Robust Pitch Detection using DCT based Spectral Autocorrelation

R. Murali shankar and A. G. Ramakrishnan

Dept. of Electrical Engg, Indian Institute of Science, Bangalore, India 560012

Abstract - **In this paper, we propose a robust algorithm for the determination of pitch in voiced speech. Lahat et al. [1] proposed a method based on spectral autocorrelation function (SAF). The SAF measures the regular harmonic spacing in the Discrete Cosine Transform (DCT) spectrum of the speech signal. Chilton and Evans [2] modified this technique by working on the linear prediction residuals. Kunieda et al [3] worked on the log spectrum. Here, we have analyzed DCT spectrum rather than the Discrete Fourier Transform (DFT) spectrum, as used by the above techniques. The pitch is detected using the DCT SAF. The algorithm is applied on noisy speech of SNR down to 0 dB to validate its robustness. The algorithm performs better than the one based on DFT in the presence of noise.**

I. INTRODUCTION

Measurement of F_0, the fundamental frequency of voiced speech, is essential for studies involving phonetics, linguistics, analysis of pathological voice, synthesis of speech and analysis-synthesis telephony. Many methods have been proposed to measure F_0. With a periodic voiced waveform, it is easy to determine the period of the waveform by observation. On the other hand, when the voiced signal is transient or noisy and the variation of the pitch values is large, accurate measurement and tracking of the pitch period have been very difficult to achieve.

The time-domain autocorrelation function (TDAF) is well known for its ability to extract the periodicity from noisy signals and it has been extensively used for pitch determination in speech because of its robustness and reliability [4]. Before computing auto-correlation, the speech signal must be suitably preprocessed in order to flatten the spectrum. The residuals from LP analysis have been used for this purpose.

Frequency domain pitch detection methods measure the harmonic spacing in the spectrum and have the advantage that the fundamental can be missing, as may occur over telephone lines for male speech [5]. Unfortunately, the harmonic structure in the spectrum tends to be poor at higher frequencies and often, only lower part of the spectrum is included [6].

The SAF was first proposed in [7]. It describes appropriate postprocessing needed to ensure accurate performance over a wide range of pitch values. We can eliminate the need for postprocessing and make the technique more robust, if we use DCT SAF. The technique can be extended to enable voiced/unvoiced decision.

The Spectral autocorrelation function

The TDAF and power spectral density function (PSD) forms a Fourier transform pair and it is to be expected that the robust properties of the autocorrelation would be preserved in the frequency domain. If P (k) is the PSD of a signal, then the SAF is defined by:

$$R(k) = 1/N \sum_{i=0}^{N-k-1} P(i)P(k+i)$$

for k = 0, 1, ... , M, where k represents the shift in the sampled frequency domain. Since P(k) is real for the DCT spectrum, the autocorrelation is performed over N samples of an N-th order DCT. For a signal containing significant harmonics, this function will produce strong correlation peaks at the average harmonic spacing in the spectrum and multiples of this value, in the shifted frequency domain. As with the TDAF, this function requires some preprocessing for it to operate successfully. The power spectrum should ideally be flat with all harmonics approximately the same height.

Inverse filtering the speech using the LP coefficients is a satisfactory method of achieving this.

II. METHODS

The residual of a voiced speech differs from the idealized model (pulsy nature) in terms of its morphology. It is bipolar and pulse heights and the pitch period may vary considerably within any analysis block. We take the sequence of maximum permitted length (say 320 samples) and pad it with zero after windowing it with hamming window. No improvement in resolution can be obtained this way but the effect is to interpolate the frequency domain sequence. The heights of the spectral peaks overall are reduced with a broadening and smoothing of their shapes. A similar effect occurs for the spectral noise in the sequence. It is observed that the shape of the spectral peaks is now determined by the transform and interpolation operation. The width of each peak towards the base is determined by the limit of the resolution. Further, for large pitch periods, i.e. when the spectral peaks are close, the base of one peak merges into the next. The only evidence of noise is in the random variations in the heights of the pitch peaks and an occasional "skewness" where a significant noise peak and a pitch peak are closely adjacent.

