AUDITORY-INSPIRED PITCH EXTRACTION USING A SYNCHRONY CAPTURE FILTERBANK AND PHASE ALIGNMENT

Ramdas Kumaresan[*], *Vijay Kumar Peddinti*

Department of Electrical Engineering
Kelley Hall, University of Rhode Island
Kingston, RI 02881

Peter Cariani

Hearing Research Center &
Department of Biomedical Engineering
Boston University, Boston, MA 02114

ABSTRACT

The question of how harmonic sounds produce strong, low pitches at their fundamental frequencies, f_0s, has been of theoretical and practical interest to scientists and engineers for many decades. Currently the best auditory models for f_0 pitch, e.g. [1], are based on bandpass filtering (cochlear mechanics), half-wave rectification and low-pass filtering (haircell transduction and synaptic transmission), channel autocorrelations (all-order interspike interval statistics) aggregated into a summary autocorrelation, and an analysis that determines the most prevalent interspike intervals. As a possible alternative to autocorrelation computations, we propose an alternative model that uses an adaptive Synchrony Capture Filterbank (SCFB) in which groups of filters or channels in a filterbank neighborhood are driven exclusively (captured) by dominant frequency components that are closest to them. The channel outputs are then adaptively phase aligned with respect to a common time reference to compute a Summary Phase Aligned Function (SPAF), aggregated across all channels, from which f_0 can be easily extracted.

Index Terms— Synchrony Capture, Frequency Capture, Pitch Extraction, Filterbank

1. INTRODUCTION

Pitch is an essential attribute of quasi-periodic acoustic signals in speech, music, and other listening contexts [2, 3, 4]. For a quasi-periodic sound, the dominant pitch is almost invariably heard at its fundamental frequency f_0. Common periodicity ("harmonicity," sharing of common subharmonics), along with common onset, play very strong roles in grouping frequency components into auditory objects, and separating out multiple objects, each evoking its own pitch. Neural pitch mechanisms thus appear to be intimately related to early auditory grouping mechanisms, which in turn render analyses of multiple objects and streams in an auditory scene much more tractable. Human listeners are presently far superior to artificial, machine listening systems when it comes to tracking and analyzing sounds in noisy, cluttered, real world acoustic environments. If the operating principles inherent in neural mechanisms for pitch and auditory grouping can be understood and emulated, better artificial speech and music recognition systems and auditory prostheses are likely to follow. With this in mind, a method for extracting pitches of harmonic sounds is proposed, that may have parallels with signal processing strategies employed by auditory systems.

Currently three broad classes of f_0 pitch models exist: spectral pattern-matching models, residue models, and temporal autocorrelation models [5]. Spectral pattern-matching models first carry out a frequency analysis and then match patterns of resolved frequency components to harmonic spectral templates, so as to infer f_0 [6, 7]. Here, sinusoidal components that are exclusively represented in the output of a filter or channel are said to be "resolved," whereas signal components that interact within the passband of a filter are referred to as "unresolved." Residue models posit that f_0 pitch arises from (beating) interactions between (nearby) unresolved harmonics that are produced by broad cochlear filtering. A temporal analysis of the resultant beating patterns produces an estimate of f_0. Thus spectral pattern models predict only the f_0 pitches of resolved harmonics, whereas residue models predict only those of unresolved harmonics. Temporal autocorrelation models analyze patterns of all-order interspike intervals produced in the auditory nerve to identify patterns of interval peaks associated with different f_0 pitches [8, 9]. These models predict f_0 pitches produced by both resolved and unresolved harmonics. The underlying neuronal mechanisms proposed for temporal autocorrelation-based analysis and separation, utilize neural delay lines and coincidence detectors [10, 11, 12].

