Real-time pitch detection on the PDA for cochlear implant applications

Rohith Ramachandran and Philipos C. Loizou, *Senior Member, IEEE*

***Abstract*— There is currently a need in cochlear implants to develop speech coding algorithms that provide better access to pitch cues, known to be critical for music perception. Such algorithms require an estimate of the fundamental frequency (F0). This paper presents the implementation of a real-time pitch (F0) detector on a Personal Digital Assistant (PDA). The pitch detection algorithm is based on the autocorrelation function and is implemented real-time on a Dell AXIM Pocket PC. Its performance, in terms of F0 accuracy, is compared against that obtained by the pitch detection algorithm used in STRAIGHT. The implementation details and real-time performance measurements are also provided.**

I. Introduction

The coding of pitch information for cochlear implant (CI) users has been a long-standing challenge. Because of that, most CI users are not able to enjoy or appreciate music [1]. Pitch information can be conveyed in cochlear implants via temporal and/or spectral (place) cues [2,3]. Temporal cues are present in the envelope modulations of the band-pass filtered waveforms. Pitch may also be conveyed by the electrode place of stimulation due to the tonotopic arrangement of the electrodes in the cochlea. Stimulation of apical electrodes elicits low pitch percepts while stimulation of basal electrodes elicits higher pitch percepts. Several researchers investigated the possibility of providing explicit pitch information to CI users by enhancing access to pitch cues present in the envelope modulations. The development of such or other novel speech coding strategies that would enhance access to pitch information requires at the very least an estimate of the fundamental frequency (F0).

The present study focuses on the development of an F0 detection algorithm that operates in real-time. An F0 estimator that operates in the frequency domain was proposed in [4] for processing of musical sounds. In our implementation, the F0 estimator operates in the time-domain and is based on the autocorrelation [5]. Arguably there is large number of (more) sophisticated pitch-detection algorithms (e.g., [6]), but we chose this algorithm because of its simplicity and ease of implementation. The implementation of this pitch-detector was carried out on a Dell AXIM Pocket PC using C. Intel Performance Primitives (IPP) library functions for the Intel processor were also used to improve performance. The developed F0 estimator was evaluated using vowel recordings and compared against the F0 estimator used in the STRAIGHT algorithm [7].

Manuscript received August 24, 2007. This work was supported by the National Institutes of Health (NIH/NIDCD) under Contract NO1-DC-6-0002.

R. Ramachandran and P. Loizou are with the Department of Electrical Engineering, University of Texas-Dallas, Richardson, TX 75083, USA (e-mail: rohith.ramachandran@student.utdallas.edu, corresponding author: loizou@utdallas.edu).

II. F0 Detection Algorithm

The vocal pitch (F0) was determined using the autocorrelation function. The autocorrelation function of the signal x[n] is given by:

$$R_{xx}[\tau] = \sum_{n=-\infty}^{n=\infty} x[n]x[n+\tau]$$

It is known that the autocorrelation function of a periodic signal produces peaks at integer multiples of the period (reciprocal of fundamental frequency F0) of the signal. Therefore, when the autocorrelation of a voiced segment of speech is computed, peaks are produced at the intervals equal to the period of the voiced signal. From this, we can determine F0.

The algorithm used to estimate F0 is based on the algorithm proposed in [5] and is outlined below:

1. Speech is windowed in 20-ms segments.
2. The number of zero-crossings is obtained for each speech segment. If the number of zero-crossings is above a certain threshold (typically 20 for a 20ms window), the segment is considered unvoiced and the pitch (F0) is set to zero.
3. If the number of zero-crossings is less than 20 then the following steps are executed. The signal is first centre-clipped based on a threshold level of 70% of the maximum amplitude within each segment. That is, amplitudes falling below the threshold level are set to zero, and amplitudes falling above the threshold level are set to the input level minus the threshold level.
4. The autocorrelation and the energy of the centre-clipped signal are computed. The maximum of the autocorrelation function is searched within an interval corresponding to an F0 range of 60 Hz to 320 Hz. If the peak amplitude is smaller than 40% of the estimated energy, then the segment is treated as silence. Otherwise, the F0 value is estimated based on the location of the maximum of the autocorrelation function.

The above procedure is repeated for all speech segments. A median filter is subsequently used to smooth the estimated F0 values. The F0 value is first passed through a 5-point median filter followed by a 3-point median filter. The resulting F0 value is subtracted from the original pitch value. The difference is again passed through a 5-point median filter followed by a 3-point median filter. The filter outputs are finally added to obtain the smoothed values of F0.

978-1-4244-1626-4/07/$25.00 ©2007 IEEE.

III. Software Implementation Details

In the PDA implementation of the pitch detection algorithm, the speech signal was acquired from the microphone of the PDA, and sampled at 22 kHz with 16-bit resolution. The volume level of the microphone was set to a minimum value above zero to prevent saturation of the speech amplitude.