High quality speech was recorded at 16 kHz sampling rate. Preprocessing of the speech signal is done to flatten the spectrum before applying autocorrelation. A frame size of 20 ms with 10 ms overlap is used to get the pitch contour. The DCT magnitude spectrum is obtained for each analysis frame (N=1024). It is observed that the harmonic structure at higher frequencies is weaker than that at the lower frequency end of the spectrum. Hence, the DCT magnitude spectrum is smoothed by the following window:

$W(k) = 1, \quad \text{for } 0 < k < N/2$

$W(k) = 0.5*(1-\cos(2*\pi*k/N)) \quad \text{for } N/2 < k < N$

SAF is applied on this smoothed DCT spectrum. Since pitch must lie below 1 kHz, only the first M=128 lag values of the function are calculated. The pitch detection is simply a matter of picking an appropriate peak, irrespective of whether the SNR is high or low. Whereas, in the case of DFT based SAF, Chilton et al [2] report that postprocessing of the SAF is necessary to obtain acceptable results for low SNR speech.

III. RESULTS AND DISCUSSION

The algorithm was tested on a number of speech segments obtained from both male and female speakers and also on male voice received over telephone lines. The validity of DCT-SAF was established by verification using the methods based on autocorrelation and cepstrum.

Fig.1 gives the results obtained with a phoneme /ae/, (300 ms long) from a female speaker. Figs. 1(a) and 1(b) show the clean speech segment and its DCT-SAF, respectively. Figs. 1(c) and 1(d) show its pitch contour estimated by DFT-SAF and DCT-SAF, respectively. Figs. 1(e) and 1(f) show the phoneme /a/ from a different female speaker and its pitch contour using our algorithm. Gaussian noise was added to this speech file to obtain different SNRs and SAF was obtained using the LP residuals derived from these noisy files.

Figs. 1(g) and 1(h) show the pitch contour obtained from DFT-SAF and DCT-SAF for this noisy file (SNR = 0 dB). Fig. 1(i) shows the DCT-SAF obtained for different SNRs. Fig. 2 gives the results obtained with a phoneme /a/, from a male speaker over the telephone lines. Figs. 2(a) and 2(b) show the noisy speech segment and its pitch contour by our algorithm. This signal does not have the fundamental frequency because the lower cutoff frequency of telephone bandwidth is above the pitch frequency range for male voices.

Voiced/unvoiced decision can also be made with the normalized SAF where the heights of the peaks indicate the degree of pitch present irrespective of the signal power. In the case of