Meddis and co-workers [1, 13] have proposed a popular autocorrelation-based model that draws upon the original work of Licklider [10]. The model simulates cochlear action (cochlear bandpass filtering, transductive half-wave rectification, and synaptic low-pass filtering in each channel) to produce spike timing probability distributions (PSTHs, post-stimulus time histograms) for auditory nerve fibers (ANFs) of all characteristic frequencies (CFs). The autocorrelation of each fiber's PSTH is computed, and all of these auditory nerve frequency-channel autocorrelations are summed to produce the summary autocorrelation function (SACF) for the entire auditory nerve array. In effect, the SACF provides an autocorrelation-like representation of the acoustic signal. Major peaks in the SACF are identified, and the resultant f_0 pitch estimates successfully predict an extremely wide range of human pitch judgments. However, neuronal mechanisms by which the auditory system might analyze SACFs in the form of population-wide interspike interval statistics have yet to be found, motivating the search for alternative signal processing strategies that realize analysis operations similar to summary autocorrelation.

Recently, we have developed signal processing algorithms that emulate the synchrony capture phenomenon in the auditory nerve [14] that may afford alternative strategies for utilizing neural spike timing information. If one examines the representation of complex sounds in the auditory nerve, a striking feature is "synchrony capture," wherein nerve fibers in an entire cochlear CF region are driven almost exclusively by one dominant local frequency component, (see, for example, [15]), such that the individual component imposes its (largely unmodulated) temporal fine structure (TFS) on the timing pattern of spikes in that region. At moderate and high sound

[*]This work was supported by the Airforce Office of Scientific Research under the grant # AFSOR FA9550-09-1-0119.

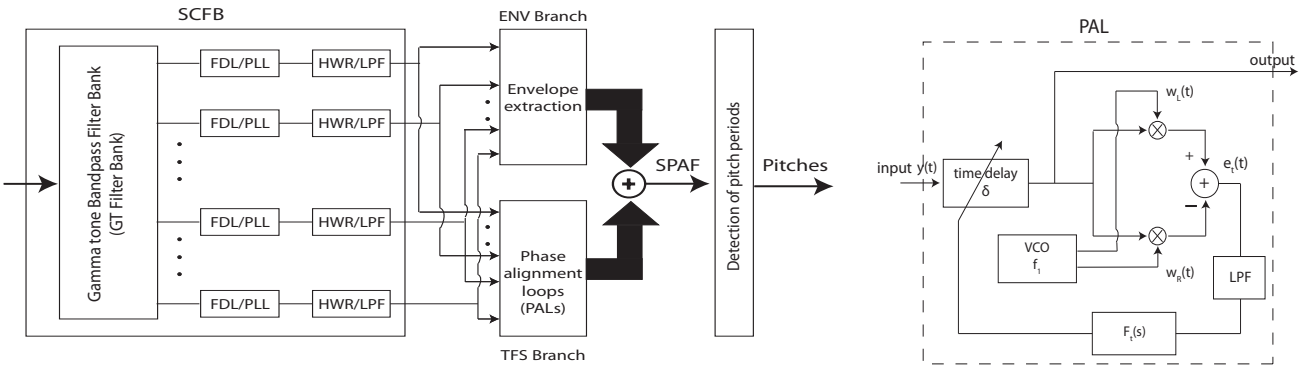


Fig. 1. **a) Schematic of the Pitch Extraction Algorithm:** The gammatone (GT) filters provide some spectral isolation in each channel. The frequency discriminator loop/phase locked loop (FDL/PLL) block achieves synchrony capture, i.e., the voltage controlled oscillator (VCO) in the block locks on to the strongest frequency component. The half wave rectified and low pass filtered (HWR/LPF) output of the channel is processed by the envelope (ENV) and the temporal fine structure (TFS) branches in parallel, which extract the envelope and phase align the dominant tonal signals, respectively. The outputs of all the channels, both the ENV and TFS branches, are summed to produce the Summary Phase Aligned Function (SPAF), the peak locations of which are used to determine f_0. **b) Phase alignment loop (PAL):** The half wave rectified tone, $y(t)$, is delayed such that it overlaps symmetrically on the left ($w_L(t)$) and right ($w_R(t)$) windows (depiction shown in figure 3(b)). Analogous to the FDL, when $y(t)$ is centered around a time reference, t_0, the error $e_t(t)$ goes zero and the loop reaches steady state. At this point, the half wave (HW) rectified tone is in cosine phase with respect to t_0. See Section 4.

pressure levels, a dominant harmonic can drive a large swath of auditory nerve fibers with CFs spanning an octave or more. For harmonics that are sufficiently separated, synchrony capture enhances their global temporal (interspike interval) representation by suppressing the temporal representation of beating interactions between harmonics. For harmonics closer together, within roughly a critical band or so, ANFs in surrounding CF regions are instead driven by the composite waveform pattern of the two interacting harmonics, such that the interspike interval representation of individual harmonics is severely degraded. Thus, neural synchrony capture appears to parallel perceptual frequency selectivity and harmonic resolution. Since resolved harmonics are known to permit separation of concurrent sounds by human listeners, synchrony capture in artificial systems may likewise be exploited for better sound separations.