F0 was computed on a frame-by-frame basis and the F0 values of 20 consecutive frames were collected and smoothed before displaying the computed F0 value on the PDA screen. Each frame consists of 512 samples, corresponding to a frame-duration of 23.2ms.

As mentioned above, each speech frame is first evaluated for being voiced or unvoiced. Voiced/unvoiced decisions are made based on zero crossings counts. To determine the appropriate threshold for making voiced/unvoiced decisions, we conducted some preliminary experiments using microphone recordings from the PDA. A 5-sec recording of speech consisting primarily of voiced utterances (such as vowels) was used to compute histogram of the number of zero-crossings for voiced segments. Similarly, a 5-sec recording of speech consisting primarily of unvoiced syllables (e.g., fricatives) was used to compute histogram of the number of zero-crossings for unvoiced segments. The histograms of voiced and unvoiced segments are shown in Fig. 1. From that histogram, we determined that an appropriate threshold value for the voiced/unvoiced distinction was 20.

Experimentally, we found that some computation can be avoided in detecting a silent frame by simply comparing the maximum magnitude of the samples in a frame with 2300 (this threshold value can arguably be modified depending on the microphone level). If the maximum magnitude of the samples in the frame is less than 2300, the frame is then assumed to be a silence frame and the F0 is set to zero. The silence detection procedure described in the previous section is implemented for frames having maximum magnitudes greater than 2300.

The F0 values of 20 successive frames of 512 samples each are then smoothed and displayed, using Windows Application Program Interfaces (APIs) functions, to the screen of the PDA (see PDA snapshot in Fig. 2). In our implementation, the linear filters used in the smoothing technique in [8] were eliminated as they were found to produce dips in the pitch contour. Intel Performance Primitives (IPP) library functions were used to implement the autocorrelation and median-filtering functions. All arithmetic operations involved were done in fixed-point.

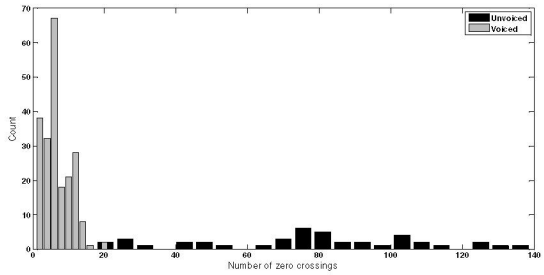

Fig.1: Histograms of zero-crossings for voiced and unvoiced utterances recorded on the PDA.

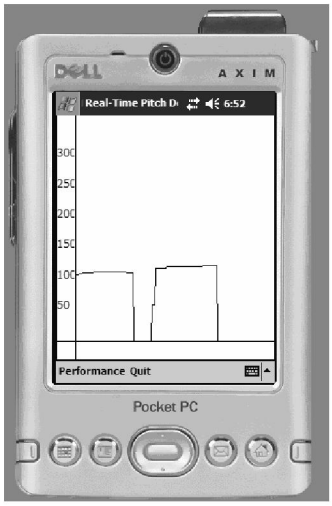

Fig. 2: Snapshot of the F0 contour of the syllable /a s a/ displayed in real-time on the PDA.

IV. Evaluation and Comparison of Pitch-Detector

The accuracy of the F0 estimate obtained with our pitch-detection algorithm was compared against the accuracy of the F0 estimator used in the STRAIGHT algorithm [7]. To compare the accuracy of the pitch-detector with STRAIGHT, we made use of an off-line version of the pitch-detector which takes as input, speech signals from a stored file (.wav format) rather than from the PDA microphone. This version still ran on the PDA, but processed stored speech files and saved the estimated F0 contours in a file. A total of 29 vowel recordings (in quiet), taken from the vowel database in [9], were used in the evaluation. Vowels in /hVd/ format produced by roughly an equal number of adult male and female speakers, and five boys and five girls were used. Assuming that STRAIGHT provides an accurate estimate of F0, we computed the normalized error (NE) between the F0 value estimated via our algorithm and that obtained by STRAIGHT:

$$NE = \frac{| f_{PDA} - f_{STRAIGHT} |}{f_{STRAIGHT}} \times 100$$

Table 1 shows the normalized error (NE) values, along with the mean F0 values computed via our implementation and that of STRAIGHT. As can be seen, the estimated F0 values computed on the PDA were quite accurate and comparable to the values obtained by STRAIGHT. Figure 3 shows the F0 contour obtained with our implementation and that of STRAIGHT for the vowel in "hood" produced by an adult female speaker. As can be seen, the F0 contour produced by our F0-detection algorithm follows closely that produced by STRAIGHT.

