Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Speech Communication 42 (2004) 143–154

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Modification of pitch using DCT in the source domain

R. Muralishankar *, A.G. Ramakrishnan, P. Prathibha

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

Received 1 September 2001; received in revised form 1 November 2002; accepted 1 May 2003

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 the discrete cosine transform (DCT) of these residual frames is taken. Based on the desired factor of pitch modification, the dimension of the DCT coefficients of the residual is modified by truncating or zero padding, and then the inverse discrete cosine transform is obtained. This period modified residual signal is then forward filtered to obtain the pitch modified speech. The mismatch between the positions of the harmonics of the pitch modified signal and the LP spectrum of the original signal introduce gain variations, which is more pronounced in the case of female speech [Proc. Int. Conf. on Acoust. Speech and Signal Process. (1997) 1623]. This is minimised by modifying the radii of the poles of the filter to broaden the otherwise peaky linear predictive spectrum. The modified LP coefficients are used for both inverse and forward filtering. This pitch modification scheme is used in our Concatenative Speech synthesis system for Kannada. The technique has also been successfully applied to creating interrogative sentences from affirmative sentences. The modified speech has been evaluated in terms of intelligibility, distortion and speaker identity. Results indicate that our scheme results in acceptable speech in terms of all these parameters for pitch change factors required for our speech synthesis work.

Keywords: Linear prediction; Concatenative synthesis; Residual signal; Resampling; 3 dB bandwidth; Spectral broadening

1. Introduction

Machine synthesis of speech (Roe and Wilpon, 1994; Syrdal et al., 1995) facilitates convenient information transmission in a number of applications, including voice delivery of text messages and email, voice response to database enquires, reading aids for the blind and mobile communications. Speech synthesis presents a key challenge when it comes to improved quality (Liberman, 1994), which is assessed by the attributes of intelligibility and naturalness. Of the various approaches to speech synthesis, concatenative synthesis has entailed speech with the highest quality to date. Concatenative synthesis involves selecting a class of basic acoustic units, creating an inventory of stored units by recording them from natural voice, and then generating utterances by concatenating appropriately modified segments from this inventory. A critical task in concatenative speech synthesis is that of modifying the prosody (pitch, amplitude and durations) of the voiced sections of the stored units and creating a concatenation of units that sounds seamless. Methods have been proposed in the literature for both time and pitch

* Corresponding author. Tel.: +91-80293-2935; fax: +91-80360-0444.
E-mail address: sripad@ee.iisc.ernet.in (R. Muralishankar).

0167-6393/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2003.05.001

scale speech modification (George and Smith, 1992; McAulay and Quatieri, 1986; Portnoff, 1981; Quatieri and McAulay, 1986; Vergin et al., 1997). Pitch scale modification or pitch modification has applications such as adjusting the pitch in a singer's voice to get the desired effect, helping hearing impaired to understand speech better and modifying speech so that it is easier to code efficiently (Anssi, 1999). The objective of pitch modification is to alter the fundamental frequency of speech without affecting the time-varying spectral envelope. Techniques exist in the literature that accomplish this in the time or frequency domain.

1.1. Time domain pitch modification

Time domain pitch synchronous overlap adding (PSOLA (Moulines and Charpentier, 1990)) is likely the simplest method that can be imagined for good quality pitch modification of speech signals. In practice, the implementation of pitch modification in time domain (TD-PSOLA) requires knowledge about the pitch pulse locations. Exact pitch pulse locations are not essential, but it is crucial to maintain an exact pitch synchronicity between successive pitch marks. The signal is windowed pitch synchronously using a Hamming window of length 2–4 pitch periods, centered around the current pitch pulse. A length of 2 periods is usually good for pure time-domain modification and a longer window (>2) is good for frequency domain PSOLA (FD-PSOLA). Because the intervals between the pitch pulses are altered, the total length of the signal is modified and thus time scale modification of speech is also usually needed in order to maintain the original length of the signal. It is implemented in a simple way: If the pitch is increased, some frames are used twice and if it is lowered, some frames from the original signal are left out in the synthesized signal.