The algorithm proposed here further develops the synchrony capture filterbank (SCFB) architecture presented in [14] and extends it to extract pitch frequencies of harmonic signals. The signal input may consist of resolved and/or unresolved components and additive noise. The key component of the algorithm, the SCFB architecture (see Figure 1), consists of a bank of broadly tuned filters (á la basilar membrane) in cascade with narrower filters (á la outer hair cells) that adaptively lock onto locally-dominant frequency components to produce synchrony capture behavior. The narrower filters constitute a frequency discriminator loop (FDL) and are able to track individual time-varying frequency components, such as low harmonics and the dominant harmonics associated with formants in speech, in the midst of noise. In this article, we modified the algorithm to work even when there are unresolved tones in the input (see section 3).

The schematic of the entire pitch extraction algorithm is shown in Figure 1 (a). The input signal, consisting of multiple, harmonically-related tones, is first filtered by a logarithmically spaced Gammatone (GT) filterbank. The output of each filter is then processed by a frequency discriminator loop/phase locked loop (FDL/PLL) block which determines the frequency of the dominant tone present in the passband of the associated GT filter. The details of this block are described in section 3 and shown in Figure 2. Each filter channel output is then halfwave-rectified and low pass filtered (by HWR/LPF block) and then delivered in parallel to the Envelope (ENV) branch and the Temporal Fine Structure (TFS) branch. In the ENV branch the HWR/LPF output is further low pass filtered to extract the en-

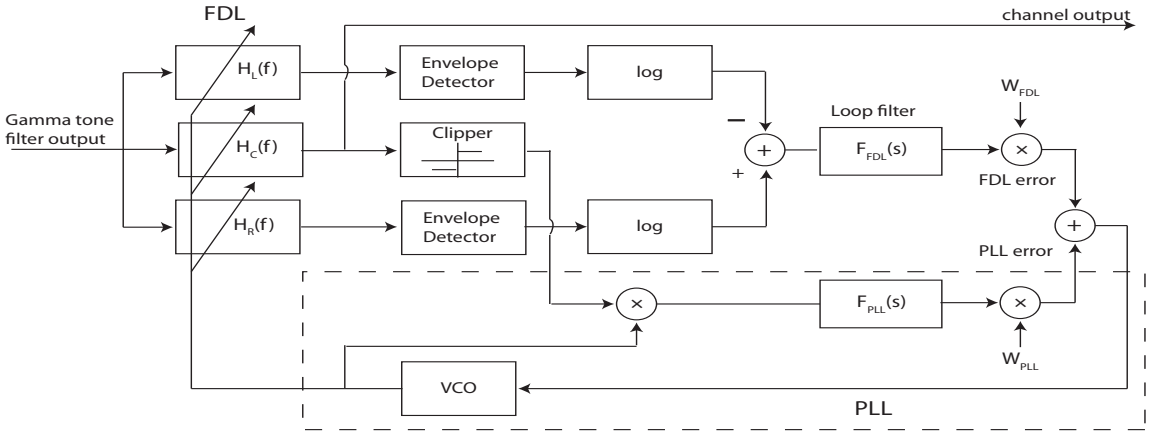
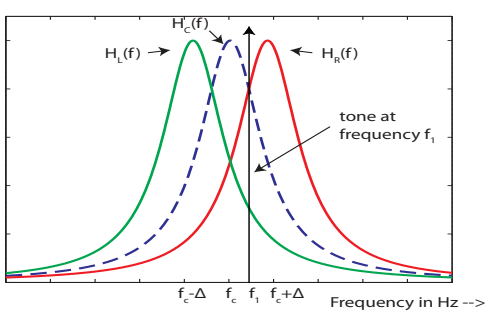
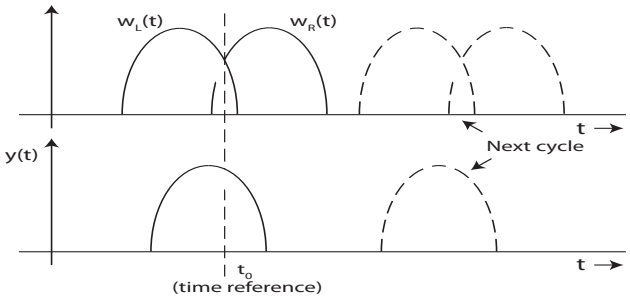


