A DCT based approach to Estimation of Pitch

Suresh K
Dept. of E&C, College of Engg., Thiruvananthapuram, India 695016
sureshkj@vsnl.com

A. G. Ramakrishnan
Dept.of Electrical Engg, Indian Institute of Science, Bangalore, India 560012.
ramkiag@ee.iisc.ernet.in

Abstract— **A simple and elegant algorithm is proposed for pitch detection, using discrete cosine transform (DCT). In the first step, DCT peaks in the probable range of frequencies are picked up. We then look for the presence of harmonics for each of these peaks. The base band peak with the maximum number of harmonics corresponds to pitch frequency. The algorithm was tested extensively on both male and female speech. It was also tested successfully on signals with very low SNR. Comparisons with other pitch detection methods confirmed the validity of the algorithm.**

I. Introduction

Pitch period estimation or equivalently, fundamental frequency estimation is an important problem in speech processing. Algorithms commonly used for pitch extraction are simplified inverse filtering technique (*SIFT*) [1] and those based on cepstral analysis [2]. In *SIFT*, the pitch corresponds to the interval between the peaks in the autocorrelation of the inverse filter output of the low pass filtered signal. In cepstral method, the pitch corresponds to the cepstral peak in the probable range of pitch frequencies. In [3], a technique for detecting the instance of glottal closure in voiced speech is proposed, which also can be viewed as a pitch detection technique. A spectral autocorrelation based method is proposed in [4] where the frequency shift between successive peaks of the autocorrelation of the spectrum corresponds to the fundamental frequency. A modification of the method proposed in [4] is cited in [5]. Here preprocessing is done to flatten the spectrum by obtaining the residual from linear prediction analysis. The preprocessed signal is used for getting the spectral autocorrelation function (SAF). In this paper, a new pitch detection algorithm using discrete cosine transform (DCT) is proposed. Though the discussion here is limited to its application for speech synthesis, the proposed method can be used for other applications as well.

II. Methods

The basis of the algorithm is the observation that the pitch can be estimated from the index of the DCT peak within the probable range of pitch frequency, which has significant harmonics. The pitch frequency is assumed have a value between 50 Hz and 400 Hz. Speech signals sampled at 16 kHz were analyzed with a frame length of 60 msec. A high resolution DCT of the frame was taken. All the DCT coefficients corresponding to frequencies below 1kHz were considered for processing. The algorithm has two main steps. In the first step, the probable coefficients/cluster of coefficients, which can be declared as the pitch coefficients are determined. In the second step, the true candidate among the coefficients selected is determined. The coefficients with absolute value below a threshold are assigned a value of zero and the rest, a value of one. The threshold is set to one fourth of the maximum magnitude of the DCT coefficients in that frame. The search space for the first step is between 50 and 400 Hz and clusters of ones in that range are determined. Once a cluster is detected, the search is restricted to 40 Hz above that. The lowest and highest indices of each cluster are stored. Number of such clusters is also noted. In the second step, the presence of harmonics is checked for each cluster detected in the first step. The search space for the harmonics is up to 1 kHz. Let hmax be the maximum number of harmonics that can exist within 1 kHz for a particular cluster. A cluster is declared as the pitch cluster, if the number of detected harmonics is greater than or equal to (hmax-2). If more than one cluster passes the test, the cluster with maximum harmonics is declared as the pitch cluster. If none of them passes the test, a recursive harmonic detection process is initiated, in which the threshold is successively reduced to three fourth of the previous value and the harmonics are detected as before. The minimum value of the threshold is restricted to ten percent of the frame maximum value. If it still fails to pass the test, there are two possibilities. The first is that the corresponding frame is unvoiced. The other possibility is that the frame is dominated by the pitch (only the first harmonic), and hence, the number of clusters detected in the first step must be unity. If this is true, the corresponding value is decided as the pitch. Otherwise, the frame is declared as unvoiced.