1.2. Frequency domain pitch modification

Historically, the FD-PSOLA was the first pitch synchronous time scale and pitch scale modification technique proposed in the literature (Charpentier and Stella, 1986). FD-PSOLA and residual domain PSOLA (LP-PSOLA) are two methods that can be adapted almost directly from the TD-PSOLA paradigm. These two methods are more flexible than the TD-PSOLA technique because they provide a direct control over the spectral envelope at both the analysis and the synthesis stages. In FD-PSOLA, prior to overlap add synthesis, each short-time analysis signal is modified; the modification is carried out in the frequency domain on the short-time Fourier transform signal. The algorithm used is basically a frequency domain resampling, which leads to some complex problems in the synthesis stage. It can be said that, if features such as speaker identity hiding are not needed, TD-PSOLA leads to the same results with a much simpler implementation. In practice, FD-PSOLA differs from TD-PSOLA only in the definition of the short-time synthesis signals for pitch scale modifications.

In LP-PSOLA, prior to PSOLA processing, the signal is split into an excitation component e(n) and the spectral envelope A(z). Pitch scale modification is then carried out on the source (residual) signal. The output is obtained by combining the modified source signal with the time-varying spectral envelope usually using linear prediction. Synthesis is again complex and the details can be found in the literature (Baastian Kleijn and Paliwal, 1995).

In this paper, we present a new method of modifying the residual obtained after inverse filtering with linear prediction coefficients. Gimenez de los Galanes et al. (1995) modified the pitch by interpolating the residual signal, realized by either upsampling or downsampling. Both upsampling and downsampling remap the 0 to π scale to the new residual length corresponding to the given pitch modification factor. Once the residual is modified, the spectral envelope responsible for the formant structure will be superimposed by forward filtering with the same LP coefficients. Our approach is similar to the one above, but differs in the interpolation of the residual signal. Interpolation is carried out using forward and inverse orthogonal transformation of the residual signal (Rao and Yip, 1990). Traditionally, low-pass filters are used in sampling rate conversions for upsampling as well as downsampling to avoid spectral repetitions and aliasing. With the help of fast transforms, computational complexity involved in sampling rate conversion can be significantly reduced. Depending on the pitch modification factor, truncation or zero padding is performed on the forward transformed residual and the modified forward transformed residual is inverse transformed.

For a time-varying pitch modification using upsampling or downsampling, the low-pass filter must be redesigned every time, because the cutoff frequency varies according to the pitch modification factor. This could very well be avoided using an orthogonal transform, irrespective of whether the pitch modification factor is constant or time varying. We have also made some modifications to the above algorithm for handling female speech. In this method, the filter parameters are modified to produce a magnitude response that is significantly less peaky than the original linear predictive model used for inverse filtering. This reduces the filter sensitivity to pitch modification (Ansari, 1997). The discrete cosine transform (DCT) (Ahmed and Rao, 1975) has been used in our algorithm for resampling the residual. Energy loss is minimal in resampling process because DCT has high energy compaction.

2. Method

As an alternative to strictly time domain techniques, the ubiquitous source-filter model of speech can be invoked (Rabiner and Schafer, 1975). 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. Ideally, the analysis would separate the excitation signal, which could be modified independent of the vocal tract response. In practice, the system *attempts* to separate the speech signal into a spectral shaping component and a residual signal. The LPC (Makhoul, 1975) residual (error, excitation) signal has a number of advantages over the speech signal in the context of pitch modification (Edgington and Lowry, 1996). The former is spectrally flat and there is little correlation within each pitch period.

2.1. 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 speech signal has been used. In the autocorrelation domain, finding the local maxima and the distance between successive local maxima gives the periodicity of the signal under consideration. After getting the pitch information, it is subjected to various periodicity constraint rules, and linked together in order to obtain a chain of pitch marks. A nonlinear processing of these marks, which includes deletion, delay and interpolation, results in the final pitch marks positioned at the peaks of the signal in the voiced segments. Fig. 1 shows a segment of a pitch marked signal. Unvoiced segments are also marked using the same approach, which results in marks that are arbitrarily positioned. For the rest of this paper, 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 systems and is employed in our algorithm. For a 10 ms pitch period and a 16 kHz sampling rate, for example, the error in the

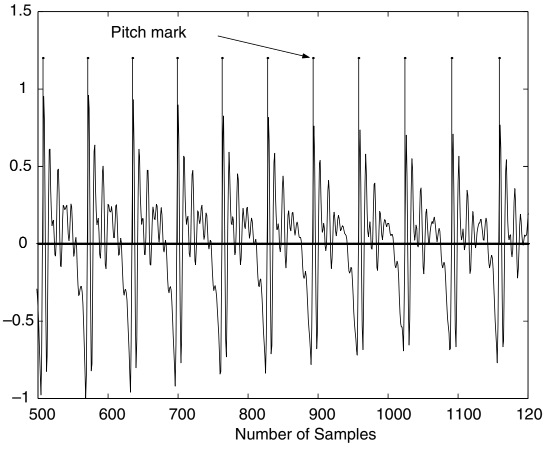

Fig. 1. Pitch marked speech '/a/'.

pitch period due to quantization is lower than 0.63%.

