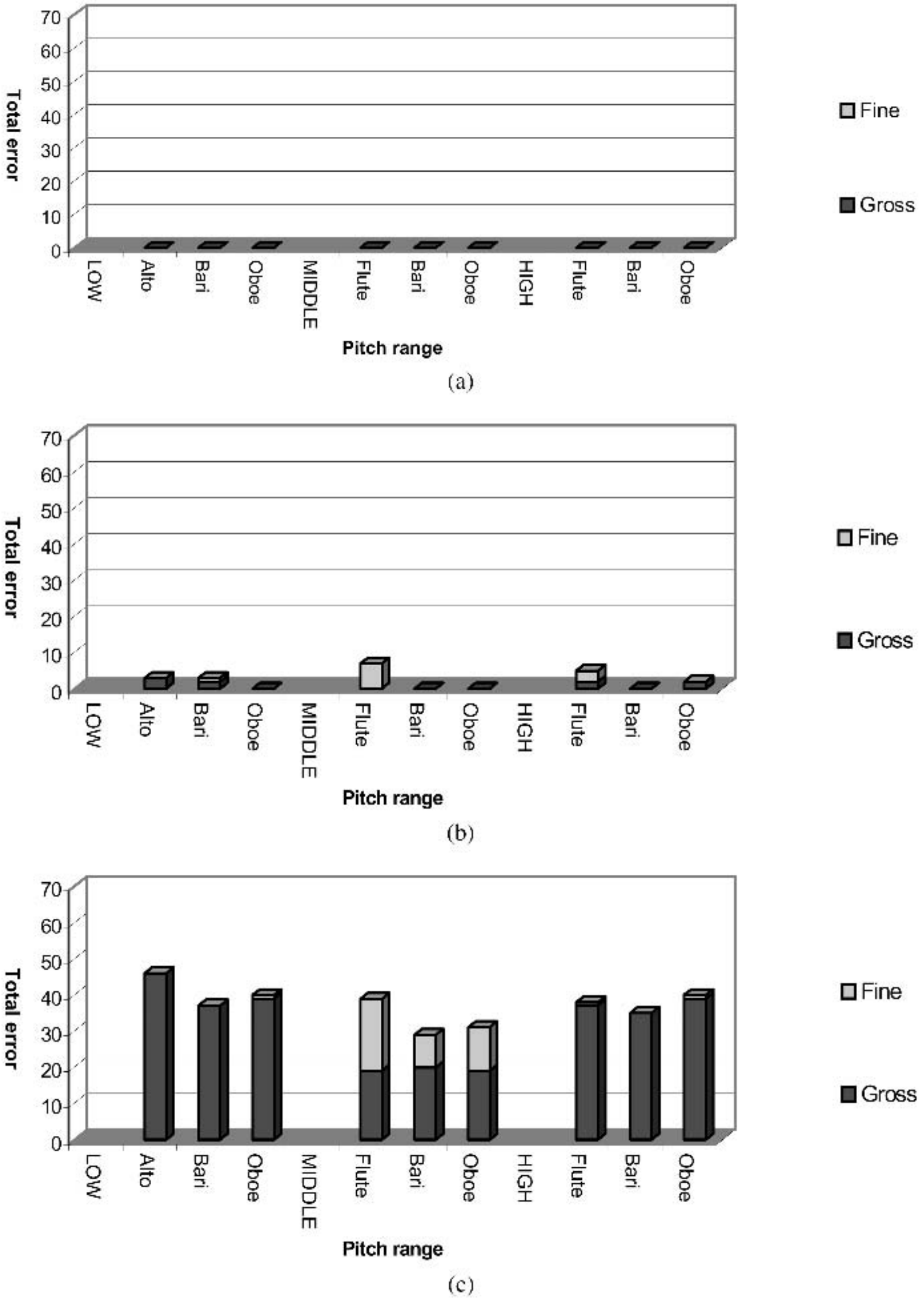
(a)

(b)

(c)

