Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 841–856

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Multi-band summary correlogram-based pitch detection for noisy speech

Lee Ngee Tan *, Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA

Received 8 November 2012; received in revised form 19 February 2013; accepted 8 March 2013
Available online 28 March 2013

Abstract

A multi-band summary correlogram (MBSC)-based pitch detection algorithm (PDA) is proposed. The PDA performs pitch estimation and voiced/unvoiced (V/UV) detection via novel signal processing schemes that are designed to enhance the MBSC's peaks at the most likely pitch period. These peak-enhancement schemes include comb-filter channel-weighting to yield each individual subband's summary correlogram (SC) stream, and stream-reliability-weighting to combine these SCs into a single MBSC. V/UV detection is performed by applying a constant threshold on the maximum peak of the enhanced MBSC. Narrowband noisy speech sampled at 8 kHz are generated from Keele (development set) and CSTR – Centre for Speech Technology Research-(evaluation set) corpora. Both 4-kHz full-band speech, and G.712-filtered telephone speech are simulated. When evaluated solely on pitch estimation accuracy, assuming voicing detection is perfect, the proposed algorithm has the lowest gross pitch error for noisy speech in the evaluation set among the algorithms evaluated (RAPT, YIN, etc.). The proposed PDA also achieves the lowest average pitch detection error, when both pitch estimation and voicing detection errors are taken into account.

Keywords: Pitch detection; Multi-band; Correlogram; Comb-filter; Noise-robust

1. Introduction

Pitch or fundamental frequency (F0) detection is important for many speech applications. These applications include speech enhancement, synthesis, coding, source separation, and auditory scene analysis. Due to the increasing use of mobile devices, speech communication seldom takes place in a noise-free environment. Although pitch detection is a well-researched subject, and existing pitch detection algorithms (PDAs) work reasonably well for clean speech, accurate pitch detection for bandpass-filtered speech (e.g., telephone speech with G.712 filter characteristics (ITU, 1996)), and noisy speech still poses a challenge.

A pitch detector performs both pitch estimation and voiced/unvoiced (V/UV) detection. In pitch estimation, the rate of vocal-fold vibration is estimated, while in V/UV detection, voiced or quasi-periodic speech frames are distinguished from the rest of the signal. In general, pitch estimation can be performed using (1) time-domain, (2) frequency-domain, or (3) time-frequency-domain signal processing techniques. Time-domain pitch estimation exploits the signal's temporal periodicity by computing a temporal correlation or difference function directly from the signal samples. Some well-known examples of time-domain pitch estimation algorithms are RAPT (Talkin, 1995), YIN (Cheveigné and Kawahara, 2002), and the average magnitude difference function (AMDF) pitch extractor (Ross et al., 1974), which are known to give accurate pitch estimates for clean speech. Frequency-domain pitch estimation relies on the presence of strong harmonic peaks near integer multiples of F0 in the short-time spectral

* Corresponding author. Address: Electrical Engineering Department, University of California, Los Angeles, 56-125B Engineering IV Building, Box 951594, Los Angeles, CA 90095, USA. Tel.: +1 310 729 1135.
E-mail addresses: ngee@seas.ucla.edu (L.N. Tan), alwan@ee.ucla.edu (A. Alwan).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.03.001

representation. Some examples of such frequency-domain pitch estimation algorithms are subharmonic-to-harmonic ratio (SHR) (Sun, 2002), dominant harmonics (Nakatani and Irino, 2004), and SWIPE' (SWIPE' is a variant of SWIPE that focuses on harmonics at prime integer multiples of F0) (Camacho and Harris, 2008). In time-frequency-domain pitch estimation algorithms, the input signal is typically decomposed into multiple frequency subbands, and time-domain techniques are applied on each subband signal. A popular time-frequency-domain technique is the auditory-model correlogram-based algorithm inspired by Licklider's duplex theory of pitch perception (Licklider, 1951), in which frequency decomposition is performed using an auditory filterbank (for which gammatone filterbanks (Patterson et al., 1992) are widely used), followed by autocorrelation (ACR) computation on each subband signal. The correlogram is formed by vertically stacking all ACR functions to form a 2-D image (Slaney and Lyon, 1990). Finally, the fundamental period (T0) of the signal is found by locating the ACR delay lag of the maximum peak in the "summary" correlogram (SC), which is typically the averaged ACR function. ACR can be applied either directly on the subband signal or its envelope. The latter is usually performed on mid- and high-frequency subbands only (Rouat et al., 1997; Wu et al., 2003). These subbands have sufficiently wide bandwidths to capture at least two consecutive harmonic peaks, such that the resulting filtered signals have an amplitude modulation frequency equal to F0 (a.k.a. beat frequency) (Delgutte, 1980). It has been shown that correlogram-based techniques can yield estimates close to human's perceived pitch for difficult signals with missing fundamental, inharmonic complexes and noise tones (Meddis and Hewitt, 1991; Cariani and Delgutte, 1996). Being a multi-band approach, correlogram-based techniques also tend to be more noise-robust than time-domain or frequency-domain algorithms whose parameters are fixed regardless of the signal's periodicity in the different subbands. This is because additional subband selection or weighting schemes, such as those in Rouat et al. (1997) and WWB (Wu et al., 2003), can be implemented to give less emphasis to the noise-dominated subbands. Since the filters in a gammatone filterbank are narrower and spaced more closely at lower linear frequencies than at higher frequencies (Patterson et al., 1992), the number of filters at lower frequencies (within the first 1 kHz) can be almost equivalent to the number filters in the mid and high frequencies. When the majority of harmonics at the low frequencies are attenuated due to the transmission channel characteristics or masked by strong low-frequency noise interference, it is challenging to design an effective subband selection and weighting scheme to select the reliable subband ACRs such that the maximum peak of the resulting summary correlogram yields the true pitch value.

As for V/UV detection in a pitch detector, it can be performed by either utilizing the information derived from the pitch estimation module, or using a separate module that is independent of the pitch estimation algorithm. The simplest V/UV detector is one that applies a constant decision threshold on a single degree-of-voicing feature computed by the pitch estimation module, e.g., ACR or cepstral peak amplitudes (Rabiner et al., 1976). To further improve detection accuracy, the initial V/UV decisions are usually smoothed via median filtering (Secrest and Doddington, 1982; Ahmadi and Spanias, 1999). A disadvantage of the constant-threshold scheme is that since the degree-of-voicing feature tends to be very noise-sensitive and dependent on the signal-to-noise ratio (SNR), a threshold level tuned for a particular SNR, generally does not work well at a different SNR. Thus, threshold adaptation techniques have been proposed to improve the noise-robustness of V/UV detectors. Typically, the threshold is adapted based on long-term statistics (min, max, mean, median, etc.) of degree-of-voicing-related features (Medan et al., 1991; Ahmadi and Spanias, 1999). In this case, V/UV detection performance would tend to degrade under a highly non-stationary noise condition. Dynamic programming – a tracking algorithm that integrates V/UV detection with pitch estimation, is another common technique used for pitch detection (Secrest and Doddington, 1983; Talkin, 1995; Luengo et al., 2007). A dynamic programming algorithm finds the least-cost path based on some pre-defined voicing and frequency transition cost functions, leading to performance improvements in both V/UV detection and pitch estimation through utilizing voicing and pitch information from multiple frames. However, when a constant value is used to control the voicing transition cost, such as the voice bias in Talkin (1995), the V/UV detection performance of these pitch detectors is also dependent on SNRs. Since it is generally difficult to perform noise-robust V/UV detection based on the single degree-of-voicing feature from pitch estimation (Atal and Rabiner, 1976), statistically-trained V/UV classifiers have also been proposed, especially for applications that do not require pitch estimates to be computed (e.g., speaker-independent speech recognition). This latter class of V/UV detectors, which can operate independently from pitch estimation, have reported robust V/UV detection performance, especially if their parameters are learned from noisy speech (Shah et al., 2004; Beritelli et al., 2007). In this paper, since pitch estimation is already part of a pitch detector, we are mainly interested in the former class of V/UV detectors. There are also algorithms that perform pitch-tracking using models trained on information extracted during pitch estimation. For example, hidden Markov models (HMMs) are used in WWB (Wu et al., 2003) to form continuous single or dual pitch contours for noisy speech. These data-driven algorithms yield robust voicing/pitch detection performance when the test data has characteristics that are similar to the data used for training the models.

In this paper, a multi-band summary correlogram (MBSC)-based pitch detection algorithm is proposed. This work is an extension of our previous algorithm in Tan and Alwan (2011) that has focused on pitch estimation only.

Note that the proposed MBSC PDA in this paper performs single-pitch detection of target speech in noise. To improve the noise-robustness of V/UV detection, novel signal processing schemes are included to enhance the maximum peak in the MBSC – the degree-of-voicing feature used for V/UV detection. The sensitivity of this voicing feature to noise is reduced, such that good V/UV detection performance is achievable with a constant threshold and median filtering scheme. Babble, car, and machine gun noises are artificially added to narrowband speech sampled at 8 kHz from Keele (Plante et al., 1995) (development set) and Centre for Speech Technology Research (CSTR) (Bagshaw et al., 1993) (evaluation set) corpora. In addition to the 4-kHz fullband version, the G.712-filtered (ITU, 1996) bandpass version of the noisy speech is also generated for performance evaluation. When compared to widely-used algorithms, and those employing similar processing techniques, the proposed MBSC PDA outperforms these comparative algorithms for the majority of the conditions evaluated.

The organization of this paper is as follows. The technical details of the MBSC PDA are explained in the next section. Section 3 describes the corpora, performance measures and comparative algorithms used to evaluate pitch estimation and pitch detection performances. Section 4 contains the results and discussions on the pitch estimation performance evaluation, while the results and discussions on the pitch detection performance evaluation are found in Section 5. Section 6 summarizes and concludes this article.

2. Proposed MBSC pitch detector

The block diagram in Fig. 1 gives an overview of the proposed MBSC pitch detector. It consists of six main signal processing stages, which are explained in the following subsections.

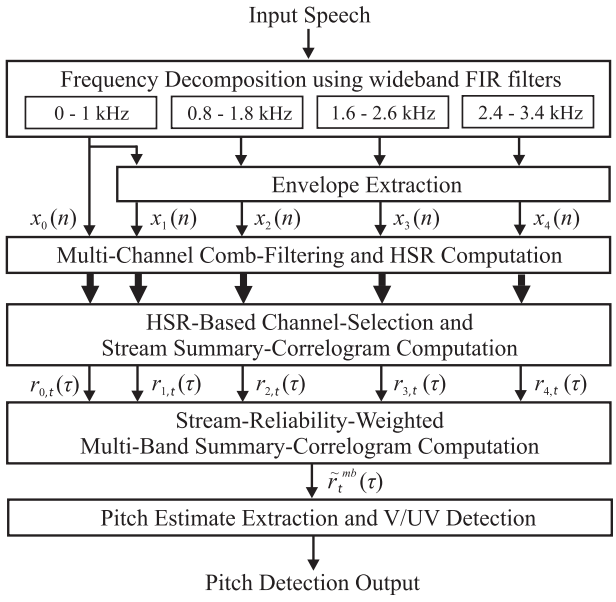

Fig. 1. Block diagram of the proposed pitch detector. Multi-channel outputs are indicated by bold arrows.