2.2. Resampling using DCT

Let $\{e_n; 0 \leqslant n \leqslant N_1 - 1\}$ be the residual signal obtained after pitch synchronous inverse filtering with LP coefficients. Signal expansion in orthogonal functions can be written as

$$e_n = \sum_{k=0}^{N_1-1} \theta_k \phi_k(n), \quad 0 \leqslant n \leqslant N_1 - 1$$

where,

$$\theta_k = \sum_{n=0}^{N_1-1} e_n \phi_k^{\star}(n), \quad 0 \leqslant k \leqslant N_1 - 1$$

The set of coefficients $\{\theta_k; 0 \leqslant k \leqslant N_1 - 1\}$ constitute the spectral coefficients of $\{e_n\}$ relative to the given orthonormal family of basis functions. In our algorithm, we use IDCT after truncating or zero padding θ_k to obtain a different pitch frequency and corresponding harmonics. This operation can be explained as a linear transformation $A' : R^{N_1} \to R^{N_2}$, where A' is the IDCT $N_2 \times N_2$ matrix. For $N_1 > N_2$, pitch frequency increases, and for $N_1 < N_2$, pitch frequency decreases. The forward transformation of the residual signal can be represented in matrix form as

$$\underline{\theta} = A\underline{e}$$

where, $\underline{e}^{\mathrm{T}}$ is the residual signal, A is the DCT $N_1 \times N_1$ matrix and $\{\theta_k; 0 \leqslant k \leqslant N_1 - 1\}$ are the DCT coefficients. The linear transformation of θ_k to $\{e'_l; 0 \leqslant l \leqslant N_2 - 1\}$ can be performed by premultiplication of $\underline{\theta}$ by the IDCT matrix. For $N_1 > N_2$, we truncate $\theta_k; 0 \leqslant k \leqslant N_1 - 1$ up to $N_2 - 1$ and premultiply with $N_2 \times N_2$ IDCT matrix:

$$\begin{pmatrix} e'_0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ e'_{N_2-1} \end{pmatrix} = \begin{pmatrix} \phi_0(0) & \cdot & \phi_0(N_2-1) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \phi_{N_2-1}(0) & \cdot & \phi_{N_2-1}(N_2-1) \end{pmatrix} \begin{pmatrix} \theta_0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \theta_{N_2-1} \end{pmatrix}$$

where ϕ_k's are the DCT basis vectors. For $N_1 < N_2$, we pad $(N_2 - N_1)$ zeros to $\theta_k; 0 \leqslant k \leqslant N_1 - 1$ to obtain $\theta_l; 0 \leqslant l \leqslant N_2 - 1$ and then premultiply with $N_2 \times N_2$ IDCT matrix:

$$\begin{pmatrix} e'_0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ e'_{N_2-1} \end{pmatrix} = \begin{pmatrix} \phi_0(0) & \cdot & \phi_0(N_1-1) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \phi_{N_2-1}(0) & \cdot & \phi_{N_2-1}(N_1-1) \end{pmatrix} \begin{pmatrix} \theta_0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \theta_{N_1-1} \end{pmatrix}$$

Components of the basis vectors after $\phi_k(N_1 - 1)$are not considered because, they multiply the padded zeros in θ_l and therefore, will not contribute to the output e'_l.

2.3. Modification of the residual signal

Fig. 2 shows the details of the proposed method. The process starts with pitch synchronous extraction of the residual signal. The length of the residual signal of each frame is modified using DCT-IDCT. N_1 point DCT of each frame of the excitation signal is obtained, where 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 divided by the pitch modification factor. In the DCT domain, for pitch increase, $N_1 - N_2$ trailing DCT coefficients are removed; whereas, for decreasing the pitch, $N_2 - N_1$ zeros are added to the DCT coefficients. Before taking IDCT, amplitude normalization must be carried out to compensate for the effect of change