Fig. 7. Error performance of AC2 PDA for various target instruments and pitch ranges with percussion instrument in background. (a) Kick drum. (b) Hi hat. (c) Low agogo.

noise partial coincides with the signal pitch peak and the peak of the noise ACF coincides with a target pitch multiple. This is true whenever an interference partial occurs at or near an odd multiple of half the target fundamental frequency. Likewise, we expect no pitch errors when the interference partial is near a multiple of the target fundamental frequency. It is easy to see from Fig. 10 that the likelihood of pitch octave error would increase as the amplitude of the noise partial increases relative to the target signal strength. This explains why the introduction of a linear filter such as the OEM filter affecting the relative amplitudes of the signal and noise partials leads to a change in the error profiles, as seen in Fig. 7. The introduction of outer ear–middle ear filtering reduces the errors in the case of the kick drum, but has the contrary effect in the case of the hi hat and, to some extent, on the low agogo. This can be explained by the low-frequency nature of the spectrum of the kick drum, which consequently is heavily attenuated by the OEM filter. The hi hat and low agogo on the other hand have brighter spectra with much middle frequency content that remains after the OEM filter. The target spectrum, because of its preponderant lower harmonics, suffers greater overall attenuation than the bright spectra percussions. The unusually sharp rise in pitch errors in the high-pitch target range with hi-hat interference was found to be due to the chance occurrence of an interference partial at an odd multiple of half the fundamental frequency of a note of recurring pitch throughout the song. This partial fell near the resonance frequency (3 kHz) of the OEM filter and was a prominent spectral component in the filtered sig-



(a)

(b)

(c)

Fig. 8. Error performance of AC3 PDA for various target instruments and pitch ranges with percussion instrument in background. (a) Kick drum. (b) Hi hat. (c) Low agogo.

nal. It resulted in an octave error in the ACF pitch estimate almost throughout the duration of the note in question.

Introducing the hair-cell model prior to ACF computation is equivalent to a nonlinear processing of the signal that, among other effects, gives rise to new frequency components located at sum and difference frequencies of the original components. In the case of a weak or missing fundamental, the creation of distortion components contributes to the enhancement of the fundamental frequency component [1]. In addition the hair-cell model introduces a dc bias and a low-pass frequency selectivity [9]. The presence of interference partials at nonharmonic locations gives rise to nonharmonic distortion components, whose magnitudes depend on the magnitudes and phases of the interacting components (both signal and interference). Due to this the distortion components affect the peak at pitch lag in the ACF in different ways. One consistent effect is the dc level introduced by the hair-cell processing that leads to a bias favoring lower pitch lags, as seen in Fig. 10(c), where we also observe the attenuation of the high-frequency partial. Such effects contribute to the overall improvement in the performance demonstrated by the AC3 PDA in Fig. 8. In particular, a more robust pitch estimator is obtained in the case of the interference partial at a high odd multiple of half the target fundamental frequency. The Meddis–Hewitt PDA is an enhancement of the AC3 algorithm in that a cochlear filter bank is included. The ACF is computed separately in each frequency channel, and summed across channels to obtain the pitch estimate as the largest peak lag in the search range. The frequency decomposition affected by the filter bank limits the number of interacting partials through the hair-cell model nonlinearity applied separately to each channel. Fig. 11, obtained for the same signal and interference as Fig. 10, illustrates the effect of this on the SACF. Shown, for two different channel configurations (of four channels