2.1. Frequency decomposition using wideband FIR filters

The input speech signal is first decomposed into four subbands using 32-point FIR filters. Filter cut-off frequencies are shown in Fig. 1. A 1-kHz filter bandwidth is chosen so that at least two harmonics are captured by each filter, for example, capturing harmonics at 400 Hz and 800 Hz in the first 1 kHz subband for an F0 of 400 Hz – the maximum pitch value of interest in our defined adult pitch range (50–400 Hz). When a signal contains more than 1 harmonic of the target voiced speech, its envelope would typically oscillate at an amplitude modulation frequency corresponding to the inter-harmonic separation. A 0.2 kHz overlap between adjacent filters results in a spectrum coverage up to 3.4 kHz with four such filters, which corresponds to the upper cut-off frequency of the G.712-characteristic filter (ITU, 1996). In our PDA, wideband FIR filters are used instead of a gammatone filterbank (e.g., in Rouat et al. (1997) and Wu et al. (2003)), so that useful pitch information residing in the low-frequency envelope can be exploited. In our previous pitch estimation algorithm in Tan and Alwan (2011), we showed that inclusion of the low-frequency band signal envelope is especially useful in improving pitch estimation accuracy for speech whose low-frequency harmonics are severely attenuated due to G.712-filtering in telephone speech. In addition, FIR filters that are equally-spaced in the linear frequency domain (instead of a warped frequency scaling) are used, to give equal emphasis to all the harmonics in the available frequency range. The noise-robustness of the algorithm should increase with the number of FIR filters up to a certain limit, but with a higher cost in computational complexity. We found that four FIR filters are sufficient to give a relatively good pitch detection performance for the development set used.

2.2. Envelope extraction per subband

The Hilbert envelope (Loughlin and Tacer, 1996) in each subband is extracted by computing the magnitude squared of the analytic signal, which is obtained by applying Hilbert transform on the FIR-filtered outputs. It is noted in Drullman (1995) that the Hilbert envelope accurately follows the amplitude modulations of a bandpass signal. The Hilbert envelope is mean-normalized on a frame-by-frame basis before subsequent processing. The four mean-normalized envelope streams are denoted as $x_s(n)$, where $s = 1$, 2, 3, and 4 is the stream index, and n refers to the sample index. The lowpass-filtered, non-envelope stream from the first subband, which we labeled as $x_0(n)$, is also used in subsequent processing (see Fig. 1). This non-envelope stream contains valuable information for pitch detection, especially when the first harmonic is not attenuated or noise-corrupted, because the

stream tends to be more periodic than the envelopes whose periodicities are more affected by signal energy variations.

2.3. Multi-channel comb-filtering per stream

Multi-channel comb-filtering is performed in the frequency domain, as shown in Eq. (1), by multiplying the input stream spectrum, $X(f)$, with a comb-function, $c(f)$, represented in the frequency domain. $X_{k,s,t}(f)$ and $Y_{k,s,t}(f)$ denote the complex discrete Fourier transform (DFT) coefficients of $x_s(n)$ in frame t, and its comb-filtered version, respectively. The comb-functions are formulated using raised-cosines, as shown in Eq. (2), where $c_k(f)$ is the kth channel's comb-function with an inter-peak frequency of F_k Hz, and f represents frequency. This comb-function enhances spectral harmonics spaced F_k apart, and suppresses the energies at the subharmonics. The raised-cosine function is selected due to its broad spectral peak lobes and smooth peak-to-valley transitions. By selecting a smooth comb-function (Camacho and Harris, 2008), a slight signal inharmonicity (harmonic frequency perturbation from multiples of F0) would not result in sharp energy attenuation, in comparison to using an impulse-train-like comb-function (Hermes, 1988; Sun, 2002), when the signal is filtered with a comb-function whose F_k is close but not exactly equal to the true F0. To reduce the dependency of the maximum ACR peak amplitude on the signal's fundamental period (ACR is applied on comb-filtered signals in a subsequent processing stage), N_k, the number of samples in $x_s(n)$ that is used to compute $X_{k,s,t}(f)$ is made proportional to the time periodicity enhanced by $c_k(f)$, i.e., $T_k = f_s/F_k$ samples, where f_s is the sampling frequency. It is found that $N_k = 4T_k$ samples are sufficient to achieve good pitch estimation and V/UV detection performance at low SNRs, but is not too large as to severely degrade the resolution of V/UV boundary detection at high SNRs. To reduce the number of unique $X_{k,s,t}(f)$ computations, N_k is quantized to the multiple of 40 samples that is closest to $4T_k$.

$$Y_{k,s,t}(f) = \begin{cases} X_{k,s,t}(f)\, c_k^d(f), & \text{if } s = 0 \\ X_{k,s,t}(f)\, c_k(f), & \text{if } s = 1, 2, 3, \text{ or } 4 \end{cases} \tag{1}$$

$$c_k(f) = \begin{cases} \frac{1+\cos(2\pi f/F_k)}{2}, & \text{if } 0.5F_k \leqslant f \leqslant 1 \text{ kHz} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$c_k^d(f) = d_k(f)\, c_k(f) \tag{3}$$

$$d_k(f) = 1 + \min(1, F_k/f) \tag{4}$$

From Eq. (1), it can be seen that constant-magnitude comb-functions – $c_k(f)$, are applied on the envelope streams, $s \geqslant 1$; while decreasing-amplitude comb-functions – $c_k^d(f)$, are used on the non-envelope stream, s=0. We obtain $c_k^d(f)$ by multiplying $c_k(f)$ with a decreasing function, $d_k(f)$, as shown in Eqs. (3) and (4). In $c_k^d(f)$, the first harmonic has the largest gain of 2, while the gains of the higher harmonics decrease towards 1. Filtering $x_0(n)$ with a $c_k^d(f)$ whose F_k is near the true pitch value would boost the strength of the first harmonic relative to the higher ones. This helps reduce over-estimation errors in the presence of a very strong non-fundamental harmonic, such as in G.712-filtered telephone speech. Boosting the first harmonic in $x_{s\geqslant 1}(n)$ is unnecessary, since the first harmonic of a Hilbert envelope is usually stronger than its higher harmonics. To reduce subharmonic estimation errors, decreasing-amplitude comb-functions have also been used in other comb-filter-based pitch estimation algorithms (Hermes, 1988; Camacho and Harris, 2008). However, these comb-filter-based algorithms apply only one group of comb-filters (with various peak intervals) that spans the same frequency range. Thus, these algorithms depend heavily on the prominence of the lower frequency harmonics to obtain an accurate pitch estimate, and these harmonics can be masked by noise or attenuated in bandpass-filtered speech (e.g., telephone, hand-held radios (Walker and Strassel, 2012), etc.). By incorporating separate comb-filters that span different subbands, the pitch detection performance of our PDA is more robust to such distortions.

2.4. HSR-based comb-channel selection per stream

2.4.1. HSR computation

Harmonic-to-subharmonic energy ratio (HSR) is a measure computed to aid the selection of reliable comb-filter channels per stream. Comb-filters defined in Section 2.3 capture the harmonic energy of the signal. To capture the subharmonic or inter-harmonic energy, inverted comb-filters, $c_k^-(f)$ and $c_k^{d-}(f)$, are designed to pair with each $c_k(f)$ and $c_k^d(f)$, as shown in Eqs. (5) and (6). The HSR of the kth channel in stream s, denoted by $q_{s,t}(k)$, is computed using Eq. (7). For a comb-function whose inter-peak frequency, F_k, is close to the input signal's true pitch value, its HSR would be high.

$$c_k^-(f) = \begin{cases} 1 - c_k(f), & \text{if } 0.5F_k \leqslant f \leqslant 1 \text{ kHz} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$c_k^{d-}(f) = d_k(f)\, c_k^-(f) \tag{6}$$

$$q_{s,t}(k) = \begin{cases} \sum_f |X_{k,s,t}(f)\, c_k^d(f)|^2 / \sum_f |X_{k,s,t}(f)\, c_k^{d-}(f)|^2, & \text{if } s = 0 \\ \sum_f |X_{k,s,t}(f)\, c_k(f)|^2 / \sum_f |X_{k,s,t}(f)\, c_k^-(f)|^2, & \text{if } s = 1, 2, 3, 4 \end{cases} \tag{7}$$

In the presence of noise, one frame of speech might be more corrupted than its neighboring frames, such that inter-frame peak consistency in $q_{s,t}(k)$ is affected. Since the pitch contour of natural speech generally varies smoothly in time, we applied a lowpass IIR filter on $q_{s,t}(k)$ to improve its inter-frame peak consistency, as shown in Eq. (8). This time-smoothed HSR is denoted by $\tilde{q}_{s,t}(k)$.

$$\tilde{q}_{s,t}(k) = 0.5\,\tilde{q}_{s,t-1}(k) + 0.5\, q_{s,t}(k) \tag{8}$$

2.4.2. Tri-stage channel selection

The proposed PDA does not use the computed HSR directly for pitch estimation, because the HSR computed at a subharmonic can sometimes be as high as or higher than the HSR at an F_k near the true F0. Instead, the HSR is used to identify reliable comb-channels in each stream. Channel selection is performed on a per stream basis (i.e., channels selected in stream $s = 0$ are independent of those selected in other streams), using a novel tri-stage selection process designed to improve both pitch estimation and voicing detection performance. The flowchart of this tri-stage channel selection is shown in row (I) of Fig. 2, while row (II) shows the HSR, $\tilde{q}_{s,t}(k)$, for a particular $x_{s,t}(n)$ in blue, with the channels selected in each stage indicated by the red asterisks. Row (III) of Fig. 2 contains the ACRs computed from the comb-filtered outputs of channels selected in Stage 2. For ease of visualization and comprehension of the selection algorithm described subsequently, F_k instead of channel index k, is used for labeling the horizontal axes in the HSR plots, and a "clean" voiced frame example is used to show a clear subharmonic relationship in the HSR peaks.

In Stage 1, channels corresponding to peaks in $\tilde{q}_{s,t}(k)$ that have an amplitude greater than 1 are selected (Fig. 2(IIa)). Taking the peaks in $\tilde{q}_{s,t}(k)$ ensures that only the best-matched comb-filters among their neighbors with similar F_k are selected, while setting an amplitude threshold of 1 selects the channels whose harmonic energy is greater than its subharmonic counterpart.

If $x_s(n)$ has a pitch value of f_0, there should be peaks in $\tilde{q}_{s,t}(k)$ at $F_k \approx f_0$, and its subharmonics, i.e., at $F_k \approx f_0/2$. Hence, in Stage 2, comb-channels with $F_k = f_a$ and $0.5f_a$ are retained (if both are selected in Stage 1, see Fig. 2(IIb)). To implement this selection scheme, HSR is also evaluated using $c_k(f)$ with F_k that are half of those defined in the original search array. Since voiced speech is generally not perfectly harmonic, especially in noise, we relax the subharmonic channel selection criterion to within $\pm 20\%$ of $0.5f_a$.