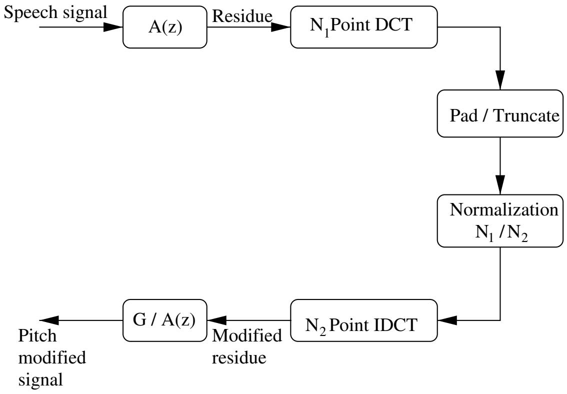

Fig. 2. Block diagram of pitch modification using DCT in the source domain.

in length of the signal. The effect of taking a 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 of the transformed residual after truncation or padding is mapped to π. Each frame of the speech signal is then synthesized by forward filtering.

2.4. Modification of all-pole filter coefficients

It is known that linear prediction using a least squares error criterion produces spectral estimates that are biased towards the pitch harmonics (Makhoul, 1975). Bandwidth estimates are typically poor. One often observes signal degradation in LPC pitch modified speech, especially for female speech. This is because of the gain variations with the new harmonic positions. Observations on the difficulty of modeling the data from a single recording are discussed in (Ansari, 1997). In their work, the filter parameters are not chosen to model the data by minimizing the residual energy, but to have sensitivity to pitch modification. The latter is accomplished by modifying the covariance matrix of the data in each frame in such a way that it produces an all-pole filter with broadened peaks. In our work, the system parameters are determined in a pitch synchronous manner. A 14th order all-pole filter was used for representing the signal in each pitch period in the voiced portion. The magnitude response is modified so as to have a significantly less peaky structure (see Fig. 3), than that which is typically obtained in LPC. In our approach, this is achieved by adaptively decreasing the radius of the corresponding pole. The polynomial in z formed by the LP filter coefficients is solved for the roots (poles), which are obtained in the polar form (angle (Θ) and radius). From theta, we get the information about the frequency of the corresponding peak in LP spectrum ($f_i = \Theta_i F_s / 2\pi$) and from radius ($r$), we get the 3-dB bandwidth ($B_i = -\ln(r_i)F_s/\pi$) of the peak, where F_s is the sampling frequency. Depending on the pitch modification factor, the radius of the pole is decreased; thus, the bandwidth is increased to accommodate the new harmonic positions. This broadening of the peaks is independent of the order used for linear prediction. The modified LP coefficients are used for both inverse and forward filtering.

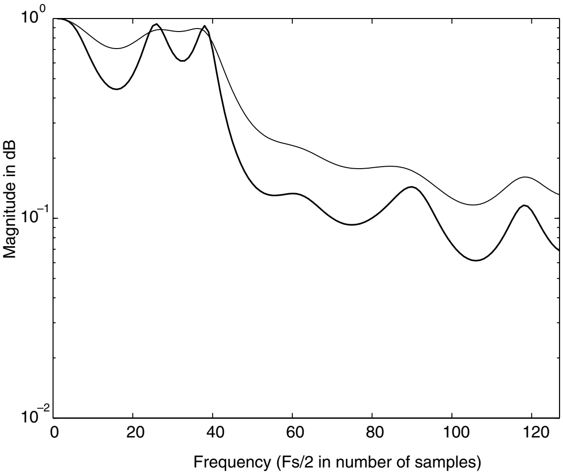

Fig. 3. LPC spectrum of a frame of speech and its peak-broadened version.

3. Results and discussion

To demonstrate the effectiveness of this technique, individual phonemes, words and sentences spoken by both male and female volunteers were recorded using SHURE mike model SM 58 in the laboratory with some ambient noise resulting in a SNR of about 25 dB. These utterances were analyzed and re-synthesized for different pitch change factors. Fig. 4(a) shows a segment of a phoneme. Fig. 4(d) gives the corresponding segment of the residual signal extracted by inverse filtering the phoneme using modified LP coefficients (model order 14). Fig. 4(e) shows the length-modified residual signal obtained through DCT-IDCT, the factor of decrease in pitch being 0.5. Fig. 4(b) shows the corresponding synthesized speech signal after forward filtering by the same modified LP coefficients. Fig. 4(f) shows the length-modified residual signal for a pitch modification factor of 1.5. Fig. 4(c) shows the corresponding synthesized speech signal after forward filtering.