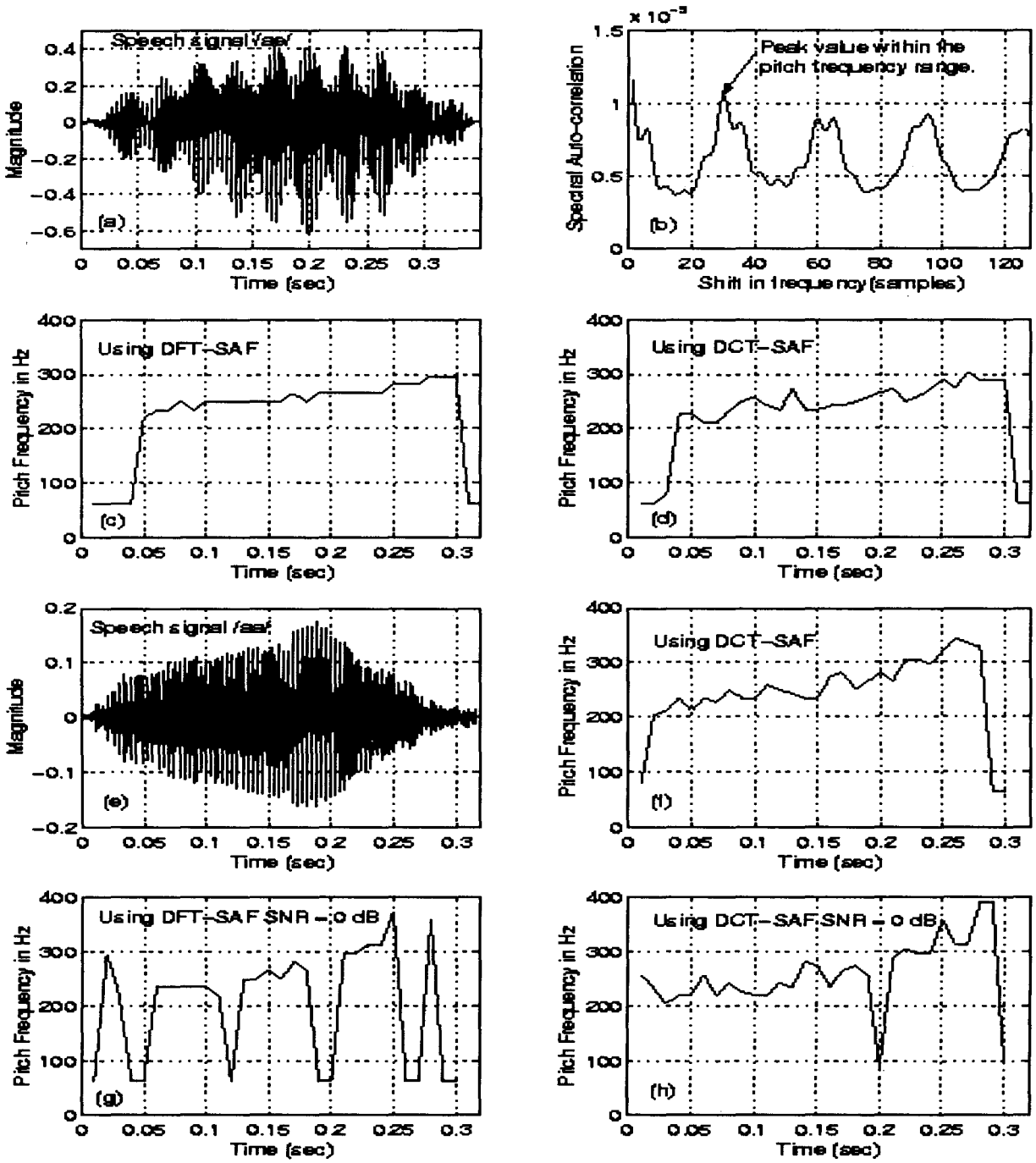

Fig 1(a) Clean speech /ae/. (b) DCT based Spectral Autocorrelation function (SAF) for clean speech. (c) Pitch contour using DFT based SAF. (d) Pitch contour using DCT based SAF (e) Clean speech /aa/, with emotion. (f) Pitch contour using DCT based SAF. (g) Pitch contour using DFT SAF for SNR = 0 dB. (h) Pitch contour using DCT SAF for SNR= 0 dB. (i) Pitch contour of noisy speech obtained using DCT SAF for different SNRs