From Fig. 2(IIb), it can be observed that besides comb-channels with $F_k = f_0$ and $f_0/2$, comb-channels with F_k corresponding to lower subharmonic frequencies $(f_0/i, i > 2)$ might also be selected. The number of these lower subharmonic frequencies within the pitch range of interest increases with f_0. Since longer frame lengths (N_k) are used in comb-channels with lower F_k, stopping channel selection at Stage 2 would result in early/late detections of a voicing onset/offset, especially when f_0 is high. Hence, a third selection stage is implemented to make the number of channels selected less dependent on f_0. An energy-normalized ACR is computed from each comb-filtered, time-domain signal, $y_{k,s,t}(n)$ (inverse-FFT of $Y_{k,s,t}(f)$), of the selected channels in Stage 2. If the pitch value is f_0, the maximum ACR peak should still be located at $\tau = f_s/f_0$ samples, even for the subharmonic comb-channels, as shown in the ACRs plotted in row (III) of Fig. 2. To retain comb-channels with $F_k = f_0$ and $f_0/2$, channels whose maximum ACR peak is located within a 20% deviation tolerance of $\tau = f_s/F_k$ or $\tau = 0.5f_s/F_k$ are selected in Stage 3, as shown in Fig. 2(IIc). Retaining an additional subharmonic channel with $F_k = f_0/2$ helps improve voicing detection at low SNRs.

2.5. SC computation per stream

After channel selection, a correlogram matrix representation, $R_{s,t}(k', \tau)$, is formed per stream by vertically stacking

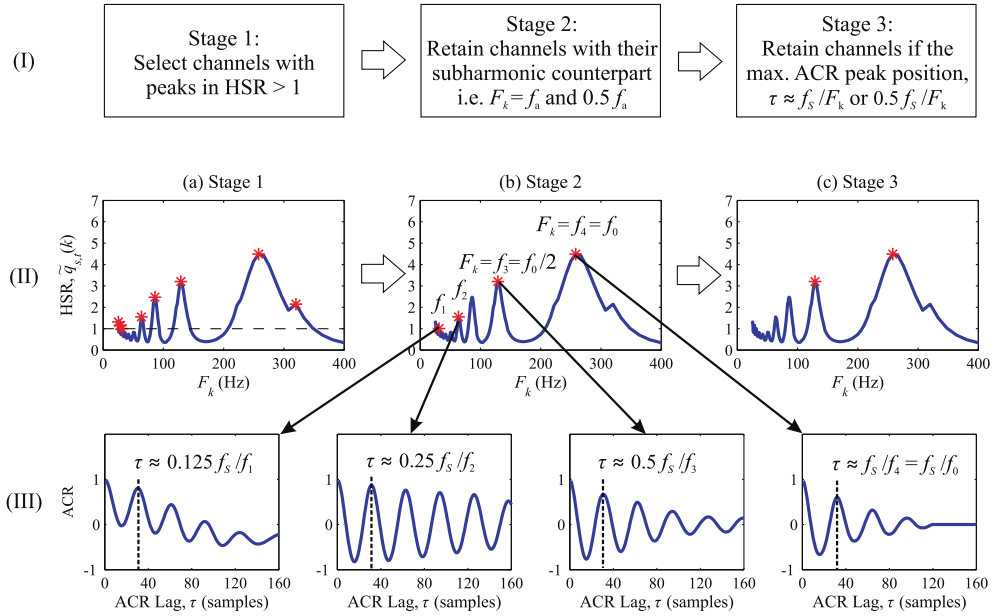

Fig. 2. (I) Flowchart of the tri-stage channel selection. (II) HSR, $\tilde{q}_{s,t}(k)$ (blue line) and selected channels (red asterisks) in (a) Stage 1, (b) Stage 2, and (c) Stage 3, of a particular $x_s(n)$ with $f_0 \approx 260$ Hz, for a clean voiced frame. (III) ACRs computed from comb-filtered outputs of channels selected in Stage 2. Channels with $F_k \approx 130$ Hz and 260 Hz are selected after this tri-stage selection process in this example. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the energy-normalized ACR functions of the selected K_s channels, where K_s is the total number of channels selected in stream s, and $k' = 1, 2, \ldots, K_s$ represents renumbered indices of the selected channels. Instead of a simple average of the selected channels' ACRs to get each stream's summary correlogram (SC), as done in Rouat et al. (1997), an HSR-based weighted averaging scheme is performed by multiplying the stream's HSR-based channel-weighting function, $\phi_{s,t}(k')$, to its corresponding correlogram matrix, $R_{s,t}(k', \tau)$, to obtain the stream SC, $r_{s,t}(\tau)$, as shown in Eqs. (9)–(11). A value of one is subtracted from $\tilde{q}_{s,t}(k')$ in Eq. (11) to increase the relative differences of the values in $\phi_{s,t}(k')$, since every value in $\tilde{q}_{s,t}(k')$ is greater than one. The process of computing the HSR-weighted stream SC from the selected channels' ACRs is also illustrated in Fig. 3, which is performed on a babble noise-corrupted voiced speech frame at 5 dB SNR that has a fundamental period ≈68 samples.

$$r_{s,t}(\tau) = \sum_{k'} \phi_{s,t}(k') R_{s,t}(k', \tau) \tag{9}$$

$$\phi_{s,t}(k') = \frac{Q_{s,t}(k')}{\sum_{k'} Q_{s,t}(k')} \tag{10}$$

$$Q_{s,t}(k') = \tilde{q}_{s,t}(k') - 1 \tag{11}$$

Through this weighting scheme, ACRs from the more reliable channels (with higher HSR) will have a greater impact on their stream SC, $r_{s,t}(\tau)$, resulting in a more prominent ACR peak at the most likely pitch period of the signal. Peak prominence in each stream SC is also due to the use of harmonically-enhanced comb-filtered signals in the individual channel's ACR computation. This effect is shown in Fig. 4 for the same 5 dB SNR babble noise-corrupted voiced speech frame used in Fig. 3. Row (I) plots the energy-normalized ACR function of pre-comb-filtered $x_s(n)$. A frame length of 4 times the true fundamental period (T0), is used to compute these ACRs for a fair comparison. Row (II) plots the SCs obtained by using a simple average of selected channels' ACRs in each stream, while row (III) plots the SC, $r_{s,t}(\tau)$, obtained using the proposed HSR-based weighting scheme. In this example, it is evident that the SCs in row (II) for $s = 1$, 3, and 4 have a more prominent peak compared to the respective stream's ACR of pre-comb-filtered $x_s(n)$ in row (I), and the HSR-weighted SCs in row (III) for $s = 0$, 3, and 4 have a more prominent peak at the true T0 compared to the equal-channel-weighted SCs in row (II).

2.6. MBSC computation

The stream SCs are further fused into a single SC, which we named the multi-band summary correlogram (MBSC), $r_t^{mb}(\tau)$. The contributions of the stream SCs are determined by a stream-reliability-weighting function, $w_t(s)$, as shown in Eq. (12). It is observed that the maximum HSRs in the more reliable streams are higher. In addition, reliable $r_{s,t}(\tau)$ tend to have similar peak locations. Hence, the values of $w_t(s)$ in Eq. (13) are made dependent on two factors: (1) a within-stream reliability factor, $\alpha_t(s)$, based on the maximum value of $Q_{s,t}(k')$ in each stream, as shown in Eq. (14); (2) a between-stream reliability factor, $\beta_t(s)$, that corresponds to the number of $r_{s,t}(\tau)$ whose maximum peak position, $g_t(s)$, falls within 10% of its own, as defined in Eqs. (15)–(17). A stricter deviation tolerance of 10% is set in Eq. (16) compared to the 20% tolerance allowed in channel selection, because there is a high tendency of assigning a large $\beta_t(s)$ to an unreliable $r_{s,t}(\tau)$ when a larger tolerance is set.

$$r_t^{mb}(\tau) = \sum_{s=0}^{4} w_t(s)\, r_{s,t}(\tau) \tag{12}$$

$$w_t(s) = \frac{\alpha_t(s)\, \beta_t(s)}{\sum_{s=0}^{4} \alpha_t(s)\, \beta_t(s)} \tag{13}$$

$$\alpha_t(s) = \max_{k'} Q_{s,t}(k') \tag{14}$$

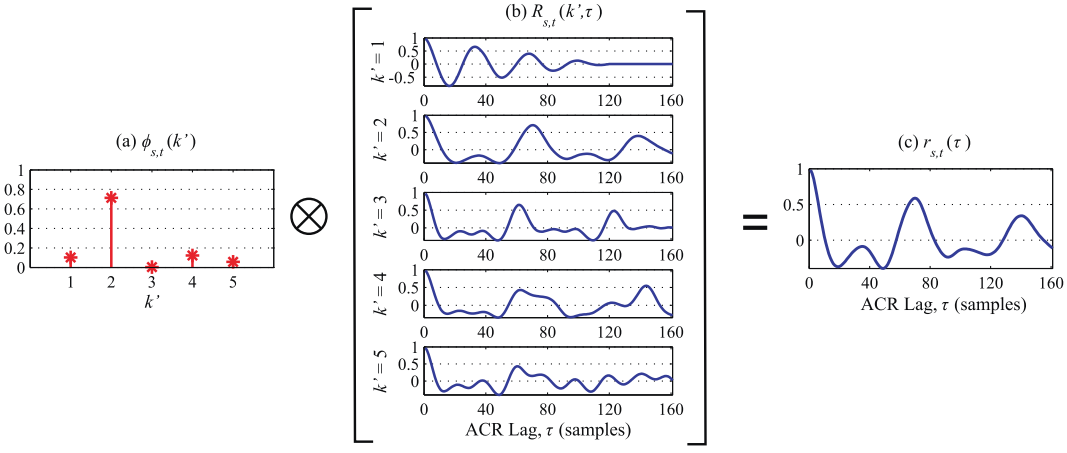

Fig. 3. (a) HSR-based channel-weighting function, $\phi_{s,t}(k')$, (b) Selected channels' ACRs in the correlogram matrix, $R_{s,t}(k', \tau)$, and (c) HSR-weighted stream summary correlogram, $r_{s,t}(\tau)$ in stream $s = 0$, for a babble noise-corrupted voiced frame with a fundamental period ≈68 samples and an SNR of 5 dB.