Fig. 2. **FDL and PLL:** Together they track the frequency of the dominant tone in the passband of the associated GT filter.



(a)



(b)

Fig. 3. **a) Frequency measurement using stagger tuned filters:** The left $H_L(f)$, right $H_R(f)$ (and center $H_C(f)$) filters are tuned by adjusting the frequency f_c of a voltage controlled oscillator (VCO), which is embedded in a feedback loop (See Figure 2). When $f_c = f_1$, the tone's amplitude at the output of the right and left filters is equal, hence the loop reaches steady state. **b) Phase measurement using staggered time windows:** The half wave (HW) rectified tone, $y(t)$ (shown in bottom panel), is multiplied separately by the left ($w_L(t)$) and right ($w_R(t)$) windows. $y(t)$ is delayed using a feedback loop until the areas under $y(t)w_L(t)$ and $y(t)w_R(t)$ are equal, and the loop reaches steady state. Then, the delayed HW rectified tone is in cosine phase with respect to a time reference, t_0.

velope. In the TFS branch the halfwave rectified signals are aligned in phase with respect to a common time reference, t_0, using a Phase Alignment Loop or PAL (see section 4). The FDL/PLL block, also provides a continuous estimate of the dominant frequency to the PAL. Once the channel outputs in the TFS branches are aligned in phase, the signals across the TFS *and* ENV branches of all channels are aggregated to obtain the Summary Phase Aligned Function (SPAF), which is then used to extract the f_0 information. Simulation results are presented in Section 5.

2. TWO MAIN IDEAS: DUALS IN TIME AND FREQUENCY

The proposed algorithm uses two basic signal processing strategies, one in the frequency domain and the other in time domain. Consider a sinusoidal signal $x(t) = A\cos(2\pi f_1 t + \theta_1)$. The first goal is to determine the frequency f_1. It is determined using a frequency discriminator loop (FDL), the details are in Section 3. The basic principle is shown in Figure 3(a). The tone, shown as an impulse in Figure 3 (a), is fed as an input to stagger-tuned left ($H_L(f)$) and right ($H_R(f)$) bandpass filters (and a center filter $H_C(f)$). Assume that these three frequency responses can be shifted in tandem along the frequency axis with the help of a VCO, whose frequency is adjustable using a feedback loop based on the difference in amplitude of the sinusoid at the output of $H_L(f)$ and $H_R(f)$. The loop ultimately settles to a steady state when the amplitude difference at the outputs of $H_L(f)$ and $H_R(f)$ is zero and the VCO frequency coincides with input frequency, i.e., $f_c = f_1$. Such FDLs have been used in automatic frequency control and communication systems for decades. We modified this basic FDL in Section 3 to work when the input consists of unresolved tones as well.

The second idea shown in Figure 3(b) is the time domain dual of the FDL. It is called a Phase Alignment Loop or PAL. The goal is to align the phase of the sinusoid such that it is in cosine phase with respect to an arbitrary time reference, t_0. This adaptive phase alignment obviates the need to compute the autocorrelation function for each channel. In Figure 3(b) we assume that f_1 is known. The half wave rectified tone, named $y(t)$, is multiplied separately by the left ($w_L(t)$) and right ($w_R(t)$) windows, which are centered around some arbitrary time reference, t_0. The error signal, e(t) is the difference in areas under the curves $y(t)w_L(t)$ and $y(t)w_R(t)$. Later e(t) is smoothed and used to delay $y(t)$. Analogous to the FDL, when $y(t)$ is centered around t_0, error approaches zero and the loop reaches a steady state. See Section 4 for details. The delayed HW rectified tone is then in cosine phase with respect to t_0.