III. Algorithm

1. $x[n]$ is the window of the speech signal, appended with sufficient zeros in order to make the DCT resolution equal 1 Hz; $X[k]$ are the corresponding DCT coefficients.
2. $j = 50$, $i = 0$.
3. Make $X[k] = abs(X[k])$ and get $X_M = max(X[k])$.
4. Set threshold, $X_T = 0.25 * X_M$.
5. $\forall k$, if $X[k] \geq X_T$, set $X[k] = 1$; else, set $X[k] = 0$.
6. While $X[j] = 0$, j^{++}. i^{++}; $LO[i] = j$; While $X[j] = 1$, j^{++}; $HI[i] = j - 1$;
7. If $j \geq LO(1) + 40$, go to step 8. Else, go to step 6.
8. $N_C = i$, is the number of clusters of significant coefficients, any one of which could be the pitch.
9. For $I = 1$ to N_C, $maxharm(I) = 1000/HI(I)$.
10. Define C_{ij} = cardinality of the set of nonzero DCT coefficients from $j * LO(i)$ to $j * HI(i)$.
11. For $i = 1$ to N_C, $harm(i) = 0$. For $i = 1$ to N_C, for $j = 1$ to $maxharm(i)$, if $C_{ij} \neq 0, harm(i) += 1$.
12. If $harm(m) \geq harm(j)$, for all $j \neq m$, then the pitch corresponds to the m^{th} cluster.
13. Let n be the highest harmonic of m^{th} cluster for which $C_{mn} \neq 0$. Let P_{mn} be the index corresponding to the maximum absolute value of the DCT coefficients between $n*LO(m)$ and $n * HI(m)$.
14. The pitch $= \frac{P_{mn}}{n}$.

IV. Results and Discussion

The speech data used for analysis consists of the basic units of a concatenation-based speech synthesis system. Thirty each of such units were recorded from two male and four female speakers and were used for testing the algorithm. On verification against the cepstral as well as the SIFT algorithms, the validity of the new technique was proved. Performance in the presence of excessive noise was tested with a segment having 0 dB SNR and the algorithm was able to detect the pitch accurately. Also, the algorithm was applied on words extracted from continuous speech and the effectiveness was established.

Figure 1(a) shows a segment of speech waveform for pitch analysis. Figure 1(b) shows the DCT of the same (first 1000 coefficients). Figure 1(c) s hows the thresholded, absolute value of DCT. It is clear that the thresholded cosine spectrum is dominated by the pitch and its harmonics. Figure 1(d) shows an example of a case, where the thresholded spectrum contains nothing other than the fundamental and second harmonic of the pitch. In Fig. 2(a), the phoneme '/e/' spoken by a female speaker is given and in Fig. 2(b), its pitch contour obtained by our algorithm is shown. A frame size of 60 ms with 10 ms overlap was used to get the contour. Fig. 3 (a) shows the phoneme in Fig. 2(a) corrupted with noise (0 dB SNR). Its pitch contour is shown in Fig. 3(b). Thus, we see that the algorithm works well even on very low SNR signals. In Fig. 4, the pitch contour of three different female speakers for the phoneme '/a/' at 5 dB SNR is shown.

The percentage error in the pitch estimation is shown in Table I. Five voiced phonemes spoken by a female speaker were taken and for each case, the pitch was estimated manually and with the algorithm. The error given in the table is the average estimation error with respect to the manually calculated pitch. There is very little change in the estimation error with increasing