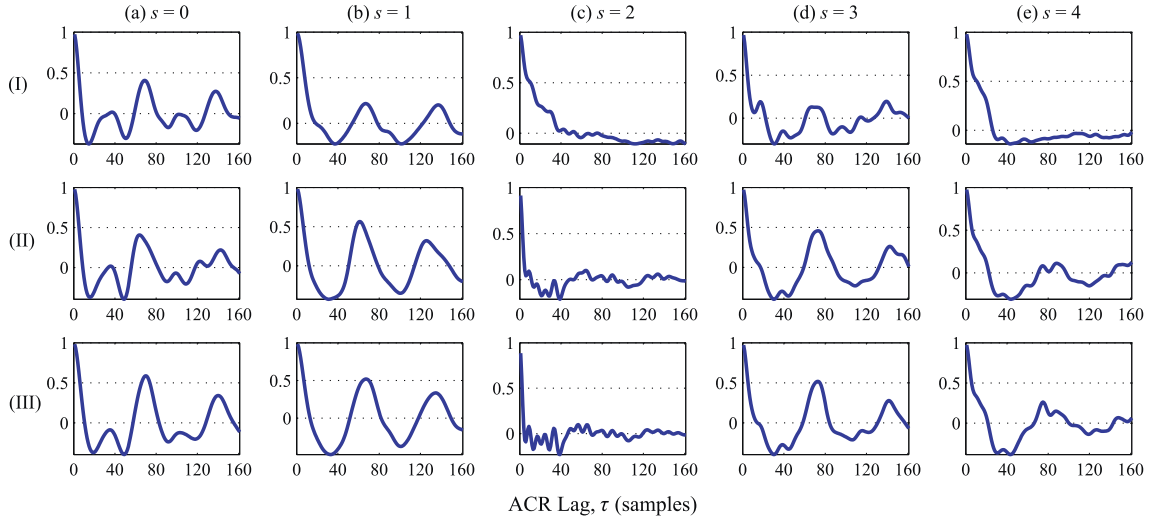

Fig. 4. (I) ACRs computed from each pre-comb-filtered $x_s(n)$, (II) Stream SCs obtained by averaging the selected channels' ACR, and (III) Stream SC obtained using the proposed HSR-weighting scheme, for $s = 0, 1, \ldots, 4$. All ACRs and SCs are computed at the same babble noise-corrupted voiced frame as that used in Fig. 3.

$$\beta_t(s) = \sum_{s'=0}^{4} p_t(s, s') \tag{15}$$

$$p_t(s, s') = \begin{cases} 1, & \text{if } \left|1 - \frac{g_t(s')}{g_t(s)}\right| < 0.1 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

$$g_t(s) = \arg\max_{20 \leqslant \tau \leqslant 160} \text{peak } r_{s,t}(\tau) \tag{17}$$

Fig. 5(a) plots the MBSC (in blue solid line) obtained by applying an equal-stream-weighting scheme on the stream SCs, $r_{s,t}(\tau)$, in row (III) of Fig. 4. On the other hand, Fig. 5(b) plots the MBSC (in blue solid line) obtained by applying the stream-reliability-weighting scheme on the same SCs. When these two multi-band SCs are compared to their counterparts (in red dashed lines) calculated from clean speech, it can be observed that difference in MBSC peak amplitudes between the clean and the 5 dB versions are larger with the equal-stream-weighting, compared to the proposed weighting technique. The differences between the two schemes will be larger if there are fewer reliable stream SCs than those present in this example. Thus, the stream-reliability-weighting scheme reduces the variability of the maximum peak amplitude in the MBSC for noisy speech, such that this amplitude becomes a robust indicator of the frame's degree of voicing.

To improve the inter-frame consistency of the MBSC's maximum peak location across a continuous voiced segment, the same lowpass IIR filter in Eq. (8) is applied on $r_t^{mb}(\tau)$ to obtain a time-smoothed $\tilde{r}_t^{mb}(\tau)$. Time-smoothing also slows down the rate of decrease of peak amplitudes, which in turn improves the detection of weak voiced frames at voicing offsets.

2.7. Pitch estimates and V/UV decisions

From $\tilde{r}_t^{mb}(\tau)$, pitch candidates corresponding to the 10 highest peaks with amplitudes greater than 0 are identified. Each peak and its immediate neighbors are fitted by a parabola, and the amplitude and lag position corresponding to the maximum point of this parabola are the refined pitch measurements for the respective pitch candidate (Cheveigné and Kawahara, 2002). To favor the pitch

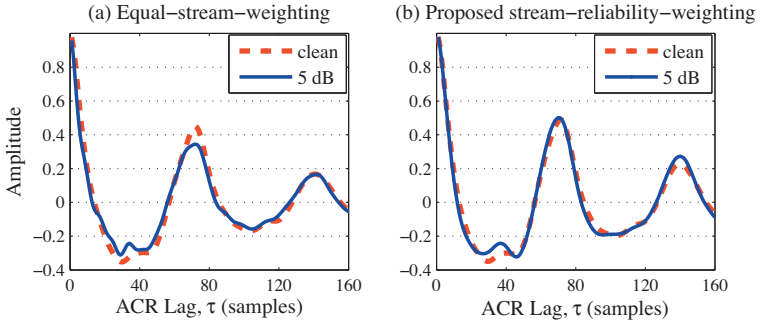

Fig. 5. MBSCs computed using (a) an equal-stream-weighting scheme, and (b) the proposed stream-reliability weighting scheme on the stream SCs, plotted in row (III) of Fig. 4.

candidate with the smallest MBSC lag position when peaks of similar amplitudes are present, $\lambda(\tau) = 1 - \frac{0.7}{\tau_{max}}\tau$, an empirically-determined, linearly decreasing lag-weighting function, is multiplied to the interpolated peak amplitudes (Talkin, 1995), where $\tau_{max} = 160$ is the maximum number of samples in one period (given an 8-kHz signal, and the minimum F0 set at 50 Hz). The lag position of the best pitch candidate (the one with the highest lag-weighted peak amplitude) gives the estimated pitch period.

As for V/UV detection, a constant threshold is applied on the maximum interpolated peak amplitude (prior to lag-weighting) to obtain the initial binary V/UV decision for each frame. This is followed by a 5-point median filtering in time on these initial decisions to get the final V/UV detection outputs.

2.8. Summary of the proposed MBSC PDA

The proposed MBSC PDA involves several algorithmic novelties to improve the robustness of its pitch detection performance. Instead of a gammatone filterbank (Slaney and Lyon, 1990; Rouat et al., 1997; Wu et al., 2003), four wideband FIR filters equally-spaced in the linear frequency are used. The low-frequency wideband FIR filter in the MBSC PDA facilitates the use of signal envelope in this band, which contributes to a higher pitch estimation accuracy for bandpass-filtered speech. The FIR filters have equal separation in the linear frequency range so that there is no bias towards harmonics in particular frequency ranges. This helps avoid a significant performance degradation when several harmonics in a particular frequency band are masked by noise. Besides the Hilbert signal envelope of each subband, the non-envelope signal stream from the low-frequency subband is also used. Multi-channel comb-filtering is performed separately for each subband stream. This is in contrast to other comb-filter-based algorithms that apply comb-functions spanning the full spectrum, making them highly dependent on prominent low-frequency harmonics to perform well, especially if the comb-functions have decreasing amplitude with frequency (e.g. Hermes, 1988; Camacho and Harris, 2008). To derive the peak-enhanced MBSC, an SC is first computed for each stream using an HSR-weighted-average of the ACRs computed from comb-filtered outputs of selected channels. These SC streams are further fused to form the MBSC based on their within-stream and between-stream reliability factors. Together, the proposed signal processing schemes (subband multi-channel comb-filtering, HSR-based channel-selection-and-weighting, stream-reliability-weighting) help to enhance the maximum MBSC peak at the most likely pitch period, which in turn improves the accuracy of pitch estimation, as well as V/UV detection. The variability of the maximum MBSC peak amplitude with SNRs is reduced, such that robust V/UV detection is achieved by simply applying a constant threshold on this single feature, followed by median filtering – without requiring additional features (Sun, 2002), a separate V/UV detection module, or a pitch continuity tracking algorithm involving dynamic programming (Talkin, 1995), or statistically-trained, data-driven modeling techniques (Wu et al., 2003). To ascertain whether or not the threshold is heavily data-dependent, our evaluations involve separate development and evaluation data sets, as described in Section 3.

3. Experimental setup

3.1. Reference pitch corpora for development and evaluation

The two popular pitch corpora for speech – Keele (Plante et al., 1995) and CSTR (Bagshaw et al., 1993) are used to generate narrowband noisy speech, with a sampling rate of 8 kHz (downsampled from the original clean version at 20 kHz). The dataset generated from Keele corpus is the development (dev) set used for algorithmic parameter tuning, while the dataset generated from the CSTR corpus is treated as the evaluation (eval) set. The reference pitch contours (provided in the corpora) are derived from laryngograph signals recorded simultaneously.

3.1.1. Dev set – noise-added Keele

The Keele corpus contains a phonetically balanced story read by five adults from each gender: one file per speaker, each about 30 seconds long. To simulate bandlimited noisy speech, we down-sampled the data to 8 kHz, mixed it with three real-world noise types – babble, car (volvo), and machine gun. These are sample noise files from the NOISEX-92 corpus (Varga and Steeneken, 1993) (available at http://spib.rice.edu/spib/select_noise.html). They are selected for their different degrees of stationarity as presented in Table 1. The noise files are also downsampled to 8 kHz before adding them to clean speech at 20, 15, 10, 5, and 0 dB SNR using the Filtering-and-Noise-adding-Tool (FaNT) (Hirsch, 2005). Speech with fullband or G.712 (ITU, 1996) spectral characteristics is generated by setting the "-m snr_4khz" or "-f g712" option, respectively in FaNT. The ITU G.712 characteristic filter has a flat bandpass response between 300 and 3400 Hz. Since fundamental harmonics below 300 Hz are attenuated, the G.712 dataset is a more challenging corpus than its fullband counterpart.

3.1.2. Eval set – noise-added CSTR

The CSTR corpus contains about 5 min of speech from an adult male and an adult female (50 sentences each). The phonemes in the sentences are biased towards voiced fricatives, nasals, liquids and glides, for which accurate pitch estimation is difficult. Since the pitch contours in the reference files are not computed at a regular rate of 100 Hz, linear interpolation and extrapolation are applied to obtain a reference pitch value at every 10 ms for performance evaluation.

3.2. Algorithms for comparison

The performance of the algorithms listed in Table 2 are evaluated on the same pitch corpora for comparison

Table 1
Characteristics of the noise-types used to generate noisy speech.

Noise type	Spectral characteristics
Babble	Non-stationary speech-shaped harmonics
Car	Highly stationary low-frequency noise
Machine gun	Highly non-stationary, containing: – No-noise time segments – Short bursts of energy covering all frequencies – Each burst is followed by a longer low-frequency noise

Table 2
Algorithms evaluated for performance comparison.