The LP spectra of the above original and pitch modified signals are shown in Fig. 5. The figures

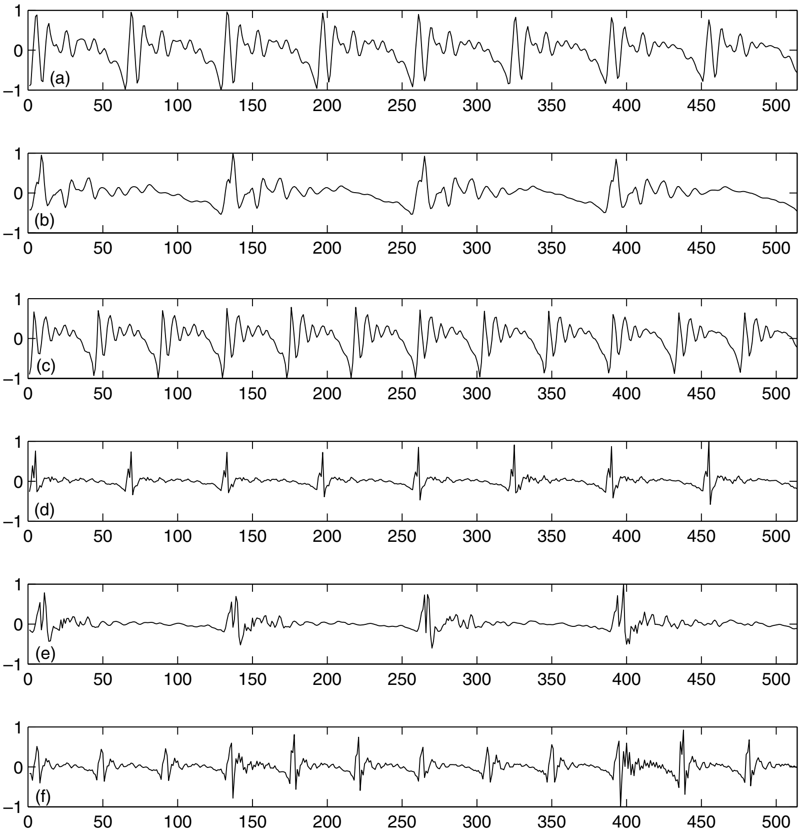

Fig. 4. (a) Few frames of the original signal '/a/'. (b) Few frames of the signal reconstructed by forward filtering the signal in (e) using modified LP coefficients. (c) Few frames of the signal reconstructed by forward filtering the signal in (f) using modified LP coefficients.(d) Few frames of the original excitation. (e) Few frames of the modified excitation for a pitch decrease factor of 0.5. (f) Few frames of the modified excitation for a pitch increase factor of 1.5.

illustrate the fact that while there is no appreciable shift in the formants for factors 0.8 and 0.5 and a minor shift for factor 1.3, there is an appreciable shift in the first formant for a pitch increase factor of 1.5. It is known that the speaker identity is not disturbed if the variation in the formant values is within ±15% (Abe, 1996) of the original values. To verify this, we evaluated the resultant speech for speaker identity too, in addition to other attributes. The intelligibility of the modified signals is found to be good. Fig. 6 shows the pitch contours for the phoneme shown in Fig. 4(a), and its pitch modified versions for factors 0.7 and 1.4. It can be seen that the shape of the pitch contour is maintained in the modified signals. Fig. 7 shows the speech signal for a whole word uttered by a female volunteer, its original pitch contour and the contours after pitch change using the technique involving modified LP coefficients. The corresponding plots for a complete sentence are shown in Fig. 8. Fig. 9(a) shows the same signal as in Fig. 8. Fig. 9(b) and (c) show the reconstructed speech for a pitch increase factor of 1.5 using direct LP method and modified LP method, respectively. The output of the modified LP method is more intelligible and has less distortion than that of the direct LP method.

The results of pitch modification by a time-varying factor using the above algorithm is shown in Fig. 10. The characteristics of interrogative sentences with an "yes or no" answer is that both the pitch contour and the amplitude rise sharply