TABLE I
PERFORMANCE OF THE ALGORITHM AT DIFFERENT NOISE LEVELS

SNR(dB)	0	5	10	15	20
% Error	5.76	5.33	5.33	5.32	5.12

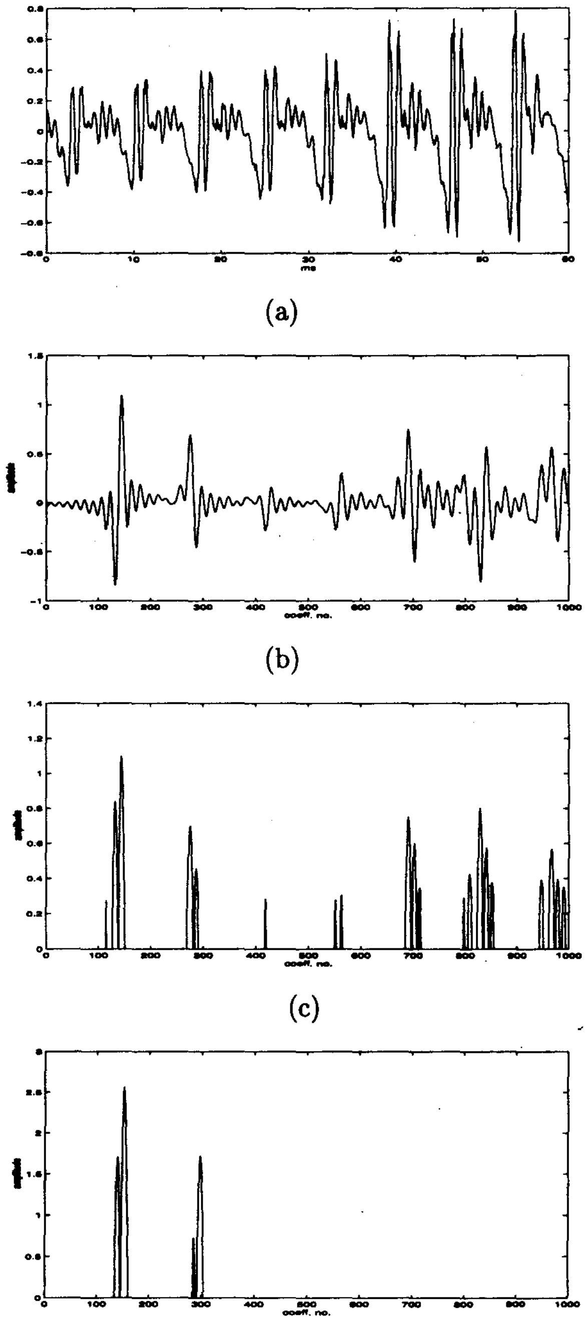
noise level. This proves the noise robustness of the algorithm.

The methods discussed in [4] and [5] are based on the autocorrelation of windowed spectrum. They essentially rely on the presence of harmonic components of the fundamental frequency of the signal and the pitch period is obtained from the harmonic spacing in the spectrum. Our algorithm looks for the harmonic structure of the signal to determine the true candidate for the pitch period. It is observed that a voiced speech segment has pitch and its harmonics as its main components. The algorithm directly searches for the harmonics in the DCT domain.

V. Conclusion

A simple and effective pitch detection algorithm has been proposed and tested extensively. It holds promise as a viable and useful technique for analyzing the speech data used for a concatenation-based speech synthesis system. The effectiveness of the technique shows that there are both linear and non-linear processes in operation during the generation of speech. Presence of a strong fundamental component confirms the existence of a linear process and the presence of a good number of harmonics with significant energy, proves the existence of a co-occurring nonlinear process in the vocal cavity.

The merit of the technique is that its computational complexity is low. It is possible to use the significant coefficients of the DCT for modelling the basic units in a concatenation based speech synthesis system, rather than the LPC coefficients or the formants. In such DCT model based synthesizers[6], since the DCT of the units is available to begin with, the additional computation required for pitch detection is minimal.



(a)

(b)

(c)

(d)

Fig. 1. Pitch detection using DCT (a) A speech segment (b) DCT of the speech segment (c) Thresholded absolute value of the DCT (d) Thresholded DCT of a pitch-dominated speech segment

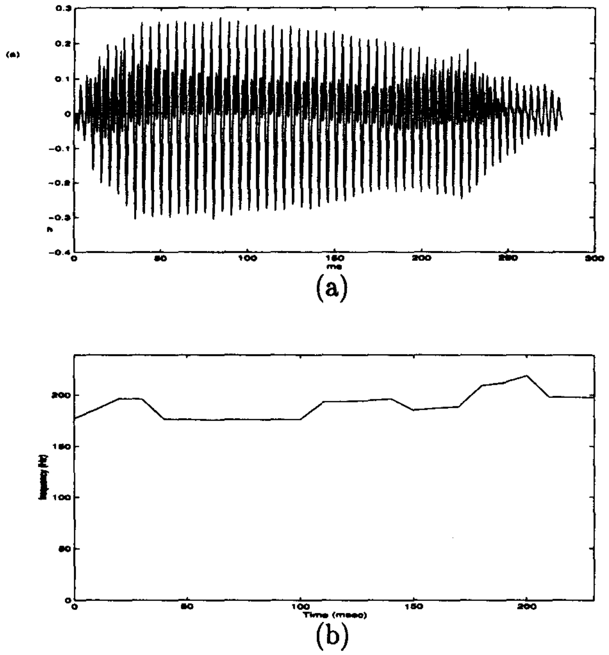


Fig. 2. (a) Speech segment '/e/'(female voice) (b) Pitch contour of '/e/'