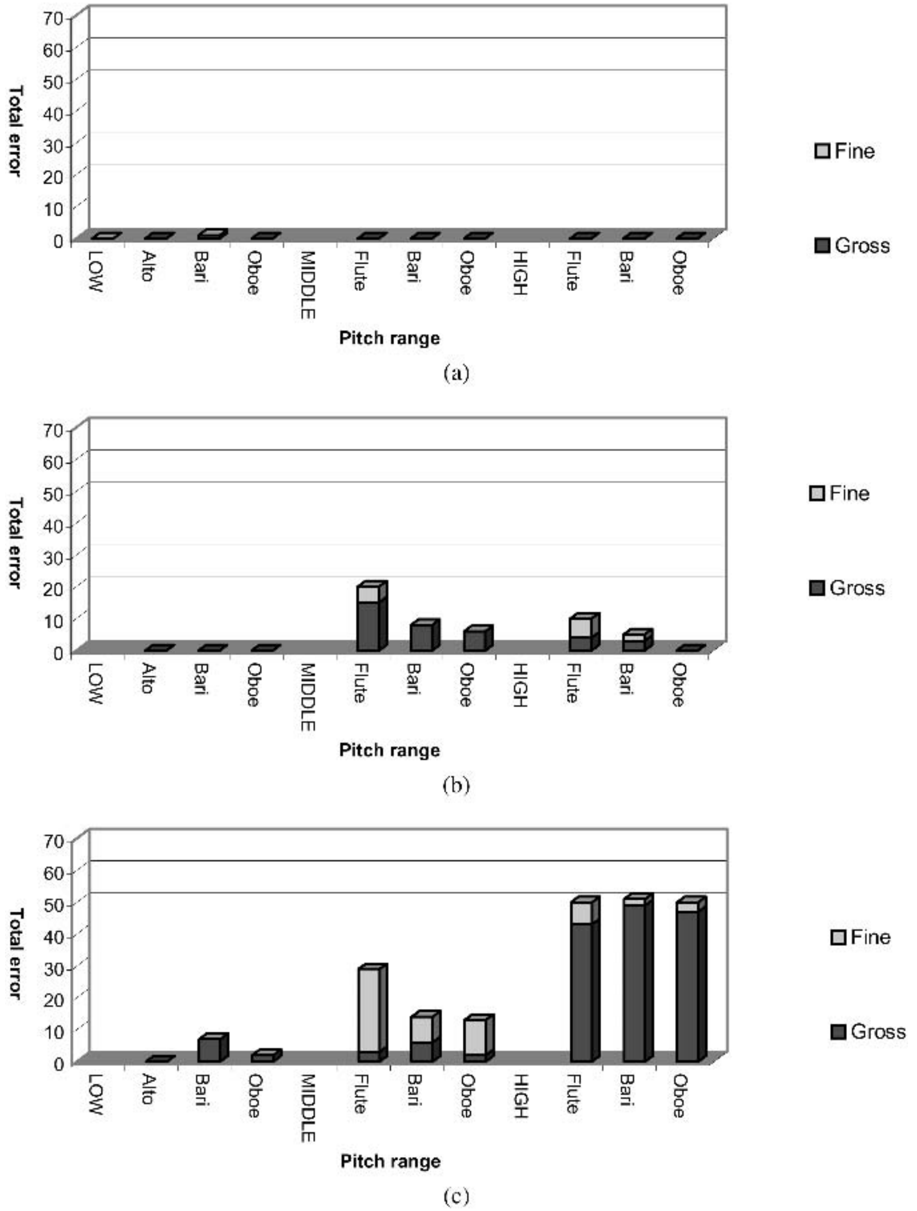


Fig. 9. Error performance of the MH1 PDA for various target instruments and pitch ranges with percussion instrument in background. (a) Kick drum. (b) Hi hat. (c) Low agogo.

simulated by ideal bandpass filters), are the signal and interference frequency components at the output of the channel nonlinearities as well as the corresponding SACF. In Fig. 11(a) the interference partial is in a separate channel by itself. This eliminates any distortion components created due to an interaction of target harmonics and interference. On the other hand, the co-occurrence of several higher harmonics of the target in a single channel strengthens the fundamental frequency component in the SACF. These two effects lead to an improved ACF peak at the signal pitch period of 73 samples. This explains the improved performance of the Meddis–Hewitt PDA for the low agogo samples for the low- and middle-pitch ranges. In the high-pitch range, however, it was observed that due to the higher interharmonic spacing, several of the channels contained only a single harmonic of the target instrument accompanied by interference components. This condition is depicted and simulated by the configuration of Fig. 11(b), where the signal partials occupy different channels and the noise partial shares a channel with a target harmonic. The last channel gives rise to inharmonic distortion components, one of which is visible in the figure. Together with the reduced contribution to the fundamental frequency due to the absence of unresolved harmonics, this leads to a degradation of the pitch estimate.