Algorithm	URL to download code
RAPT (Talkin, 1995)	www.speech.kth.se/wavesurfer/download.html
YIN (Cheveigné and Kawahara, 2002)	http://audition.ens.fr/adc/sw/yin.zip
SHR (Sun, 2002)	Search for "pitch determination algorithm" at www.mathworks.com/matlabcentral/fileexchange
SWIPE and SWIPE' (Camacho and Harris, 2008)	www.cise.ufl.edu/~acamacho/publications/swipep.m
WWB (Wu et al., 2003)	www.cse.ohio-state.edu/~dwang/pnl/shareware/wu-tsap03/,
MBSC	http://www.ee.ucla.edu/~spapl/shareware.htm

purposes. These algorithms are selected because they are popular benchmarking algorithms – RAPT (Talkin, 1995) used in Wavesurfer (Sjölander and Beskow, 2000), YIN (Cheveigné and Kawahara, 2002)) – or they employ signal processing techniques that have similarities to our proposed MBSC pitch detector – SHR (Sun, 2002), SWIPE, SWIPE' (Camacho and Harris, 2008), WWB (Wu et al., 2003). The default parameter values proposed in these algorithms are used, unless specified otherwise in this paper. The following parameters are set common in all algorithms – sampling rate = 8 kHz, pitch range from 50 to 400 Hz, and frame rate = 100 Hz. In WWB, the original values for these parameters are maintained, because altering them in its header file"common.h" either results in poor performance or runtime errors. Since WWB is designed for 16-kHz data, the 8-kHz data is upsampled to 16 kHz before running this algorithm. The audio files in the Keele corpus are also broken into 5-s segments with 100 ms overlap before running WWB, because its code can process at most 6 s of speech (even after its "MAX_SAMPLES" parameter is maximized to avoid insufficient memory problem at runtime). The multi-segment outputs are then concatenated by deleting the last four 10ms-frames in the former segment, and the first five frames in the latter segment for each overlapping section. This concatenation procedure results in the correct number of output frames in the final results, and minimizes pitch detection error over other deletion variations for the 4 kHz fullband clean Keele corpus. Audio segmentation is not required to run WWB for the CSTR corpus whose longest utterance lasts only 5 s.

3.3. Experiments and performance measures

We conducted two experiments (Expts): Expt. 1 focuses only on pitch estimation accuracy, and all voiced frames in the reference pitch contours are evaluated, i.e. assuming V/UV detection is perfect. Expt. 2 evaluates pitch detection accuracy by taking into account both pitch estimation and V/UV detection errors. The measure used for performance evaluation in Expt. 1 is the gross pitch error for all reference voiced frames (GPE_{refV}), while the key performance measure in Expt. 2 is the pitch detection error (PDE). These performance measures are explained in the following subsections.

3.3.1. Expt. 1: pitch estimation

Since perfect V/UV detection is assumed in Expt. 1, only pitch estimation errors can occur under this pitch detection scenario. Hence, the gross pitch error computed over all voiced frames noted in the reference files is the performance measure used in this experiment. This popular performance measure, denoted by GPE_{refV} in Eq. (18), has also been used by developers of comparative algorithms in Cheveigné and Kawahara (2002); Sun (2002); Camacho and Harris (2008). N_V and $N_{P,refV}$ are the number of reference voiced frames, and the number of frames with gross pitch errors (estimated and reference F0 differ by more than 20%), respectively.

$$GPE_{refV} = \frac{N_{P,refV}}{N_V} \times 100\% \tag{18}$$

To ensure there is an F0 estimate for almost every reference voiced frame, voicing-related parameters in RAPT, SHR and WWB are altered – RAPT's "VO_BIAS" to 1; SHR's "CHECK_VOICING" to 0; and state transitions probabilities ($\Omega_i \to \Omega_j$) in WWB are set to 0, except for $\Omega_0 \to \Omega_1, \Omega_1 \to \Omega_1$, and $\Omega_2 \to \Omega_1$, which are set to 1. For the proposed MBSC algorithm, the estimate given by the best pitch candidate of each frame is used for computing GPE_{refV}.

3.3.2. Expt. 2: pitch detection

Both pitch estimation and V/UV detection accuracies are of interest in Expt. 2. Three performance measures are computed – gross pitch error (GPE), voicing decision error (VDE) (Rabiner et al., 1976), and pitch detection error (PDE) (Chu and Alwan, 2009). In contrast to Expt. 1, GPE in Eq. (19) is only computed over the reference voiced frames that is detected by the respective algorithm, whose count is represented by N_{VV}, and N_P is the number of frames with gross pitch errors.

VDE in Eq. (20) is the percentage of V/UV detection errors. $N_{V\to UV}$ is the number of voiced frames misclassified as unvoiced, and vice versa for $N_{UV\to V}$, while N is the number of frames in the utterance. A pitch detector can have a low GPE, but a high VDE because many challenging voiced frames are misclassified as unvoiced, and vice versa, which makes pitch detection performance comparison

Table 3
Degree-of-voicing feature and the constant threshold level applied in Expt. 2.

Algorithm	Degree-of-voicing feature	Value
YIN	Aperiodicity measure	0.425
SWIPE	Max. normalized inner product	0.175
SWIPE'	Max. normalized inner product	0.175
MBSC	Max. peak amplitude in $\tilde{r}_t^{mb}(\tau)$	0.375

based on these two separate measures difficult. Thus, *PDE* in Eq. (21), which represents the percentage of pitch detection errors is the key performance measure used for comparative evaluation in this paper. The *PDE* takes into account all possible mutually exclusive pitch detection errors that can occur in any given frame, i.e., voiced-to-unvoiced, unvoiced-to-voiced, or gross pitch error. This performance measure is formally defined in Chu and Alwan (2009), and it has also been used for pitch detection performance evaluation in Oh and Un (1984) and Rouat et al. (1997).

$$GPE = \frac{N_P}{N_{VV}} \times 100\% \tag{19}$$

$$VDE = \frac{N_{V\to UV} + N_{UV\to V}}{N} \times 100\% \tag{20}$$

$$PDE = \frac{N_{V\to UV} + N_{UV\to V} + N_P}{N} \times 100\% \tag{21}$$

For RAPT, SHR, and WWB, pitch detection performance evaluation is conducted using the algorithms' default settings. As for pitch estimation algorithms — YIN, SWIPE, and SWIPE', V/UV detection is performed using the same constant thresholding and median filtering scheme as that used in our proposed algorithm. The threshold level (varied in steps of 0.025) that gives the lowest averaged *VDE* for noisy speech in the Keele (dev) corpus is used. Table 3 shows the threshold levels tuned for these algorithms and the proposed MBSC PDA.

4. Expt. 1: pitch estimation performance evaluation

4.1. Results

The GPE_{refV} for the CSTR corpus (eval set) are shown in this section. To reduce the amount of data presented, the GPE_{refV} results for the noisy CSTR corpus are averaged on a per noise-type or a per SNR basis. Table 4 contains the results for clean CSTR speech, while Table 5 and Table 6 show the mean GPE_{refV} obtained by averaging on a per noise-type and a per SNR basis, respectively. The lowest value in each column is boldfaced to indicate the best-performing algorithm in each case. Note that the GPE_{refV} obtained for the eval and dev sets have similar trends. For the Keele (dev) corpus results, please refer to Tables (A.10–A.12) in Appendix.

SWIPE' has the lowest GPE_{refV} for fullband clean speech. For G.712 clean speech, WWB has the lowest GPE_{refV}. In these cases, the absolute difference between MBSC and the best performing algorithm is small, less than 0.4%.

Table 4
Average GPE_{refV} (%) for clean CSTR (eval) corpus, assuming perfect V/UV detection. The result of the best performing algorithm in each case is boldfaced.

Algorithm	Clean, fullband	Clean, G.712
RAPT	9.63	10.4
YIN	4.02	7.76
SHR	3.78	11.0
SWIPE	3.69	10.2
SWIPE'	**3.49**	13.7
WWB	4.11	**4.19**
MBSC	3.83	4.50

Table 5
GPE_{refV} (%) averaged across SNRs from 20 to 0 dB per noise-type for noise-corrupted, fullband/G.712 CSTR (eval) corpus, assuming perfect V/UV detection.

Algorithm	Babble	Car	Mach. Gun	Avg. all
RAPT	34.2/31.1	11.3/12.7	25.5/19.2	23.7/21.0
YIN	18.9/22.6	9.68/16.4	12.0/15.2	13.5/18.1
SHR	16.7/24.9	6.35/17.5	11.0/17.5	11.4/19.9
SWIPE	19.3/24.4	7.41/27.0	11.3/18.6	12.7/23.3
SWIPE'	17.6/28.6	7.23/33.2	10.7/23.8	11.9/28.5
WWB	20.9/18.8	6.46/9.71	7.39/8.54	11.6/12.3
MBSC	**12.5/13.7**	**5.06/8.47**	**5.95/6.76**	**7.85/9.66**

Table 6
GPE_{refV} (%) averaged across noise-types per SNR for noise-corrupted, fullband/G.712 CSTR (eval) corpus, assuming perfect V/UV detection.

Algorithm	20 dB	10 dB	0 dB	Avg. all
RAPT	13.3/11.2	21.9/18.7	37.4/35.6	23.7/21.0
YIN	4.63/8.88	9.39/15.4	30.4/32.8	13.5/18.1
SHR	4.64/12.7	8.98/18.0	22.8/31.0	11.4/19.9
SWIPE	5.09/14.5	10.1/21.8	25.2/35.2	12.7/23.3
SWIPE'	4.73/19.7	8.99/27.7	24.5/39.1	11.9/28.5
WWB	5.63/6.18	9.71/9.70	21.1/23.7	11.6/12.3
MBSC	**4.16/5.34**	**6.00/8.24**	**15.1/17.1**	**7.85/9.66**

For noisy speech, MBSC achieves the lowest mean GPE_{refV}, whether it is averaged over SNRs for each noise-type or averaged over noise-types at each SNR. A minimum overall GPE_{refV} of 7.85% and 9.66% are achieved for fullband and G.712-filtered noisy speech in the CSTR (eval) corpora, respectively.

4.2. Discussion

In general, GPE_{refV} is higher for the G.712-filtered data compared to its fullband version because of an increase in over-estimation (doubling/tripling) error, especially for low-pitched utterances whose fundamental harmonic is severely attenuated after G.712-filtering. The reverse trend is observed in some cases for the RAPT and WWB algorithms when the reductions in under-estimation errors for G.712-filtered, high-pitched utterances exceed increases in over-estimation errors for the low-pitched utterances.

The low GPE_{refV} achieved by SWIPE' and SWIPE for clean fullband speech shows that their comb-filter-based pitch estimation algorithm is effective in reducing harmonic and subharmonic errors, which are usually the main error contributors for clean speech. However, for the G.712 version, SWIPE' and SWIPE are not among the top three best-performing algorithms, and the performance of SWIPE' is worse than SWIPE. This is because SWIPE and SWIPE' use a decreasing amplitude comb-filter, so its performance degrades significantly when the lower harmonics in the signal are missing or heavily attenuated in the G.712-filtered data. The degradation is more severe for SWIPE' that heavily relies on the first and prime harmonics (many non-prime harmonics of higher frequencies are ignored) in its comb-filter.

Although MBSC is not the best performing algorithm for clean speech in the eval set, it still has the lowest average GPE_{refV} at a high SNR of 20 dB, as shown in Table 6. MBSC's low GPE_{refV} under different noise-types and SNR levels shows that the proposed algorithm gives robust pitch estimation accuracy under a wide range of noise conditions. From Table 5, it is also observed that accurate pitch estimation is challenging for babble noise-corrupted speech for all the algorithms investigated. This is because babble noise also has a harmonic structure – it is difficult to estimate a correct pitch from the speech harmonics when an interfering set of harmonics is present.

