DCT BASED PITCH MODIFICATION

R. Murali Shankar[*], M. Anoop[#], T. Harish[*], A. K. Rohit Prasad[**] and A.G. Ramakrishnan[*]
[*]Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India.
Department of Electrical Engineering, Arizona State University, USA
[**]Infosys Technologies Ltd., Bangalore 561 229, India
Email: {sripad, ramkiag}@ee.iisc.ernet.in

Abstract
In this paper, we propose a novel algorithm for pitch modification. The linear prediction residual is obtained from pitch synchronous frames by inverse filtering the speech signal. Then Discrete Cosine Transform (DCT) is applied on these pitch synchronous frames. Based on the desired factor of pitch modification, the dimension of the DCT vector is changed by truncation or zero padding, and then Inverse Discrete Cosine Transform is applied. This period modified residual signal is then forward filtered to obtain the pitch modified speech. The quality of the speech thus obtained is better for increased pitch frequency than for decreased pitch by the same factor. The method can be applied pitch asynchronously also, with some distortion.

1 Introduction

The object of pitch-scale modifications is to alter the fundamental frequency of a speech segment without affecting its spectral envelope (more precisely, the location and bandwidths of the formants). Pitch-Scale modification of speech is a subject of major theoretical and practical interest. Applications are numerous such as text-to-speech synthesis (TTS) based on acoustic unit concatenation and transformation of voice characteristics.

In TTS, the phones must have their duration and pitch modified in order to fulfill the prosodic constraints of the word containing those phones. This processing is necessary to avoid the production of monotonous sounding synthesized speech. In a TTS system, this function is performed by a prosodic module. To allow the duration and pitch modifications in the recorded sub-units, many concatenation-based TTS systems employ the time domain pitch synchronous overlap add (TD-PSOLA) [1] model of synthesis. In the TD-PSOLA model, the speech signal is first submitted to a pitch marking algorithm. This algorithm assigns marks at the peaks of the signal in the voiced segments and assigns marks 10 ms apart in the unvoiced segments. The synthesis is made by a superposition of Hamming windowed segments centered at pitch marks and extending from the previous pitch mark to the next one. The duration modification is provided by deleting or replicating some of the windowed segments. The pitch period modification, on the other hand, is provided by increasing or decreasing the superposition between windowed segments. This algorithm is known to suffer from spectral and phase distortions. These are partly due to the time-domain nature of processing, in that the spectral envelope cannot be adequately controlled. Another computational disadvantage is that an accurate pitch marking is required. More complex methods, such as sinusoidal-model [3] based approaches, are gaining in popularity, but tend to be computationally intensive, especially at the synthesis stage. Our aim here is to describe an elegant algorithm that provides more flexibility, computational efficiency and low distortion.

2 Method

As an alternative to strict time domain techniques, the ubiquitous source-filter model of speech can be invoked [4]. Prosody modification then becomes a task of separating the excitation and vocal tract components from speech, modifying the excitation, and then recombining with the vocal tract component. In principle, this allows retaining the vocal tract response without any modifications.

In an ideal world, the analysis would separate the excitation signal, which could be modified independently of the tract response. However, in reality, there is dependence between the excitation signal and tract response. As a compromise, the system attempts to separate the speech signal into a

spectral shaping component, and an excitation (or residual) signal, ensuring in the process that the original signal's temporal detail is preserved.

According to the generally accepted engineering model for speech production, the sampled speech waveform is modeled as the output of a time varying filter driven by an excitation signal, which is either a sum of narrow-band signals with harmonically related instantaneous frequencies (voiced speech) or a stationary random sequence (unvoiced speech). The linear predictive coding (LPC) [5] residual (error, excitation) signal has a number of advantages over the speech signal in the context of modification [6]. It is spectrally flat and there is little correlation within each pitch period.