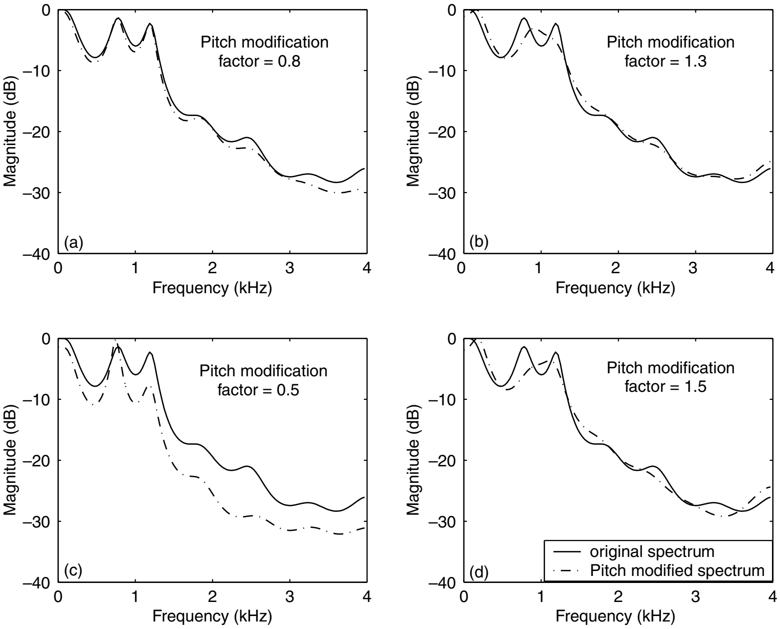

Fig. 5. LP spectra of the original signal overlapped with the LP spectra of the modified signals. (a) Pitch modification factor = 0.8. (b) Pitch modification factor = 1.3. (c) Pitch modification factor = 0.5. (d) Pitch modification factor = 1.5.

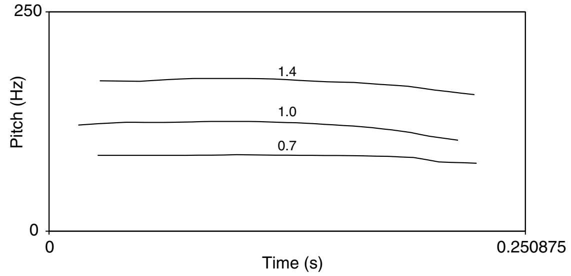

Fig. 6. Pitch contours of the original and modified phoneme '/a/'.

for the last syllable (Abe, 1996). With a linearly increasing pitch modification factor (in addition to linearly increasing amplitude modification), we have raised the pitch of the last syllable of the affirmative sentence up to a factor of 1.3 to obtain an effect of interrrogation. Fig. 11 shows energy loss due to the truncation of the DCT coefficients for pitch modification factors ranging from 1 to 2. We can observe that the energy loss is a monotonically increasing function of the pitch modification factor. It can be seen from Fig. 11. that the

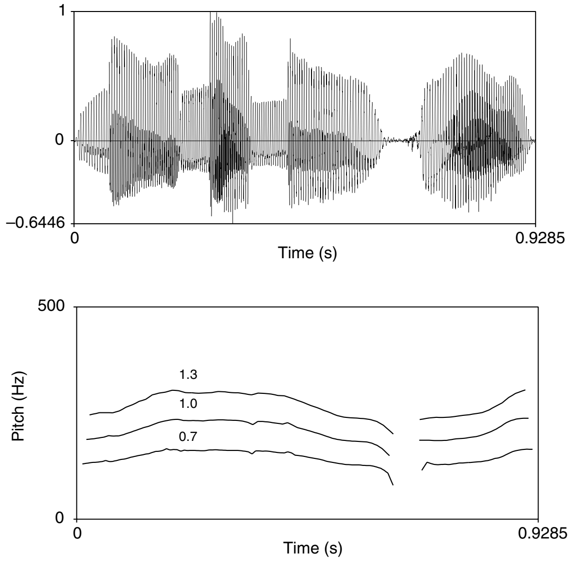

Fig. 7. Speech signal for the word niilamegha spoken by a female volunteer and pitch contours of the original and pitch modified signals.

energy loss with the modified LP method is consistently lower than that with the direct LP method,

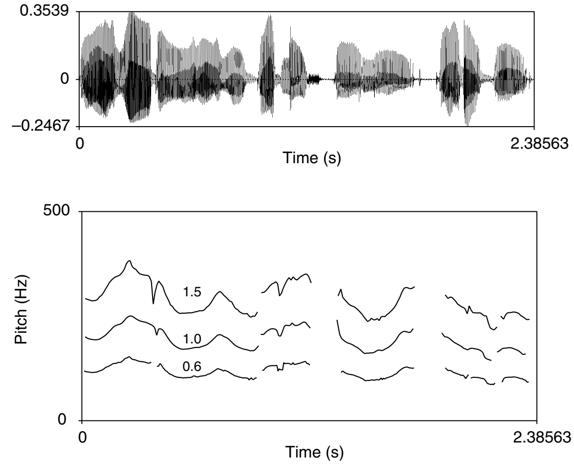

Fig. 8. Speech signal of a sentence kaaveeriya ugama sthana kodagu spoken by a female volunteer and its original and modified pitch contours.