We had also assessed the pitch estimation performance of the algorithms on frequency-shifted speech (results are not shown here), i.e., harmonics are present at $nf_0 + \delta$, where n is a positive integer, f_0 is the pitch value, and δ is the amount of frequency offset. This frequency-shift phenomenon can occur due to a carrier frequency offset between the transmitter and receiver. It is commonly found in received speech of communication systems that use the single sideband suppressed carrier (SSB-SC) modulation scheme, which are popular for voice transmission in the high frequency (HF) radio spectrum by amateur radio, commercial, and military operators (Frerking, 1994). The usage of signal envelopes in all subbands of the MBSC-based pitch estimation algorithm leads to a significantly ($> 30\%$) lower GPE_{refV} in comparison to other algorithms, since the inter-harmonic frequency separation is invariant for frequency-shifted speech. Thus, the subband signals' amplitude modulation frequency will still be equal to f_0. Other algorithms generally output a pitch estimate that is either the first positive frequency of the shifted harmonics, or the frequency of the strongest shifted harmonic. For WWB, its mid- and high-frequency subbands' ACRs are also computed from the envelopes of subband signals containing multiple harmonics. Thus, these ACRs contain information for accurate pitch estimation. However, ACRs computed from the individual shifted, low-frequency harmonics, (captured by the narrowband gammatone filters at the low frequencies) provide pitch information that is inconsistent with the envelope ACRs. As a result, the pitch estimation accuracy of WWB is only slightly better than the other time-, or frequency-domain algorithms evaluated in this case.

5. Expt. 2: pitch detection performance evaluation

5.1. Results

Tables 7–9 present the *GPE*, *VDE*, and *PDE* obtained for the CSTR (eval) corpus, arranged according to the data characteristics and noise-types – Table 7 for clean speech, and Tables 8 and 9 for fullband and G.712-filtered noisy speech, respectively. The lowest value in each column is boldfaced. Note that similar trends in pitch detection performance are observed for the eval and dev sets. For pitch detection results of the Keele (dev) corpus, please refer to Tables A.13–A.15 in Appendix.

For clean fullband and G.712-filtered speech in the eval set, it can be observed in Table 7 that RAPT has the lowest *VDE* and *PDE*.

For noisy speech in the eval set, although MBSC does not always have the lowest *GPE* and *VDE* among the algorithms at each individual noise-condition, it still has the lowest *PDE* because both its *GPE* and *VDE* are low. MBSC also has the lowest *GPE*, *VDE* and *PDE* when averaged over all noise-types investigated at each SNR (tabulated in the bottom row of Tables 8 and 9), thus it also has the best overall pitch detection performance (tabulated in the bottom-right column of Tables 8 and 9), for both fullband and G.712 noisy speech.

5.2. Discussion

For clean speech, MBSC has a slightly higher *PDE* than RAPT because the longer frame length defined in MBSC can cause an early onset and late offset detection of voiced segments. RAPT's good pitch detection performance for clean speech is mainly attributed to its low *VDE*. The usage of a short 7.5 ms frame in RAPT produces sharper transitions at voicing boundaries. In addition, RAPT derives its voicing transition costs in its dynamic programming algorithm based on inter-frame energy ratio and spectral difference, which provide good indications of voicing transition boundaries for clean and some cases of high SNR noise conditions.

Table 7
Average *GPE*, *VDE*, and *PDE* (%) for clean CSTR (eval) corpus.

Algorithm	Clean, fullband			Clean, G.712		
	GPE	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*
RAPT	2.55	**4.99**	**5.93**	2.86	**5.70**	**6.75**
YIN	1.98	7.38	8.10	4.86	7.97	9.79
SHR	2.27	7.95	8.81	9.43	9.32	13.1
SWIPE	2.00	6.18	6.93	7.46	9.05	11.8
SWIPE'	1.95	7.01	7.75	10.7	9.85	13.9
WWB	1.91	8.13	8.82	1.93	8.25	8.95
MBSC	**1.85**	5.82	6.52	**1.90**	6.67	7.37

Table 8
Average *GPE*, *VDE*, and *PDE* (%) for noise-corrupted, fullband CSTR (eval) corpus.

SNR (dB)	20			10			0			Avg. 20 to 0		
Algo.	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*
Babble noise-corrupted CSTR, fullband												
RAPT	2.62	8.12	9.05	7.95	**12.3**	14.9	35.9	**25.7**	33.8	13.7	**14.6**	18.2
YIN	1.84	9.60	10.3	5.90	14.1	16.0	28.7	28.9	34.6	10.8	16.7	19.4
SHR	2.82	**7.76**	8.78	6.09	12.8	14.7	16.0	30.1	33.0	7.70	16.0	8.78
SWIPE	2.75	12.6	13.6	8.18	18.0	20.6	30.2	33.3	39.3	12.3	20.6	23.6
SWIPE'	2.41	13.5	14.4	5.90	18.8	20.7	26.4	35.6	40.3	10.2	21.7	24.1
WWB	4.78	14.3	15.9	9.18	17.9	20.7	21.2	28.7	33.7	11.0	19.7	22.8
MBSC	**1.71**	7.98	**8.61**	**2.48**	12.4	**13.1**	**9.30**	26.8	**28.5**	**3.98**	14.9	**15.8**
Car noise-corrupted CSTR, fullband												
RAPT	2.17	4.68	5.46	1.59	6.36	6.90	1.73	15.5	16.0	3.35	14.3	15.0
YIN	1.57	4.97	5.53	1.18	8.04	8.43	3.40	22.2	22.9	1.75	10.9	11.4
SHR	2.19	7.92	8.76	2.25	12.3	13.0	3.61	26.2	27.2	2.51	14.8	15.6
SWIPE	1.33	4.79	5.27	0.73	6.86	7.09	0.91	16.7	16.9	0.92	8.83	9.13
SWIPE'	**1.24**	4.82	5.26	**0.65**	7.31	7.52	**0.58**	18.3	18.4	**0.75**	9.53	9.77
WWB	1.95	7.38	8.07	1.82	7.26	7.88	2.86	**8.36**	9.27	2.10	7.53	8.24
MBSC	1.70	**4.57**	**5.21**	1.54	**4.98**	**5.54**	1.35	8.55	**8.99**	1.54	**5.76**	**6.31**
Machine gun noise-corrupted CSTR, fullband												
RAPT	3.80	8.86	10.3	10.2	11.0	14.7	19.1	16.3	22.4	10.9	11.7	15.5
YIN	2.04	7.48	8.21	3.76	10.6	11.9	7.03	18.5	20.4	4.04	11.9	13.1
SHR	3.29	12.1	13.3	7.50	21.8	24.5	18.8	31.6	36.6	17.0	26.0	30.1
SWIPE	2.43	6.04	6.94	3.81	9.28	10.6	3.95	16.9	17.9	3.43	10.4	11.5
SWIPE'	2.12	6.69	7.47	2.61	10.7	11.6	3.00	18.7	19.5	2.57	11.7	12.6
WWB	1.97	8.34	9.02	2.07	9.75	10.5	3.47	12.5	13.6	2.35	10.1	10.9
MBSC	**1.74**	**5.19**	**5.84**	**1.69**	**6.39**	**7.00**	**1.75**	**10.4**	**10.9**	**1.72**	**7.12**	**7.72**
CSTR, fullband – Avg. across noise-types												
RAPT	2.86	7.22	8.25	6.58	9.88	12.2	18.9	19.2	24.1	8.79	11.5	14.1
YIN	1.82	7.35	8.00	3.62	10.9	12.1	13.1	23.2	26.0	5.53	13.2	14.6
SHR	2.77	9.26	10.3	5.28	15.6	17.4	12.8	29.3	32.3	9.06	18.9	21.2
SWIPE	2.17	7.82	8.61	4.24	11.4	12.7	11.7	22.3	24.7	5.55	13.3	14.8
SWIPE'	1.92	8.34	9.03	3.05	12.3	13.2	9.98	24.2	26.1	4.49	14.3	15.5
WWB	2.90	10.0	11.0	4.36	11.7	13.0	9.17	16.5	18.9	5.16	12.5	14.0
MBSC	**1.71**	**5.91**	**6.55**	**1.90**	**7.91**	**8.55**	**4.13**	**15.2**	**16.1**	**2.41**	**9.24**	**9.96**

For noisy speech, MBSC's low overall *PDE* is attributed to its good pitch detection performance over various SNRs and noise-types. MBSC has the lowest average *PDE* at both high (20 dB) and low (0 dB) SNRs, which is observable in the bottom block of Table 8 and 9. Similarly, MBSC also has the lowest average *PDE* for each noise-type, as seen in the last column of these two tables. These results affirm that the pitch detection performance of the proposed MBSC is robust against a variety of noise conditions. This also shows that the proposed signal processing schemes are generally effective in enhancing MBSC's peak amplitude at the true pitch period, since low *VDE* is achievable with the simple constant V/UV threshold detection scheme.

For babble noise-corrupted speech, MBSC has a slightly higher *VDE* than RAPT because some of the harmonic-containing babble noise frames are erroneously enhanced and misclassified as voiced speech. However, MBSC still has the lowest *PDE* because of its higher pitch estimation accuracy for babble noise-corrupted speech compared to other algorithms. In the case of car noise-corrupted speech, some comparative algorithms have lower *GPE*s than MBSC because there is a higher proportion of frames with a pitch estimation error for the additional weakly voiced frames detected by MBSC, but missed by the other algorithms (thus the *VDE*s of these algorithms are higher than MBSC's). Machine gun noise is highly non-stationary, such that there is a large variation in SNR within each machine gun noise-corrupted speech utterance. Since the MBSC peak enhancement schemes improve the algorithm's noise-robustness under a wide range of SNRs, MBSC gives superior performance in *GPE*, *VDE* and *PDE* compared to other algorithms for machine gun noise-corrupted speech at all SNR levels.

Besides V/UV detection in a pitch detection application, the maximum MBSC peak amplitude can also be a reliable frame-level degree-of-voicing feature to enhance the noise-robustness of other speech applications. For example, it can be used for speech activity detection (SAD). When the maximum MBSC peak amplitude feature with its deltas, and delta-deltas in time are appended to MFCCs in the SRI speech activity detector developed for the Robust Automatic Transcription of Speech (RATS) project (Walker and Strassel, 2012), the combined feature vector gave a significant performance gain over MFCCs in an SAD task.

Table 9
Average *GPE*, *VDE*, and *PDE* (%) for noise-corrupted, G.712 CSTR (eval) corpus.