TABLE I
AVERAGE VALUES OF PITCH AND NORMALISED ERROR

Speaker type	AVERAGE PITCH FROM STRAIGHT	Average pitch from PDA	Average normalised error (%)
Boys	246.13 Hz	232.34 Hz	3.65 %
Girls	224.76 Hz	214.64 Hz	4.67 %
Men	122.22 Hz	117.35 Hz	3.94 %
Women	221.42 Hz	211.49 Hz	4.67 %

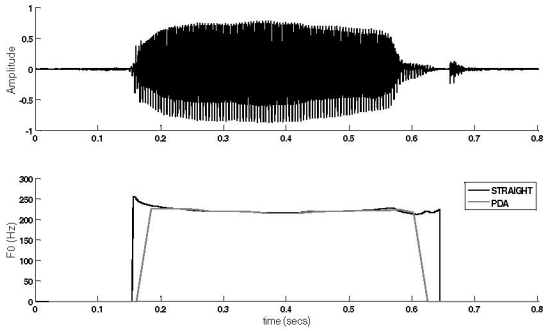

Fig. 3: F0 contour (bottom panel) obtained using our F0-detection algorithm and STRAIGHT for the word "hood" produced by an adult female speaker.

V. PROFILING RESULTS

The implementation of the F0 detection algorithm involves three components: zero-crossings estimation, autocorrelation estimation and post-processing of the estimated F0 values. To assess the computational load involved in each of the three components, we performed profiling on the timing involved in each component. Table II shows the profiling measurements. This table gives the time taken to process 20 frames by each of the three components.

TABLE II
PROFILING MEASUREMENTS

Speaker	Zero-crossings		Autocorrelation and voiced-unvoiced detection		Median-filtering & post-processing		Total
	Time (ms)	% load	Time (ms)	% load	Time (ms)	% load	Time (ms)
Boys	1.6	11.3	13.33	94.1	0.83	5.8	14.16
Girls	1.6	12.6	11.83	93.4	0.83	6.5	12.67
Men	1.6	13.6	11.27	93.0	0.83	6.9	12.10
Women	1.6	11.9	12.97	94.1	0.79	5.8	13.75

As expected, most of the time is spent on computing the autocorrelation function (note that we used Intel's IPP routine to implement the autocorrelation function). The total duration of 20 frames is 464.4ms. Therefore, to process 464.4 ms of input, it takes on the average 13.17ms. This implies that F0 is estimated about 35 times (=464.4/13.17) real-time.

VI. COCHLEAR IMPLANT APPLICATIONS

The F0 estimate obtained via the above algorithm can be used in several cochlear implant applications. For one, it can be used in the implementation of the original channel vocoder proposed by Dudley in the 1940s [10]. Several variations of the channel vocoder are currently being used to study the performance of cochlear implant patients in the absence of confounding factors (e.g., insertion depth, duration of deafness, etc). Use of vocoder simulations has proven to be a valuable tool in cochlear implant research. Results with vocoder simulations obtained with normal-hearing listeners have been found to be consistent with outcomes with cochlear implant patients.

The original channel vocoder [10] consists of a speech analyzer and a speech synthesizer. The incoming signal is first filtered into a number of contiguous frequency channels using a bank of band-pass filters and the envelope of the signal in each channel is estimated by full-wave rectification and low-pass filtering. In addition to envelope estimation, the vocoder analyzer makes a voiced/unvoiced decision and estimates the vocal pitch (F0) of the signal. These two pieces of information are transmitted alongside the envelope information. The synthesizer modulates the received envelopes by the appropriate excitation as determined by the voiced/unvoiced (binary) signal. The excitation signal consists of random noise for unvoiced speech segments and a periodic pulse generator for voiced speech, with the period of the pulse generator being controlled by F0. The modulated signals are subsequently bandpass-filtered by the same filters and then added together to produce the synthesized speech waveform.

Current cochlear implant processors (sixty years later) utilize the same blocks of the channel vocoder analyzer. At present, only the vocoder analyzer is used for transmitting envelope information to the individual electrodes, but recently there has been a shift in research focus toward implementing blocks of the synthesizer as well [11,12]. Interestingly, early devices based on feature extraction strategies modulated the estimated formant amplitudes by F0 [1]. These strategies, however, were abandoned due to the inherent difficulties associated with F0 extraction in noisy environments. It is also interesting to note that the acoustic simulations often used to study performance of cochlear implant patients in the absence of confounding factors (e.g., duration of deafness, insertion depth) utilize the synthesizer. By choosing random noise as the excitation signals for all segments of speech, we get the noise-band cochlear implant simulations [13]. Similarly, by choosing sine waves with frequencies set to the center frequencies of the bandpass filters as the excitation signals, we get the sine wave simulations [14].