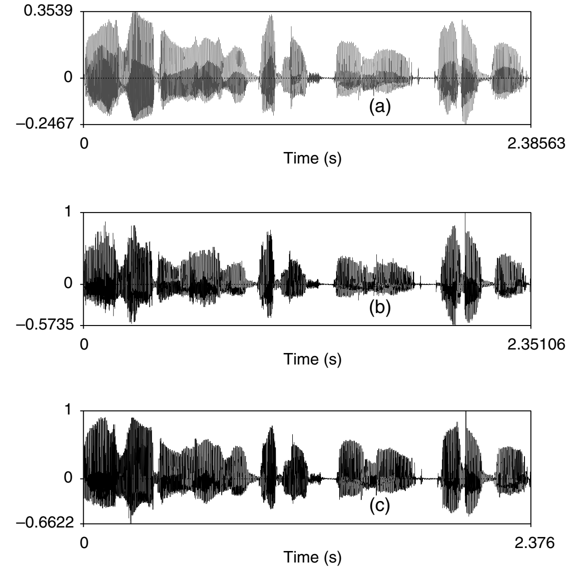

Fig. 9. Results of pitch modification using direct LP and modified LP methods for a factor of 1.5. (a) Original speech signal kaaveeriya ugama sthana kodagu spoken by a female volunteer. (b) Speech signal in (a) after pitch modification using direct LP method. (c) Speech signal in (a) after pitch modification using modified LP method.

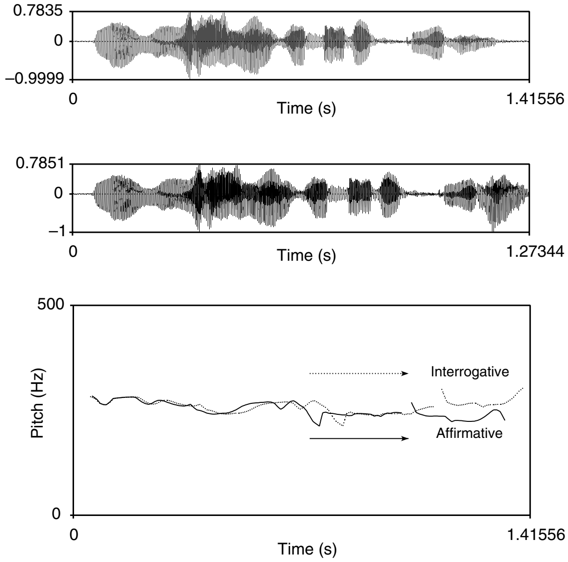

Fig. 10. Modification of affirmative sentence niivu yaavaga baruthira/ to interrogative one (female voice).

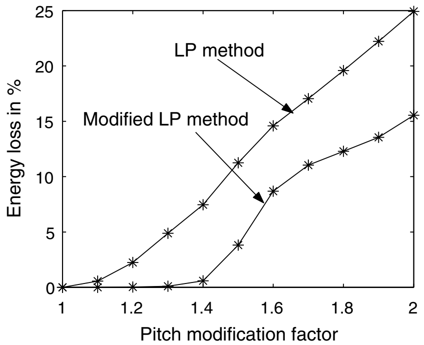

Fig. 11. Energy loss with respect to the total residual energy as a function of pitch modification factor.

across pitch modification factors ranging from 1 to 2. For example, at a pitch modification factor of 1.6, the energy loss is around 10% for the modified LP method, whereas the direct LP method results in a loss of around 15%.

To evaluate the performance of the proposed technique, perceptive evaluation tests were carried out. The pitch contours of 50 phonemes, 20 words and 10 sentences spoken by both male and female volunteers were modified by different factors. Ten people were asked to rate the intelligibility, speaker identity and distortion after the modification. Speech intelligibility is the measure of the effectiveness of speech and does not imply speech quality. We measured it by the number of evaluators who understood the message correctly. The pitch modified speech may be completely understood by the listener, but may be judged to have distortion and some change of speaker identity, and hence, to be of low quality. Speaker identity does change with large changes in the pitch. However, in our case, we are primarily interested in pitch modification to be used for speech synthesis, in which the modified speech must be perceived as belonging to the same speaker. This is what we refer to as the speaker identity. The results shown in Tables 1 and 2 clearly reflect the fact that the perceived speaker identity is drastically changed for pitch modification factors outside the range of factors 0.8–1.3. On the other hand, within the range required for speech synthesis, the speaker identity has been rated by the evaluators to be "good" for 60–70% of the cases, meaning negligible modification in the perceived speaker, and minimal observable change (rated "fair") in the remaining 30–40%. Pitch modification scheme may introduce certain unnaturalness in the resultant speech, in terms of buzziness, clicks, etc. We asked our evaluators to score the distortion present in the pitch modified speech due to the above factors.