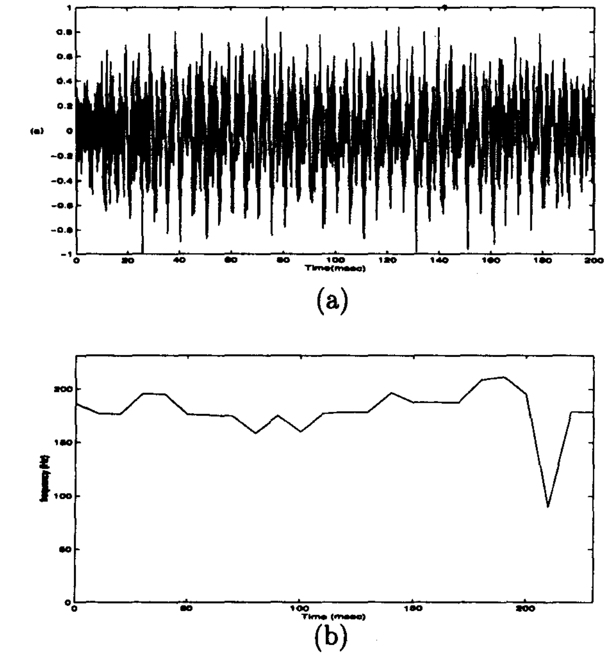


Fig. 3. Performance of our pitch detection algorithm on noisy speech. (a) Speech segment '/e/' (female voice) having 0 dB SNR. (b) Pitch contour obtained by our technique.

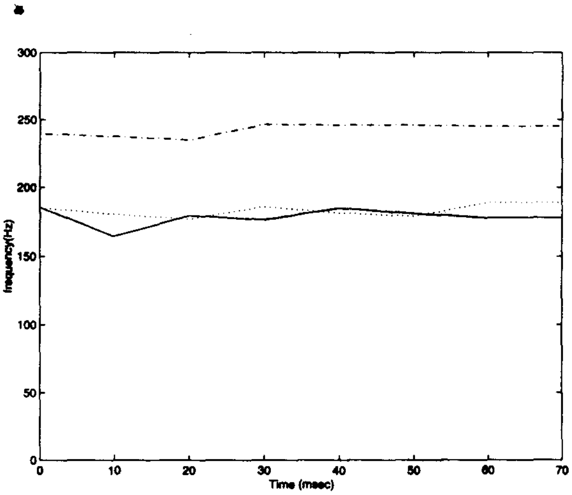


Fig. 4. Pitch contour of the phoneme '/a/' at 5 dB SNR spoken by three different female speakers.

References

[1] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio, Electroacoust., Vol. 20, No. 5, pp. 367-377, 1972.*

[2] L.R.Rabiner,R.W.Schafer,"Digital Processing of Speech Signals," *Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978, pp 372-378.*

[3] B.Yegnanarayana and R.L.H.M.Smits, "A Robust method for determining instants of major excitations in vocal speech", *Proc. of ICASSP-95, Detroit, USA, May 8-12, 1995.*

[4] M.Lahat et. al., "A spectral autocorrelation method for the measurement of the fundamental frequency of noise-corrupted speech," *IEEE Trans. ASSP. vol 35, No. 6, pp 741-750, June 1987.*

[5] E. Chilton and B.G.Evans, "The spectral autocorrelation applied to the linear prediction residual of speech for robust pitch detection," *Proc. of ICASSP '88, pp 385-361, 1988.*

[6] Suresh K. "Speech synthesizer for Malayalam," *M.E. Thesis, ECE Dept., IISc., Bangalore, India 560012, Jan. 2000.*