3. FREQUENCY TRACKING BY THE FDL/PLL BLOCK

The key element of the SCFB is the FDL/PLL block, the details of which are shown in Figure 2. It consists of three filters with frequency responses $H_L(f)$, $H_C(f)$, and $H_R(f)$ (which are spaced Δ Hz apart as shown in Figure 3 (a)). These filters are gang-tuned with the help of the VCO (See [14], our prior work for details). The difference between the log-amplitudes at the output of $H_L(f)$ and $H_R(f)$ (called FDL error) is used to adjust the VCO frequency which then moves these frequency responses in such a way to drive the log-amplitude difference to zero. If the input is a single tone then the FDL error approaches zero and the VCO frequency coincides with input frequency, i.e., $f_c = f_1$. These FDLs are used in cascade with a GT filterbank in the SCFB [14]. A key attribute of the FDL is that it exhibits the synchrony capture property similar to that seen in the auditory nerve. However, when the input signal consists of interfering sinusoids, i.e., more than one tone falls within the passband of a filter channel (unresolved case) the FDL tends to produce a biased estimate of the dominant frequency. Hence, we propose a combination of an FDL and a PLL which operates on the output of the center filter $H_C(f)$ (see Figure 2) to ameliorate this problem (details are described in [16]). The PLL shown within dashed lines in Figure 2 is a standard PLL, but the PLL error and the FDL errors are weighted to emphasize the importance of one or the other. For low frequency channels the GT filters are narrow and hence the interfering components are already sufficiently attenuated. Therefore, for these low frequency channels the FDL error alone is adequate. However, for high frequency channels the GT filters are wider and hence invariably have interacting tones within their passband. In this (high frequency channels) case, the FDL drives the VCO within the lock-in range of the PLL and the PLL plays a vital role by homing in on the nearby dominant tone (See Figure 5 for simulation result). Hence, the PLL error is weighted more to reduce the bias in frequency estimate.

4. PHASE ALIGNMENT LOOP (PAL)

Figure 1 (b) shows the schematic of the PAL. The half wave rectified tone (output of the SCFB), $y(t)$, is multiplied by the left ($w_L(t)$) and right ($w_R(t)$) windows, which are centered around some arbitrary time reference, t_0. These windows are supplied by the VCO at the rate of frequency f_1, which is obtained from the FDL/PLL block.

The area under the error waveform $e_t(t) = y(t)\{w_L(t) - w_R(t)\}$ is smoothed by the LPF and used to adjust the time delay δ. Analogous to the FDL, when the HW rectified tone $y(t)$ is centered around t_0 (actually, $t_0 + n/f_1$, where n is an integer), the error approaches zero and the loop reaches steady state. At that point the rectified tone is in cosine phase with respect to t_0 (simulation shown in Figure 4 (c)).

5. SIMULATION RESULTS

The pitch extraction algorithm has been tested with and without noise on several synthetic signals. Here, we present some results for a single tone, two unresolved tones and musical notes composed of several harmonic components, but without noise. The SCFB used here has 64 logarithmically spaced GT filters spanning 100 to 3827 Hz. The sampling frequency is 16 kHz. The filter Q values are all 6. Each GT filter is in cascade with the three filters in the FDL/PLL block, and each of the triplet filters have half the bandwidth of the GT filter. First, SCFB is input with a single tone at 800 Hz. Figure 4 (a) shows the converged frequency tracks of the VCOs of four channels in the neighborhood of 800 Hz (the center frequencies of the four GT filters and the VCOs' initial frequencies are 660, 696, 734, 773 Hz). As the tone is passed through the four GT filters it gets attenuated and phase delayed by each filter differently (figure 4 (b)). At the time reference, t_0, the PALs start to phase align these four channel outputs. Figure 4 (b) & (c) shows the HW rectified tones before and after phase alignment. These phase aligned outputs (figure 4 (c)) can be coherently added (SPAF) to obtain the pitch of the tone.