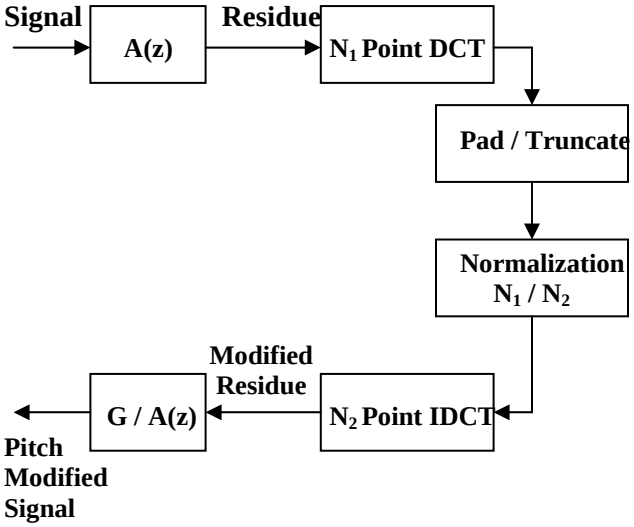

Fig 1. Block Diagram of entire process.

(a) Pitch marking

The first step in our analysis is to pitch mark the speech signal. For this task, an algorithm based on the autocorrelation of the negative part of the speech signal has been carried out. In the autocorrelation domain, finding local maxima and the distance between successive local maxima gives the periodicity of the signal under consideration. After getting the pitch information, it is submitted to various periodicity constraint rules, and linked together in order to obtain a chain of marks. A nonlinear processing of these marks that includes deletion, delay, interpolation and extrapolation, results in the final pitch marks positioned at the peaks of the signal in the voiced segments. Figure 2 shows a segment of a pitch marked signal. In the unvoiced segments, unvoiced marks are positioned 10 msec apart. The voiced and unvoiced marks are both called as pitch marks. Because the marks are positioned at specific samples of the speech signal, the resulting period is quantized to an integer number of samples. This is a common procedure in pitch synchronous TTS system and is employed in our algorithm. For a 10 msec pitch period and a 16 kHz-sampling rate, for example, the error in the pitch period due to quantization is lower than 0.63%.

(b) Modification of Residual Signal.

Figure 1 shows the details of the proposed method. The process of DCT based pitch modification involves pitch synchronous extraction of the residual signal. The length of the residual signal of each frame is modified using DCT-IDCT. Each frame of the speech signal is synthesized by forward filtering.

N_1-point DCT of each frame of the excitation signal is obtained. N_1 corresponds to the actual number of samples in each extracted frame. An N_2-point IDCT is then obtained, where N_2 corresponds to N_1 multiplied by the ratio of modified to original pitches.

The effect of taking an N_1-point DCT followed by an N_2-point IDCT has an effect almost amounting to resampling the excitation signal. This occurs because, while taking the IDCT, the new length is mapped to π.

3 Results

To demonstrate the effectiveness of this technique, individual phonemes, words and sentences spoken by a male were analyzed and re-synthesized by different factors.

Figure 3(a) shows a segment of a phoneme. Figure 3(b) gives the corresponding segment of the residual signal extracted by inverse filtering the phoneme (LP model order 18). Figure 3(c) shows the modified residual signal obtained making use of DCT – IDCT, the factor of increase in the pitch being 1.3. Figure 3(d) shows a segment of the resynthesized speech signal after forward filtering.

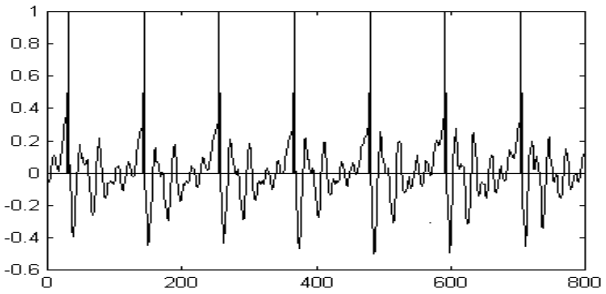

Fig 2. Pitch synchronous frames.

Figure 4 shows the pitch contours for the phoneme shown in Fig. 3(a), and its pitch modified version, shown in Fig. 3(d). It can be seen that the pitch contour is maintained in the modified signal. In order to obtain the same result by TD-PSOLA, not only pitch marking is required, but also the overlap lengths will change from one pitch period to the next.