SNR (dB)	20			10			0			Avg. 20 to 0		
Algo.	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*
Babble noise-corrupted CSTR, G.712												
RAPT	2.59	10.6	11.5	6.73	15.2	17.4	28.4	28.3	35.0	11.4	17.4	20.5
YIN	4.58	11.9	13.6	9.36	17.1	20.2	26.9	30.3	36.0	12.5	19.0	22.5
SHR	11.0	11.2	15.5	13.0	15.5	19.8	20.2	30.5	34.1	14.2	18.3	22.4
SWIPE	9.33	13.7	17.1	13.9	18.7	23.2	24.9	31.2	36.3	15.3	20.7	25.0
SWIPE'	13.2	15.5	20.3	19.5	20.7	26.8	31.5	32.5	38.6	20.8	22.4	28.1
WWB	3.62	14.0	15.1	7.88	18.7	21.1	20.5	30.5	35.1	10.1	20.3	22.9
MBSC	**2.06**	**7.75**	**8.49**	**2.55**	**13.2**	**14.0**	**4.75**	**26.4**	**27.3**	**2.96**	**15.1**	**15.9**
Car noise-corrupted CSTR, G.712												
RAPT	2.76	**5.03**	6.03	2.18	7.45	8.13	**1.32**	18.3	18.5	**2.03**	9.58	10.2
YIN	4.68	5.64	7.33	5.24	9.60	11.3	3.24	22.7	23.2	4.34	11.9	13.2
SHR	9.93	11.0	14.9	9.84	15.2	18.7	9.78	28.5	30.8	9.86	17.6	20.9
SWIPE	11.4	7.22	11.0	12.9	12.7	15.8	9.43	24.2	25.0	11.5	14.2	16.9
SWIPE'	18.7	8.00	13.9	21.0	14.0	18.3	16.9	25.3	26.4	19.3	15.3	19.2
WWB	**1.61**	6.92	7.48	**2.06**	8.54	9.21	3.45	15.1	16.0	2.35	9.80	10.5
MBSC	1.97	5.06	**5.77**	3.19	**6.67**	**7.78**	2.10	**13.4**	**13.9**	2.55	**7.92**	**8.75**
Machine gun noise-corrupted CSTR, G.712												
RAPT	3.78	18.6	20.0	9.02	22.1	25.5	20.2	27.2	34.7	10.4	22.5	26.4
YIN	5.17	10.6	12.5	7.21	13.7	16.2	9.79	18.8	21.7	7.37	14.3	16.7
SHR	10.3	18.3	22.4	12.7	24.1	28.8	22.5	31.3	37.4	14.5	24.4	29.4
SWIPE	8.42	9.80	12.8	9.96	12.4	15.7	11.2	17.1	20.1	9.81	13.0	16.1
SWIPE'	13.0	11.3	15.9	14.9	14.5	19.1	16.0	19.4	23.4	14.7	14.9	19.4
WWB	1.93	8.78	9.47	2.22	10.6	11.3	3.51	13.6	14.7	2.46	10.9	11.7
MBSC	**1.92**	**6.20**	**6.89**	**1.71**	**8.32**	**8.92**	**1.60**	**11.5**	**12.0**	**1.73**	**8.58**	**9.18**
CSTR, G.712 – Avg. across noise-types												
RAPT	3.04	11.4	12.5	5.98	14.9	17.0	16.6	24.6	29.4	7.95	16.5	19.1
YIN	4.81	9.37	11.1	7.27	13.4	15.9	3.24	22.7	23.2	8.07	15.1	17.5
SHR	10.4	13.5	17.6	11.8	18.3	22.4	17.5	30.1	34.1	12.9	20.1	24.2
SWIPE	9.71	10.2	13.6	12.2	14.6	18.2	15.2	24.2	27.2	12.2	16.0	19.3
SWIPE'	15.0	11.6	16.7	18.5	16.4	21.4	21.5	25.8	29.5	18.3	17.5	22.2
WWB	2.38	9.89	10.7	4.05	12.6	13.9	9.15	19.7	21.9	4.96	13.7	15.1
MBSC	**1.99**	**6.34**	**7.05**	**2.48**	**9.39**	**10.2**	**2.81**	**17.1**	**17.7**	**2.41**	**10.5**	**11.3**

Table A.10
Average GPE_{refV} (%) for clean Keele (dev) corpus, assuming perfect V/UV detection.

Algorithm	Clean, fullband	Clean, G.712
RAPT	5.08	5.32
YIN	3.04	6.83
SHR	2.18	8.18
SWIPE	2.09	8.46
SWIPE'	2.07	11.46
WWB	6.56	6.49
MBSC	**2.01**	**3.33**

Table A.11
GPE_{refV} (%) averaged across SNRs from 20 to 0 dB per noise-type for noise-corrupted, fullband/G.712 Keele (dev) corpus, assuming perfect V/UV detection.

Algorithm	Babble	Car	Mach. Gun	Avg. all
RAPT	22.6/25.4	9.56/11.8	20.8/17.3	17.7/18.2
YIN	15.0/23.0	10.3/18.9	11.5/16.5	12.2/19.4
SHR	12.8/22.7	5.75/17.5	9.28/16.0	9.28/18.7
SWIPE	11.7/21.4	5.97/26.5	10.1/17.3	9.28/21.7
SWIPE'	11.1/25.9	6.15/32.9	9.52/21.8	8.92/26.9
WWB	15.6/14.1	7.95/10.4	9.38/10.0	11.0/11.5
MBSC	**8.29/11.9**	**3.54/7.44**	**3.85/5.83**	**5.23/8.39**

6. Summary and conclusion

A multi-band summary correlogram (MBSC)-based pitch detection algorithm is proposed. The proposed MBSC pitch detector uses four wideband FIR filters to capture multiple harmonics in every subband. This facilitates the use of signal envelopes for pitch estimation in all subbands, including the low-frequency band. Other algorithmic novelties include the comb-filter channel selection and weighting schemes to obtain a peak-enhanced summary correlogram per stream, and the stream-reliability-weighting scheme to give more weight to reliable subband summary correlogram streams when they are combined to form the MBSC. These methods enhance the maximum peak at the most likely pitch period in the MBSC, such that robust voiced/unvoiced (V/UV) detection performance is achieved with a simple constant

Table A.12
GPE_{refV} (%) averaged across noise-types per SNR for noise-corrupted, fullband/G.712 Keele (dev) corpus, assuming perfect V/UV detection.

Algorithm	20 dB	10 dB	0 dB	Avg. all
RAPT	7.80/7.76	15.3/15.6	32.0/33.6	17.7/18.2
YIN	3.70/8.44	8.46/16.1	28.5/36.9	12.2/19.4
SHR	2.97/10.2	6.83/16.5	20.5/31.8	9.28/18.7
SWIPE	2.96/11.9	6.97/19.6	20.4/35.6	9.28/21.7
SWIPE'	2.74/16.3	6.52/25.7	20.0/39.9	8.92/26.9
WWB	7.21/7.44	9.09/9.61	18.5/19.2	11.0/11.5
MBSC	**2.22/4.00**	**3.64/6.59**	**11.4/16.5**	**5.23/8.39**

Table A.13
Average *GPE*, *VDE*, and *PDE* (%) for clean Keele (dev) corpus.

Algorithm	Clean, fullband			Clean, G.712		
	GPE	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*
RAPT	2.24	**3.95**	**5.07**	2.74	**4.94**	**6.28**
YIN	1.67	4.54	5.34	3.98	6.15	7.94
SHR	1.97	13.2	14.2	7.59	13.4	17.1
SWIPE	1.05	5.19	5.70	4.61	8.41	10.4
SWIPE'	**1.02**	5.38	5.89	7.13	9.55	12.5
WWB	2.91	8.92	10.2	3.03	9.06	10.4
MBSC	1.19	4.97	5.55	**1.59**	5.56	**6.28**

threshold detection scheme on the maximum peak amplitude of the MBSC.

When the pitch estimation accuracy is evaluated on every voiced frame (assuming perfect V/UV detection) in Expt. 1, noise-robustness of the proposed algorithm's pitch estimation accuracy is shown through the low average gross pitch error (GPE_{refV}) achieved for noisy speech at various SNRs and noise-types.

When pitch detection performance is assessed based on both pitch estimation and voicing detection accuracies in Expt. 2, the MBSC-based pitch detector has the lowest average pitch detection error (*PDE*) under various SNRs and noise types. This shows that the novel peak-enhancement schemes are generally effective in boosting the MBSC's peak at the true pitch period, since the simple constant threshold V/UV detection scheme still yields good pitch detection performances for many of the noise conditions.

The proposed MBSC-based pitch detector does not differentiate between a harmonic structure arising from speech or noise. One way to improve the algorithm's pitch detection performance in the presence of harmonic noise is

Table A.14
Average *GPE*, *VDE*, and *PDE* (%) for noise-corrupted, fullband Keele (dev) corpus.

SNR (dB)	20			10			0			Avg. 20 to 0		
Algo.	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*
Babble noise-corrupted Keele, fullband												
RAPT	1.72	**5.30**	**6.13**	5.15	**10.5**	12.8	22.2	**29.1**	36.0	8.68	**13.8**	17.0
YIN	1.62	6.71	7.46	4.90	12.8	15.0	21.0	32.5	38.1	8.00	16.3	19.0
SHR	2.55	14.6	15.9	5.88	16.2	19.0	19.8	30.3	37.1	8.52	19.3	22.8
SWIPE	1.08	7.64	8.16	3.30	13.7	15.1	16.4	32.8	37.3	5.96	16.9	19.0
SWIPE'	**0.99**	7.98	8.46	2.62	14.4	15.5	15.0	34.1	37.8	5.26	17.8	19.4
WWB	2.82	12.3	13.5	4.59	17.4	19.3	15.7	31.2	36.0	6.83	19.6	22.1
MBSC	1.16	6.60	7.15	**1.42**	11.2	**11.8**	**5.29**	30.3	**31.7**	**2.27**	14.9	**15.7**
Car noise-corrupted Keele, fullband												
RAPT	1.75	**4.25**	**5.11**	1.28	7.35	7.92	2.21	21.0	21.6	1.65	10.0	10.7
YIN	1.61	5.00	5.74	1.70	10.7	11.4	4.45	30.3	31.3	2.31	14.2	15.0
SHR	2.06	11.4	12.4	2.07	18.9	19.8	3.17	39.1	40.1	2.37	22.3	23.3
SWIPE	0.79	5.49	5.87	0.49	9.05	9.26	0.52	21.5	21.7	0.57	11.3	11.6
SWIPE'	**0.78**	5.65	6.03	**0.36**	10.0	10.2	**0.25**	23.9	23.9	**0.47**	12.4	12.6
WWB	2.68	8.48	9.67	2.44	8.49	9.54	2.15	10.4	11.3	2.44	8.98	10.0
MBSC	1.15	4.70	5.26	1.18	**5.70**	**6.24**	1.26	**10.2**	**10.7**	1.21	**6.58**	**7.12**
Machine gun noise-corrupted Keele, fullband												
RAPT	3.15	**5.07**	6.64	9.67	8.38	13.1	18.2	16.0	23.9	10.3	9.43	14.2
YIN	1.91	5.53	6.41	2.79	10.2	11.4	6.69	19.7	22.1	3.68	11.4	12.9
SHR	3.05	17.5	19.0	6.78	24.1	27.4	17.7	33.4	40.4	8.56	24.7	28.5
SWIPE	1.46	6.05	6.76	2.87	10.1	11.4	3.58	18.8	20.1	2.64	11.3	12.4
SWIPE'	**1.11**	6.34	6.88	1.64	11.1	11.8	1.88	20.7	21.3	1.52	12.3	13.0
WWB	2.76	9.20	10.4	2.82	11.0	12.2	3.46	14.9	16.2	2.97	11.6	12.8
MBSC	1.18	5.16	**5.71**	**1.20**	**6.88**	**7.42**	**1.25**	**11.2**	**11.7**	**1.20**	**7.51**	**8.04**
Keele, fullband – avg. across all four noise-types												
RAPT	2.20	**4.87**	**5.96**	5.37	8.74	11.3	14.2	22.0	27.1	6.87	11.1	14.0
YIN	1.71	5.75	6.54	3.13	11.2	12.6	10.7	27.5	30.5	4.66	14.0	15.6
SHR	2.55	14.5	15.8	4.91	19.7	22.1	13.6	34.3	39.2	6.48	22.1	24.9
SWIPE	1.11	6.39	6.93	2.22	10.9	11.9	6.83	24.4	26.4	3.05	13.2	14.3
SWIPE'	**0.96**	6.66	7.12	1.54	11.8	12.5	5.72	26.2	27.7	2.42	14.2	15.0
WWB	2.75	9.98	11.2	3.29	12.3	13.7	7.10	18.9	21.2	4.08	13.4	15.0
MBSC	1.17	5.49	6.04	**1.27**	**7.92**	**8.48**	**2.60**	**17.2**	**18.0**	**1.56**	**9.67**	**10.3**

Table A.15
Average *GPE*, *VDE*, and *PDE* % obtained for noise-corrupted, G.712-filtered Keele (dev) corpus.