Finally we return to the AC1 PDA and explain the low-frequency errors due to the kick drum in Fig. 6. The autocorrelation method of AC1 leads to a significant number of gross errors for all three percussions, depending on the frequency relation between target and interference partials. The low-pitch-range errors are the most pronounced in the case of the kick drum due to a strong low-frequency partial from this percussion. Fig. 12 illustrates this effect on ACF peak-based pitch estimation by simulating the kick drum by a strong interference tone at 68 Hz. Shown in Fig. 12 are ACFs for signals of fundamental frequency 600 Hz and 200 Hz with the same harmonic amplitudes as the signal of Fig. 10. Both clean signals yield accurate pitch peaks in the SACF (at lags of 73 samples and 220 samples, respectively). However, the addition of the 68-Hz low-frequency interference tone (with amplitude 26, corresponding to SIR 1.1) introduces a low-lag bias in the overall ACF in both cases. This leads to a gross pitch error (pitch submultiple selected) in the case of the lower fundamental frequency signal since its pitch period is comparable to that of the interference.

5 CONCLUSIONS

In this engineering report an experimental investigation is presented of the performance of pitch-detection algorithms based on temporal autocorrelation for the pitch tracking of a melodic signal with percussive accompaniment characterized by inharmonic partials. The performance of the autocorrelation pitch detector as well as its enhancements based on the Meddis–Hewitt auditory model are studied experimentally on synthetic musical signals. The ACF peak-based pitch detector incurs pitch es-