Figure 5(a) shows the speech signal for a whole word. Its original pitch contour and the contour after pitch modification by our technique are shown in Fig. 5(b).

To evaluate the overall performance, perceptive evaluation tests were carried out. The pitch contours of many phonemes, words and sentences were modified by different factors. Eleven people were asked to rate the audio quality and the change of pitch after the modification. The results of the evaluation tests are given in Tables 1 and 2, for pitch synchronous, and pitch asynchronous modifications, respectively. From these tables, we can see that the pitch synchronous scheme in general results in speech that is perceptively better than that obtained by pitch asynchronous scheme. Further, the quality of the speech is better for increased pitch frequency than for decreased pitch by the same factor.

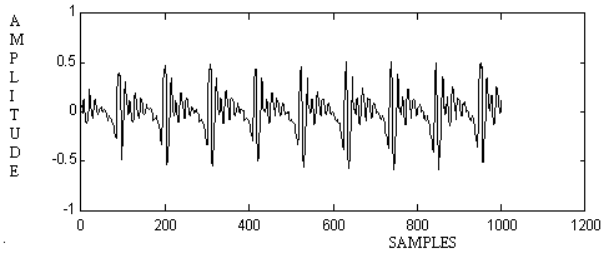

Fig 3(a). Few frames of the original signal 'a'.

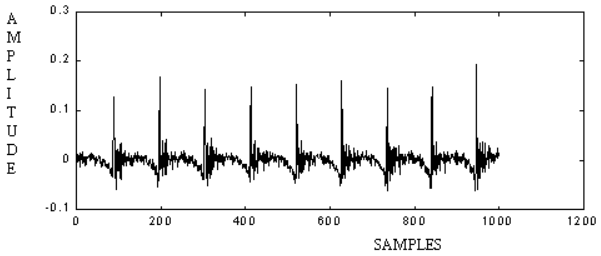

Fig 3(b). Few frames of the original excitation.

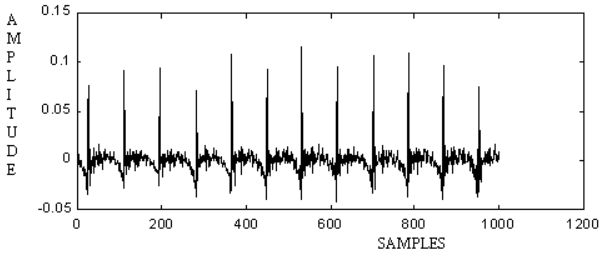

Fig 3(c). Few frames of the modified excitation.

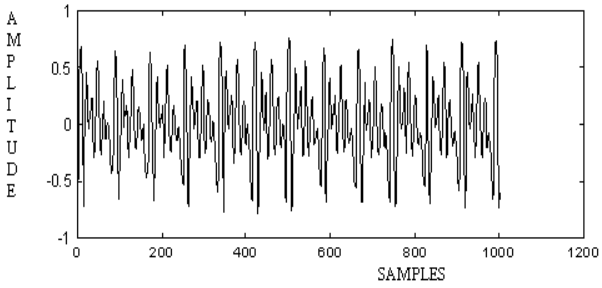

Fig 3(d). Few frames of the reconstructed signal.