SNR (dB)	20			10			0			Avg. 20 to 0		
Algo.	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*	*GPE*	*VDE*	*PDE*
Babble noise-corrupted Keele, G.712												
RAPT	2.55	**7.69**	8.89	6.81	**13.7**	16.7	24.8	**30.7**	38.5	10.3	**16.7**	20.5
YIN	4.04	10.5	12.2	9.41	17.2	21.0	27.4	34.2	41.2	12.5	19.9	24.0
SHR	9.25	17.2	21.6	14.8	19.8	26.4	29.8	32.4	41.7	17.1	22.3	29.0
SWIPE	5.04	12.2	14.3	8.33	19.4	22.3	24.0	35.1	39.8	11.35	21.6	24.7
SWIPE'	8.40	14.3	17.6	14.2	22.3	26.5	31.4	36.6	42.4	17.1	23.9	28.3
WWB	2.66	12.3	13.5	3.70	17.2	18.7	13.8	31.5	35.6	5.95	19.7	21.8
MBSC	**1.49**	7.70	**8.34**	**2.00**	13.8	**14.5**	**5.23**	31.9	**33.1**	**2.56**	16.9	**17.7**
Car noise-corrupted Keele, G.712												
RAPT	2.05	**5.46**	**6.42**	1.69	9.89	10.6	**1.54**	25.5	25.8	**1.64**	12.7	13.3
YIN	3.71	7.02	8.60	4.20	13.7	15.1	5.42	31.4	32.3	4.36	16.5	17.8
SHR	7.96	13.7	17.3	9.02	22.7	26.0	10.9	41.4	43.8	9.10	25.3	28.4
SWIPE	5.99	10.4	12.5	7.35	18.7	20.4	10.6	33.8	34.4	7.76	20.4	21.8
SWIPE'	10.5	12.7	15.8	15.8	21.0	23.4	18.2	35.0	35.7	15.1	22.4	24.6
WWB	2.32	8.50	9.47	2.05	10.8	11.6	1.92	18.0	18.7	2.09	12.0	12.8
MBSC	**1.60**	5.75	6.45	**1.63**	**8.51**	**9.16**	1.92	**16.8**	**17.4**	1.74	**9.84**	**10.5**
Machine gun noise-corrupted Keele, G.712												
RAPT	3.87	10.8	12.7	10.3	15.7	20.9	22.2	22.9	33.7	11.6	16.3	22.1
YIN	4.70	8.80	10.9	6.75	13.3	16.1	11.6	20.9	25.0	7.55	14.0	17.0
SHR	8.91	19.0	23.2	12.7	24.9	30.7	24.6	34.2	42.8	14.6	25.8	31.9
SWIPE	4.95	10.0	12.0	6.47	13.8	16.2	7.87	20.9	23.2	6.46	14.7	16.9
SWIPE'	8.29	11.9	15.0	10.8	16.3	19.8	12.6	23.2	26.6	10.6	16.9	20.3
WWB	2.92	9.68	10.9	2.76	12.2	13.3	3.80	16.4	17.7	3.05	12.7	13.9
MBSC	**1.44**	**6.16**	**6.79**	**1.38**	**8.39**	**8.97**	**1.45**	**12.8**	**13.3**	**1.41**	**8.92**	**9.51**
Keele, G.712 – avg. across noise-types												
RAPT	2.82	7.97	9.33	6.28	13.1	16.1	16.2	26.4	32.7	7.84	15.2	18.6
YIN	4.15	8.76	10.6	6.79	14.7	17.4	14.8	28.8	32.8	8.14	16.8	19.6
SHR	8.71	16.6	20.7	12.2	22.5	27.7	21.8	36.0	42.8	13.6	24.5	29.8
SWIPE	5.33	10.9	12.9	7.39	17.3	19.6	14.2	29.9	32.5	8.52	18.9	21.2
SWIPE'	9.07	13.0	16.1	13.6	19.8	23.2	20.7	31.6	34.9	14.3	21.1	24.4
WWB	2.63	10.2	11.3	2.84	13.4	14.5	6.50	22.0	24.0	3.70	14.8	16.2
MBSC	**1.51**	**6.54**	**7.20**	**1.67**	**10.2**	**10.9**	**2.87**	**20.5**	**21.3**	**1.90**	**11.9**	**12.6**

to incorporate a speech activity detector that can differentiate harmonic noise from voiced speech. Future work will explore these ideas further.

Acknowledgments

This material is based on work supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA or its Contracting Agent, the US Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the US Government. Approved for Public Release, Distribution Unlimited. The authors thank the editor and the reviewers for their comments.

Appendix A. Results for Keele (dev) corpus

See Tables A.10–A.15.

References

Ahmadi, S., Spanias, A.S., 1999. Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. IEEE Trans. Speech Audio Process. 7, 333–338.

Atal, B., Rabiner, L., 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Trans. Acoust. Speech Signal Process. 24, 201–212.

Bagshaw, P., Hiller, S., Jack, M., 1993. Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. In: Proc. European Conference on Speech Communication and Technology, pp. 1003–1006.

Beritelli, F., Casale, S., Serrano, S., 2007. Adaptive V/UV speech detection based on acoustic noise estimation and classification. Electron. Lett. 43, 249–251.

Camacho, A., Harris, J., 2008. A sawtooth waveform inspired pitch estimator for speech and music. J. Acoust. Soc. Am. 124, 1638–1652.

Cariani, P., Delgutte, B., 1996. Neural correlates of the pitch of complex tones. I: Pitch and pitch salience. Neurophysiology 76, 1698–1716.

Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111, 1917–1930.

Chu, W., Alwan, A., 2009. Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In: Proc. IEEE ICASSP, pp. 3969–3972.

Delgutte, B., 1980. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. J. Acoust. Soc. Am. 68, 843–857.

Drullman, R., 1995. Temporal envelope and fine structure cues for speech intelligibility. J. Acoust. Soc. Am. 97, 585–592.

Frerking, M.E., 1994. Digital Signal Processing in Communication Systems. Springer.

Hermes, D.J., 1988. Measurement of pitch by subharmonic summation. J. Acoust. Soc. Am. 83, 257–264.

Hirsch, H.G., 2005. Fant – filtering and noise adding tool. <http://dnt.kr.hs-niederrhein.de/download.html>.

ITU, 1996. Recommendation G.712: transmission performance characteristics of pulse code modulation channels.

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

Loughlin, P., Tacer, B., 1996. On the amplitude-and frequency-modulation decomposition of signals. J. Acoust. Soc. Am. 100, 1594–1601.

Luengo, I., Saratxaga, I., Navas, E., Hernaez, I., Sanchez, J., Sainz, I., 2007. Evaluation of pitch detection algorithms under real conditions. In: Proc. IEEE ICASSP, pp. 1057–1060.

Medan, Y., Yair, E., Chazan, D., 1991. Super resolution pitch determination of speech signals. IEEE Trans. Signal Process. 39, 40–48.

Meddis, R., 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.

Nakatani, T., Irino, T., 2004. Robust and accurate fundamental frequency estimation based on dominant harmonic components. J. Acoust. Soc. Am. 116, 3690–3700.

Oh, K., Un, C., 1984. A performance comparison of pitch extraction algorithms for noisy speech. In: Proc. IEEE ICASSP, pp. 85–88.

Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M., 1992. Complex sounds and auditory images. Auditory Physiol. Percept. 83, 429–446.

Plante, F., Meyer, G., Ainsworth, W.A., 1995. A pitch extraction reference database. In: Proc. European Conference on Speech Communication and Technology, pp. 837–840.

Rabiner, L., Cheng, M., Rosenberg, A., McGonegal, C., 1976. A comparative performance study of several pitch detection algorithms. IEEE Trans. Acoust. Speech Signal Process. 24, 399–418.

Ross, M., Shaffer, H., Cohen, A., Freudberg, R., Manley, H., 1974. Average magnitude difference function pitch extractor. IEEE Trans. Acoust. Speech Signal Process. 22, 353–362.

Rouat, J., Liu, Y., Morissette, D., 1997. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. Speech Commun. 21, 191–207.

Secrest, B., Doddington, G., 1982. Postprocessing techniques for voice pitch trackers. In: Proc. IEEE ICASSP, pp. 172–175.

Secrest, B., Doddington, G., 1983. An integrated pitch tracking algorithm for speech systems. In: Proc. IEEE ICASSP, pp. 1352–1355.

Shah, J., Iyer, A., Smolenski, B., Yantorno, R., 2004. Robust voiced/unvoiced classification using novel features and gaussian mixture model. In: Proc. IEEE ICASSP, pp. 17–21.

Sjölander, K., Beskow, J., 2000. WaveSurfer – an open source speech tool. In: Proc. Interspeech, pp. 464–467.

Slaney, M., Lyon, R.F., 1990. A perceptual pitch detector. In: Proc. IEEE ICASSP, pp. 357–360.

Sun, X., 2002. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In: Proc. ICASSP, pp. 333–336.

Talkin, D., 1995. Speech Coding and Synthesis, Elsevier, pp. 497–518 (Chapter: Robust Algorithm for Pitch Tracking (RAPT)).

Tan, L.N., Alwan, A., 2011. Noise-robust F0 estimation using SNR-weighted summary correlograms from multi-band comb filters. In: Proc. IEEE ICASSP, pp. 4464–4467.

Varga, A., Steeneken, H., 1993. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12, 247–251.

Walker, K., Strassel, S., 2012. The RATS radio traffic collection system. In: ISCA Odyssey.

Wu, M., Wang, D., Brown, G., 2003. A multipitch tracking algorithm for noisy speech. IEEE Trans. Speech Audio Process. 11, 229–241.