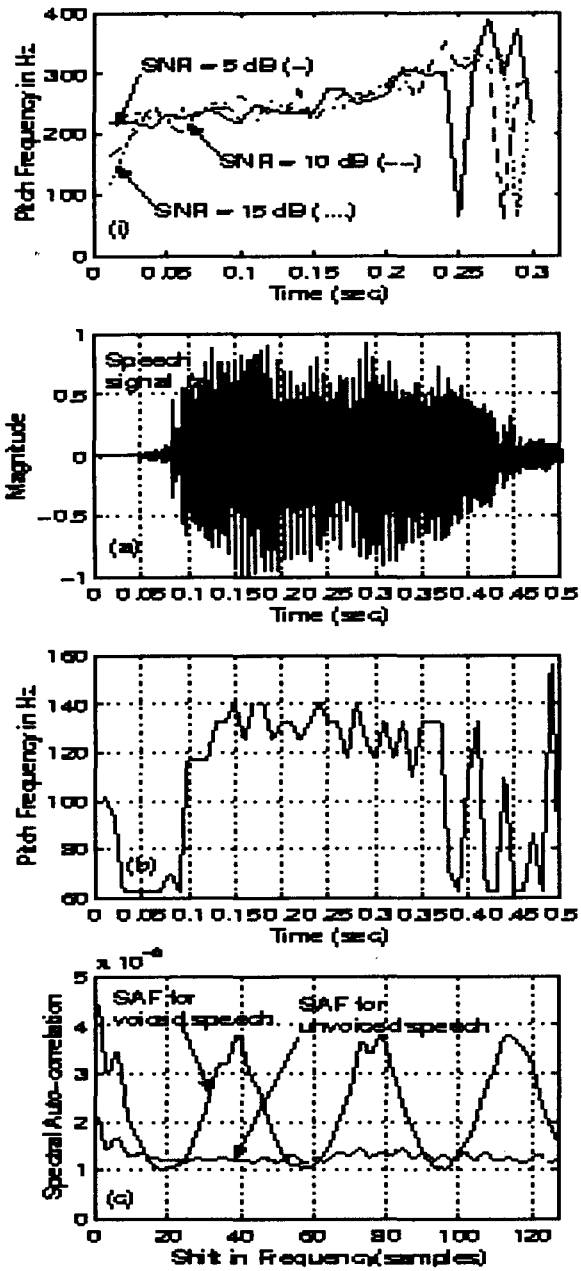

Fig. 2(a) Noisy speech from telephone line.
2(b) Pitch contour using DCT based SAF.
2(c) Spectral autocorrelation function for voiced and unvoiced speech.

low SNR, the maximum deviation of the second order differential of the normalized SAF provides a consistent measure of the ratio of periodic to random energy in speech, over a wide range of pitch values. Fig. 2(c) shows the SAF for voiced as well as unvoiced speech.

IV. CONCLUSIONS

Results indicate that the proposed method has the potential to be a robust tool for obtaining a good estimate of the fundamental frequency, even if it is high. It can be seen that DCT- SAF performs well down to a long term SNR of 0 dB.

V. REFERENCES

[1] M. Lahat et al., "A Spectral autocorre-lation method for the measurement of the fundamental frequency of noise-corrupted speech," IEEE Trans. ASSP. Vol. 35, No. 6, June 1987.

[2] E. Chilton and B. G. Evans, "The spectral autocorrelation applied to the linear prediction residual of speech for robust pitch detection", Proc. *ICASSP '88,* 1988, pp. 358-361.

[3] N. Kunieda, T. Shimamura, and J. Suzuki, "Robust method of measurement of fundamental frequency by ACLOS- Autocorrelation of Log Spectrum", Proc. *ICASSP '96,* 1996, pp. 232-235.

[4] L. R. Rabiner. "On the use of autocorrelation analysis for pitch detection," IEEE Trans. ASSP. Vol. 25, No. 1, Feb 1997.

[5] I. Boyd and R. Linggard, "Voice pitch extraction using a real-time digital filter-bank," Int. J. Electronics, Vol 57, No 5, pp 633-640: 1984.

[6] S. Seneff, "Real-time harmonic pitch detector," IEEE Trans. ASSP, Vol 26, No 4: Aug. 1978.

[7] E. Chilton and B. G. Evans, "Performance comparison of five pitch determination algorithms onthe linear prediction residual of speech", European Conf. on Speech Technology, Edinburgh: Sept. 1987.