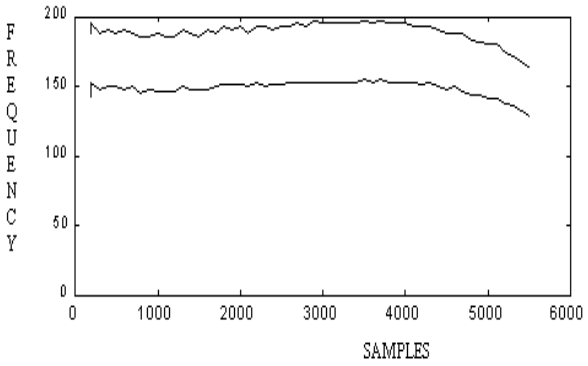

Fig 4. Pitch contours of the original and modified phoneme

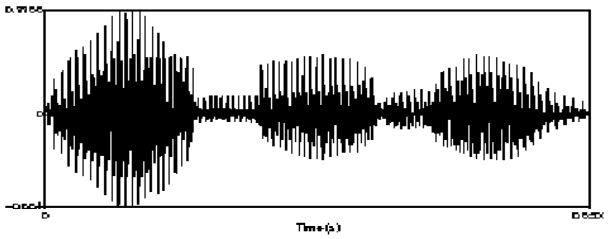

Fig 5(a). Speech signal of word "aamele".

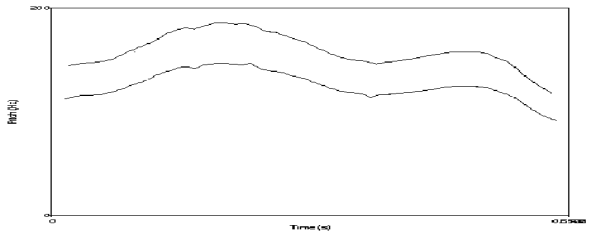

Fig 5(b). Pitch contours of the original and the modified word "aamele".

Also, we used pitch modification to convert an emotional utterance to a non-emotional one, and vice versa. Thus, a sentence spoken in surprise was converted to a normal one by reducing the pitch by a suitable factor. Since the sentence spoken in surprise is naturally of shorter duration than a normal one, there was no need for duration modification. Similarly, normal utterances have been modified to generate emotions such as surprise and anger.

4 Conclusion

This algorithm is simple and elegant. It directly follows from the basic model of speech. Perceptive evaluation shows that this performs well for the range of pitch change factors sufficient for a TTS system. The algorithm uses DCT – IDCT, and thus, is not computationally intensive. Further, it has the distinct advantage that this algorithm can be used for pitch asynchronous pitch modification, unlike the TD-PSOLA, which can only be applied pitch synchronously. Even though pitch asynchronous pitch modification requires synchronization of speech signals, the latter is an offline approach with no computational burden on the synthesis [7]. Further, the proposed scheme maintains the relative pitch contour of the original signal, without any additional processing or precautions to be taken care of.

Table 1. Perceptive Evaluation of pitch synchronous pitch modified utterances.

Pitch change Factor	Good	Fair	Bad
0.6	74%	17%	9%
0.8	91%	9%	
1.2	100%		
1.3	100%		
1.4	91%	9%	
1.6	84%	16%	
1.9	80%	10%	10%

Table 2. Perceptive Evaluation of pitch asynchronous pitch modified utterances.

Pitch change Factor	Good	Fair	Bad
0.6	74%	17%	9%
0.8	83%	7%	10%
1.2	100%		
1.3	82%	9%	9%
1.4	82%	9%	9%
1.6	73%	27%	
1.9	64%	18%	18%

References

1. Moulines, E., and Charpentier, F., "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones", *Speech Comm.*, Vol. 9, pp. 453-467, Dec. 1990.
2. N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform", IEEE Trans. Comput., vol. C-23, no. 1, pp. 90-93, Jan. 1974.
3. McAulay, R.J., and Quatieri, T.F., "Shape invariant time-scale and pitch modification of speech", *IEEE Trans. Signal Processing*, Vol. 40, pp. 497-510, March 1992.
4. Rabiner L.R., and Schafer R.W., "Digital processing of speech signals". *Prentice-Hall, Inc* Englewood Cliffs, New Jersy 07632
5. Makhoul, J., "Linear prediction: a review", *Proc. IEEE*, Vol. 63, pp. 561-580, April 1975.
6. M. Edgington and A. Lowry. "Residual-based speech modification algorithms for TTS Synthesis". *BT laboratories*, Martlesham Health IPSWICH, IP5 7RE, U.K.
7. Yannis Stylianou "Removing phase mismatches in concatenative speech synthesis" AT&T Laboratories – Research.