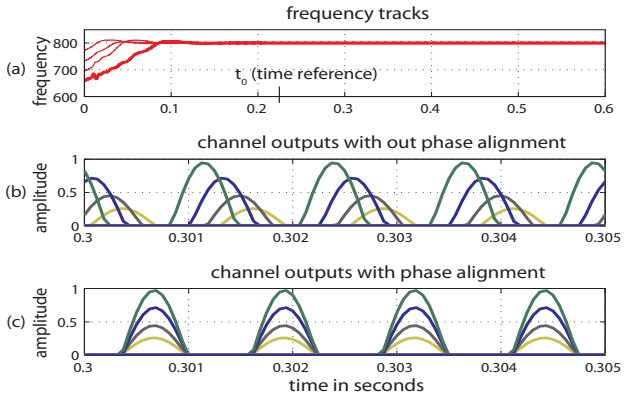


Fig. 4. **Single tone:** (a) shows the frequency trajectories of the 4 VCOs near 800 Hz, (b) and (c) show the HW rectified channel outputs without and with phase alignment, respectively.

In the second example, the input signal consists of two equal amplitude tones at frequencies 2300 and 2500 Hz. Around 2400 Hz, the 10 dB bandwidth of the GT filters is about 200 Hz. Since the frequencies fall in the same filter, this is an example of the unresolved case. Figures 5 shows the VCO frequency tracks for four channels with center frequencies 2234, 2338, 2447, and 2560 Hz, with and with out the inclusion of PLL error, figure 5 (a) & (b) respectively. It can be seen that the bias in frequency estimates is zero when the PLL error is included, and there is a noticeable bias when PLL error is included {figure 5 (a) & (b) respectively}.

The final example is a musical note with resolved harmonics, here the sum of the first six harmonics of the fundamental $f_0 = 440$ Hz. The tones are chosen to have a Schroeder phase, that is, the input is $x(t) = \sum_{k=1}^{6} \cos(2\pi k f_0 t + \pi k^2/6)$. Such signals typically exhibit little envelope variability in a pitch period. Figure 6 shows the frequency tracks of the VCOs of different channels. The synchrony capture phenomenon, (i.e., the VCO frequencies in the neighborhood of a dominant tone lock on to that tone's frequency,) is obvious. The sum of all the phase aligned signals, i.e., the SPAF is shown in Figure 7. The SPAF clearly shows peaked periodic pattern (unlike the input signal) and the pitch information can be extracted from the SPAF by finding the interval between the largest peaks.

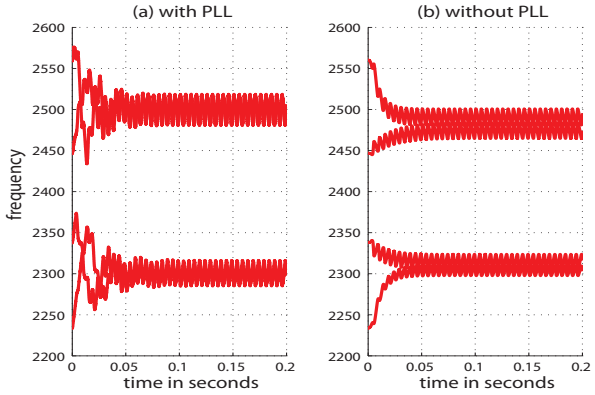


Fig. 5. **Unresolved tones:** Tracks of VCO frequencies for CFs around 2400 Hz with PLL (a) and without PLL (b, notice the bias.)