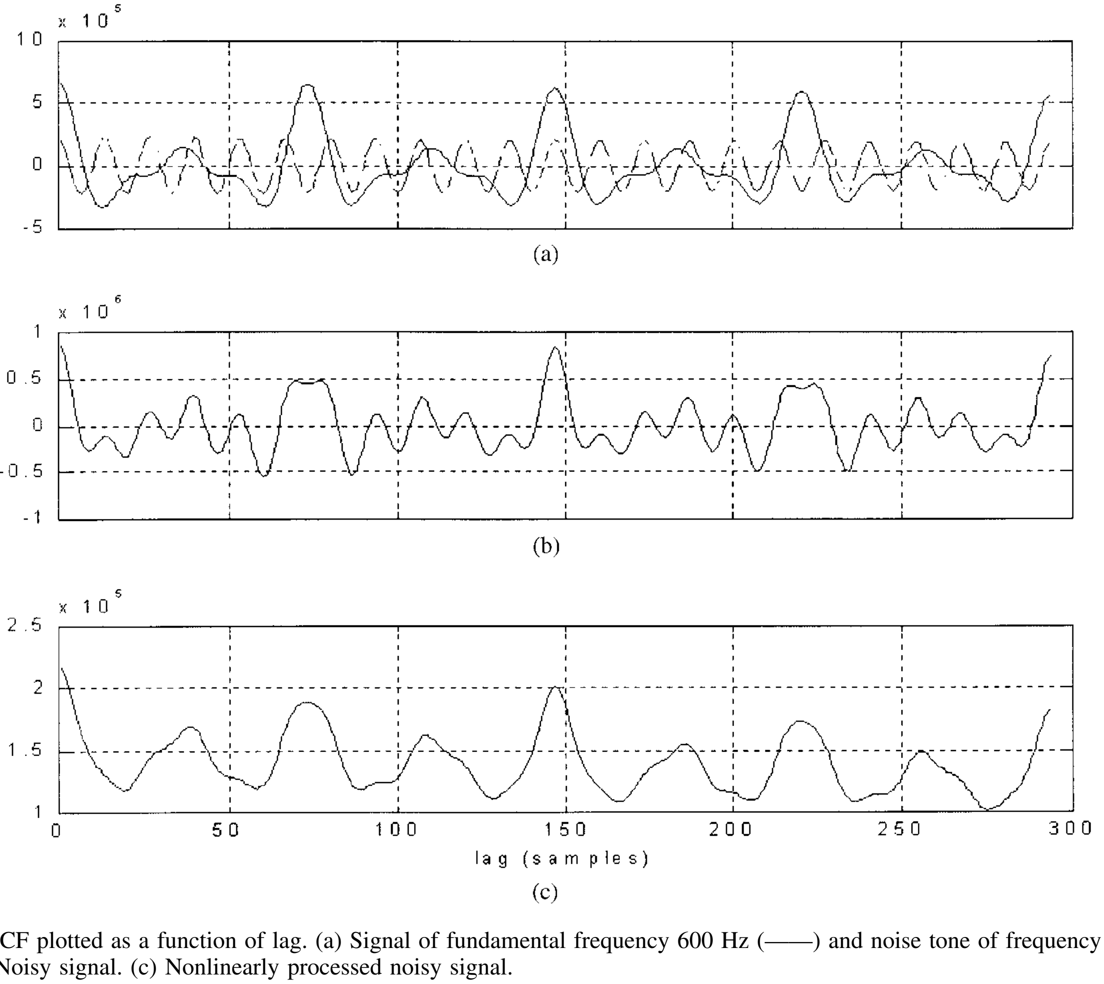
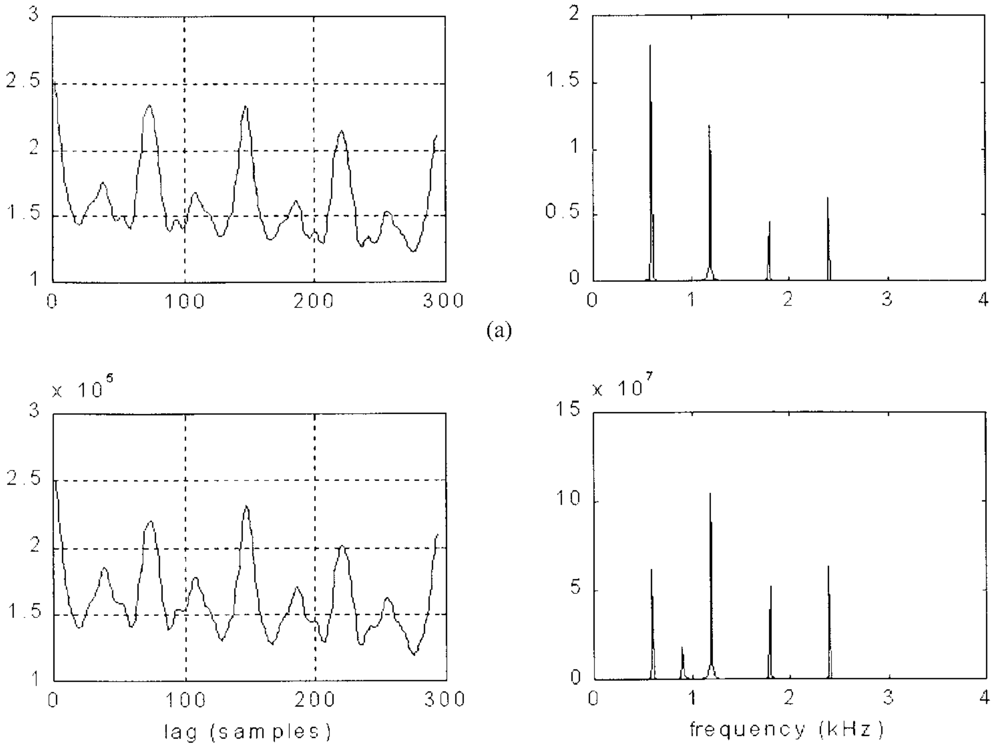