The above strategies were originally designed to convey speech information but fall short on many respects in conveying adequate vocal pitch (F0) information. Speakers of tonal languages, such as Cantonese and Mandarin, make

use of vocal pitch variations to convey lexical meaning. Several researchers have demonstrated that CI users fitted with current strategies have difficulty discriminating between several tonal contrasts. Also, CI users are not able to perceive several aspects of music including identification of familiar melodies and identification of musical instruments. Hence, strategies designed to improve coding of F0 information are critically important for better tonal language recognition and better music perception. Such strategies would require an estimate of F0.

Pitch information can be conveyed in cochlear implants via temporal and/or spectral (place) cues. Temporal cues are present in the envelope modulations of the band-pass filtered waveforms Pitch can be elicited by varying the stimulation rate (periodicity) of a train of stimulus pulses presented on a single electrode, with high pitch percepts being elicited by high stimulation rates, and low pitch percepts being perceived by low stimulation rates. Once the stimulation rate increases beyond 300 Hz, however, CI users are no longer able to utilize such temporal cues to discriminate pitch [3]. Pitch may also be conveyed by electrode place of stimulation. Access to spectral cues is limited however, by the number of electrodes available (ranging from 12-22 in commercial devices), current spread causing channel interaction and possible pitch reversals due to suboptimal electrode placement. A number of strategies have been proposed to enhance spectral (place) cues and/or temporal cues, and these strategies are reviewed in [1]. Access to an estimate of F0 is needed in all these strategies.

VII. SUMMARY

The present paper presented the development of a pitch-detection algorithm which operates in real-time on the PDA. Such an implementation can be used and incorporated in speech coding algorithms aimed at improving or enhancing the pitch perception ability of cochlear implant users. Such algorithms will hold promise in allowing cochlear implant users better appreciate music as well as for improving tonal language recognition. The developed pitch-detection algorithm will eventually be incorporated on a portable research platform based on the PDA [15]. This portable platform will allow for chronic studies of novel strategies designed to enhance access to F0 cues.

REFERENCES

[1] P. Loizou, "Speech Processing in Vocoder-Centric Cochlear Implants," in *Cochlear and Brainstem Implants*, (Moller, Ed.), Adv Otorhinolaryngol., Basel, Karger, 2006; vol. 64, pp. 109-143.

[2] B. Townshend, N. Cotter, D. Compernolle, R. L. White, "Pitch perception by cochlear implantees," *J Acoust Soc Am.,* vol. 82, pp. 106-115, 1987.

[3] F. G. Zeng, "Temporal pitch in electric hearing," *Hear Res.,* vol. 174, pp. 101-106, 2002.

[4] J. A. Zakis, H. J. McDermott, A. E. Vandali, "A fundamental frequency estimator for the real-time processing of musical sounds for cochlear implants," *Speech Communication,* February 2007; vol. 49, no. 2, pp. 113-122.

[5] J. J. Dubnowski, R. W. Schafer, L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-24,no-1,February 1976.

[6] D. Talkin, "A Robust Algorithm for Pitch Tracking," in *Speech Coding and Synthesis,* W. B. Kleijn, K. K. Paliwal, Eds., Elsevier, November 1995, pp. 497-515.

[7] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: VOCODER revisited," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol. 2, 21-24 April 1997, pp. 1303–1306.

[8] L. R. Rabiner, R. W. Schafer. *Digital Processing of Speech Signals*, New Jersey: Prentice Hall, 1978, pp. 150-160.

[9] J. M. Hillenbrand, L. Getty, M. Clark, K. Wheeler, "Acoustic characteristics of American English vowels," *J Acoust Soc Am.,* vol. 97, pp. 3099-3111, 1995.

[10] H. Dudley, "Remaking speech," *J Acoust Soc Am.,* vol. 11, pp. 1969-1977, 1939.

[11] T. Green, A. Faulkner, S. Rosen, "Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants", *J Acoust Soc Am.,* vol. 116, pp. 2298-2310, 2004.

[12] A. Vandali, C. Sucher, D. Tsang, C. McKay, J. Chew, H. McDermott, "Pitch ranking ability of cochlear implant recipients: A comparison of sound-processing strategies", *J Acoust Soc Am.,* vol. 117, no. 5 ,pp. 3126-3138, 2005.

[13] R. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, "Speech recognition with primarily temporal cues", *Science.,* vol. 270, pp. 303-304, 1995.

[14] M. Dorman, P. Loizou, R. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs", *J Acoust Soc Am.,* vol. 102, pp. 2403-2411, 1997.

[15] A. Lobo, P. Loizou, N. Kehtarnavaz, M. Torlak, H.Lee, A. Sharma, P. Gilley, V. Peddigari and L. Ramanna, "A PDA-based research platform for cochlear implants", *3rd International IEEE/EMBS Neural Engineering, Conference*, 2-5 May 2007, pp. 28 – 31.