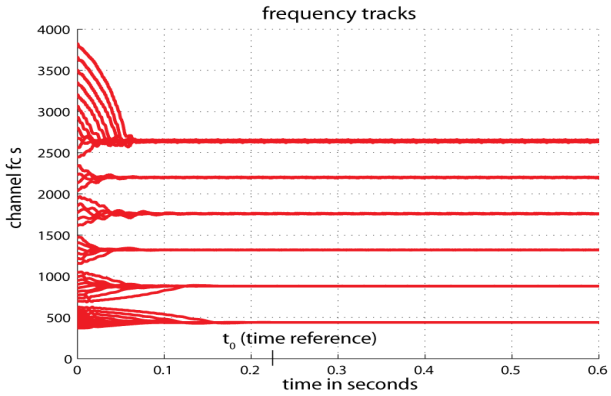


Fig. 6. **Frequency tracks of the Schroeder-phase signal**

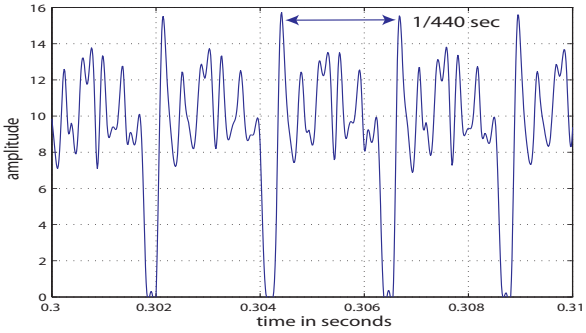


Fig. 7. **SPAF of Schroeder-phase signal**

6. CONCLUSION

We proposed a new Summary Phase Aligned Function (SPAF) as an alternative to Summary Autocorrelation Function (SACF) for computing the fundamental frequency f_0 of a periodic signal. SCAF requires autocorrelation computations and SPAF does not. We also modified our previous SCFB algorithm [14] to improve tone resolution. The simulation results on synthetic signals are promising but need to be performed on real world signals.

7. REFERENCES

[1] 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, no. 6, pp. 2866–2882, 1991.

[2] C. J. Plack and A. J. Oxenham, "Psychophysics of Pitch," in *Pitch: Neural Coding and Perception*, C. J. Plack, A. J. Oxenham, R. R. Fay and A. N. Popper, Ed., Springer Handbook of Auditory Research, chapter 2, pp. 7–55. Springer-Verlag, New York, NY, 2005.

[3] A. de Cheveigne and H. Kawahara, "Yin: A fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, 2002.

[4] D. L. Wang and G. J Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, pp. 684–697, May 1999.

[5] A. de Cheveigne, "Pitch perception," in *The Oxford Handbook of Auditory Science*, Christopher J. Plack, Ed., pp. 71–104. Oxford, Oxford University Press Inc, New York, 2010.

[6] J.L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1496–1516, 1973.

[7] P. Srulovicz and J.L. Goldstein, "A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum," *J. Acoust. Soc. Am.*, vol. 73, no. 4, pp. 1266–1276, 1983.

[8] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. I. pitch and pitch salience. II. pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch.," *J. Neurophysiology*, vol. 76, no. 3, pp. 1698–1734, 1996.

[9] R. Meddis and M. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 233–245, 1992.

[10] J.C.R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. VII, no. 4, pp. 128–134, 1951.

[11] P. Cariani, "Neural timing nets," *Neural Networks*, vol. 14, no. 6-7, pp. 737–753, 2001.

[12] P Cariani, "A temporal model for pitch multiplicity and tonal consonance," in *Proceedings of the 8th International Conference on Music Perception & Cognition (ICMPC8, Evanston, IL)*, S.D. Lipscomb, R. O. Ashley, R. O. Gjerdingen, and P. Webster, Eds., pp. 310–314. Causal Productions, Adelaide, Australia, 2004.

[13] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J Acoust Soc Am*, vol. 102, no. 3, pp. 1811–20, 1997.

[14] R. Kumaresan, Vijay Kumar Peddinti, and P. Cariani, "Synchrony Capture Filterbank: Auditory-inspired signal processing for tracking individual frequency components in speech," *J. Acoust. Soc. Am.*, vol. 133, Issue 6, pp. 4290–4310, 2013.

[15] B. Delgutte and N.Y.S. Kiang, "Speech coding in the auditory nerve: I. Vowel-like sounds," *J. Acoust. Soc. Am.*, vol. 75, pp. 866–878, 1984.

[16] R. Kumaresan, Vijay Kumar Peddinti, and P. Cariani, "Auditory-inspired pitch extraction using synchrony capture filterbank and phase alignment (To be submitted)," *J. Acoust. Soc. Am.*, 2013.