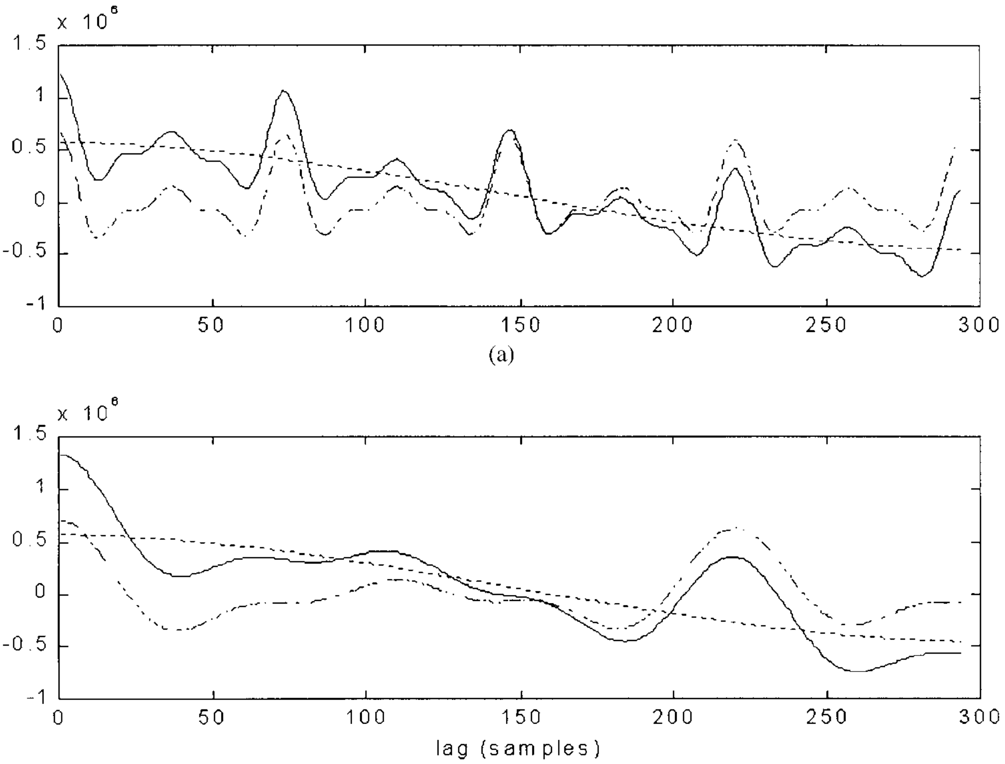
Fig. 10. ACF plotted as a function of lag. (a) Signal of fundamental frequency 600 Hz (——) and noise tone of frequency 3300 Hz (---). (b) Noisy signal. (c) Nonlinearly processed noisy signal.



(a)

(b)

Fig. 11. SACF and spectral components of noisy signal after channel filtering and nonlinear processing corresponding to different four-channel groupings of signal harmonics and noise. (a) SACF for **/h1/h2/h3+h4/n/** and corresponding power spectrum. (b) SACF for **/h1/h2/h3/h4+n/** and corresponding power spectrum.



(a)

(b)

Fig. 12. ACF plotted as a function of lag for signal and interference tone of 68 Hz. - · - · - ACF of signal; --- ACF of interference; —— ACF of noisy signal. (a) Signal fundamental frequency 600 Hz. (b) Signal fundamental frequency 200 Hz.

timation errors when interference partials cooccur with signal harmonics. The following enhancements to the basic ACF PDA improve the robustness of pitch extraction in the presence of percussive interference in the form of inharmonic tones. Outer-ear-middle-ear prefiltering, cochlear bandpass filtering, and hair-cell nonlinear processing represent a combination of linear and nonlinear preprocessing of the signal before computing the ACF to estimate pitch periodicity. The sum of channel autocorrelations (SACF) would simply be proportional to the autocorrelation of the input signal to the cochlear filter bank if it was not for the hair-cell nonlinearity [12]. Of significance then is the combined role of channel separation and hair-cell nonlinearity. The experimental results provide important insights into the nature of the pitch errors and their dependence on the relative frequencies of the signal and interference. The noise sensitivity of ACF peak-based pitch detection is highest when interfering partials fall exactly between signal harmonics. The hair-cell nonlinearity serves to increase the accuracy of pitch detection primarily via an increased bias of lower lag peaks in the ACF and the attenuation of high-frequency partials. The separate processing of frequency channels by the introduction of the cochlear filter bank is crucial in reducing the distortion components from the interaction of signal harmonics and interference partials while reinforcing the contribution to the fundamental frequency component from unresolved higher harmonics of the signal.

It would be interesting to explore the perceptual implications of the preceding observations. That is, since the observations on the accuracy of the Meddis–Hewitt PDA in the current context are explained by the signal-processing algorithms used, it is relevant to wonder whether these specific predictions of the model hold for subjective pitch perception. From a practical viewpoint, the results of this study may be applied to construct pitch detectors for musical signals that are robust to the presence of nonpitched percussion. It is interesting to consider tuning the several available parameters of the hair-cell model nonlinearity to increase its effectiveness for known signal and interference characteristics, including possibly the presence of broad-band noise. Alternative means of obtaining the SACF, such as summing with channel weighting, provide further promising directions for future work.

6 REFERENCES

[1] W. Hess, *Pitch Determination of Speech Signals* (Springer, New York, 1983).

[2] J. D. Wise, J. R. Capiro, and T. W. Parks, "Maximum Likelihood Pitch Estimation," *IEEE Trans. Acoustics, Speech, Signal Process.,* vol. ASSP-24 (1976 Oct.).

[3] G. Monti and M. Sandler, "Monophonic Transcription with Autocorrelation," in *Proc. COST G-6 Conf. on Digital Audio Effects, DAFX-00* (Verona, Italy, 2000).

[4] D. J. Hermes, "Pitch Analysis," in *Visual Representations of Speech Signals,* M. Cooke, S. Beet, and M. Crawford, Eds. (Wiley, New York, 1993).

[5] R. Meddis, and M. J. Hewitt, "Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery. I: Pitch Identification," *J. Acoust. Soc. Am.,* vol. 89 (1991 June).

[6] M. Pflueger, R. Höldrich, and W. Riedler, "A Nonlinear Model of the Peripheral Auditory System," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 45, p. 1002 (1997 Nov.), preprint 4510.

[7] "HUTear—Matlab Toolbox for Auditory Modeling," available at www.acoustics.hut.fi/software/HUTear/.

[8] A. Klapuri, "Wide-Band Pitch Estimation for Natural Sound Sources with Inharmonicities," presented at the 106th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 47, p. 520 (1999 June), preprint 4906.

[9] R. Meddis, "Simulation of Mechanical to Neural Transduction in the Auditory Receptor," *J. Acoust. Soc. Am.,* vol. 79 (1986, Mar.).

[10] R. Meddis, M. Hewitt, and T. Shackleton, "Implementation Details of a Computation Model of the Inner Hair-Cell/Auditory-Nerve Synapse," *J. Acoust. Soc. Am.,* vol. 87 (1990 Apr.).

[11] M. Slaney, Auditory Toolbox, version 2, Interval Research Corporation.

[12] P. Cariani, M. Tramo, and B. Delgutte, "Neural Representation of Pitch through Temporal Autocorrelation," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 45, pp. 1021–1022 (1997 Nov.), preprint 4583.

THE AUTHORS

P. Rao

S. Shandilya

Preeti Rao received a Bachelor degree in Electrical Engineering from the Indian Institute of Technology, Bombay, in 1984, and a Ph.D. degree specializing in signal processing from the University of Florida, Gainesville, in 1990. She taught in the Electrical Engineering department at the Indian Institute of Technology, Kanpur from 1994 to 1999. Following a six-month visiting position at the Institute of Perception Research, Eindhoven, The Netherlands, she joined the Indian Institute of Technology, Bombay, where she is presently an associate professor. Her current research interests include speech and audio signal compression and audio content retrieval.

●

Saurabh Shandilya received a B.E. degree in electrical engineering from the Government Engineering College, Bilaspur, India, in 2001 and an M.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay in 2003. Currently he works for Neomagic Semiconductors Inc., Noida, India. His interests include speech and audio processing, video coding, and associative computing.