The results of the perceptive evaluation tests for utterances from 5 male and 5 female volunteers are given in Tables 1 and 2, respectively. Perceptual evaluation revealed considerable performance differences among the different male voices tested. Similarly, there were performance differences among the various female voices tested. Of course, the performance was distinctly different for males and females, because of which we reported the results separately for males and females. From these tables, we can see that the intelligibility of the pitch modified speech is better and the distortion is less compared to TD-PSOLA and direct LP method for the modification range of 0.8–1.3. This is a reasonably sufficient range for speech synthesis. Even for the case of modification to obtain interrogative effect (shown in Fig. 10), the maximum factor we needed to use was only 1.3. The perceptual evaluation shows that even for higher pitch modification factors, the reconstructed speech has better intelligibility. From the tables, we also note that even in the case of female voice, modified LP method consistently gives better intelligibility, lower distortion and lower loss of speaker identity than the direct LP method for the entire range of pitch modification factors tested. Currently, we are using our algorithm to convert an emotional utterance to a non-emotional one,

Table 1
Perceptual evaluation after pitch modification by different techniques

Pitch change factor	Intelligibility			Distortion			Speaker identity		
	Good	Fair	Bad	Low	Medium	High	Good	Fair	Bad
(1) Pitch modification using TD-PSOLA									
Below 0.8	60	30	10	60	30	10	10	60	30
0.8–1.3	80	10	10	70	20	10	40	40	20
Above 1.3		70	30	30	30	40		40	60
(2) Direct LP method									
Below 0.8	40	40	20	50	30	20	10	50	40
0.8–1.3	70	20	10	70	30		40	50	10
Above 1.3		60	40		80	20		30	70
(3) Modified LP method									
Below 0.8	50	50		50	40	10	10	60	30
0.8–1.3	80	20		80	20		60	40	
Above 1.3		80	20	10	80	10		40	60

Results are shown as a percentage of the responses of 10 evaluators. Input utterances are from 5 males.

Table 2
Perceptual evaluation after pitch modification by different techniques

Pitch change factor	Intelligibility			Distortion			Speaker identity		
	Good	Fair	Bad	Low	Medium	High	Good	Fair	Bad
(1) Pitch modification by TD-PSOLA method									
Below 0.8	50	50			100			60	40
0.8–1.3	80	20		40	60			70	30
Above 1.3		70	30		30	70		50	50
(2) Direct LP method									
Below 0.8	50	50		20	60	20		50	50
0.8–1.3	80	20		60	40		40	40	20
Above 1.3	20	60	20	20	60	20		30	70
(3) Modified LP method									
Below 0.8	60	40		20	60	20		60	40
0.8–1.3	90	10		80	20		70	30	
Above 1.3	30	50	20	20	70	10	10	40	50

Results are shown as a percentage of the responses of 10 evaluators. Input utterances are from 5 females.

and vice versa. Thus, a sentence spoken in surprise is converted to a normal one by reducing the pitch contour by progressively decreasing factors. Since the sentence spoken in surprise is naturally of shorter duration than a normal one, there is no need for duration modification. Similarly, normal utterances have been modified to generate emotions such as surprise and anger.

Our scheme modifies the unvoiced regions too. This may not be necessary, since the unvoiced components do not carry any pitch information. However, most algorithms, including PSOLA modify the unvoiced regions too. In our case, the obtained output is not perceivably unnatural for pitch change factors from 0.8 to 1.3. We have carried out experiments on modifying only the voiced components and leaving the unvoiced portions as they are. However, in these experiments, we have not observed any perceptible improvements in the resulting signal for the range of pitch modification factors that we are interested in, as far as speech synthesis is concerned.

Further, linear prediction all-pole modeling results in a residual spectrum that contains zeros and possible spectral errors. The consequent remnant vocal tract contribution in the residual signal will be unwantedly altered by our algorithm. One cannot completely get rid of this problem. Especially, if the pitch modification factor is very high, the speaker identity might be changed. However, for the pitch modification factors we use for our synthesis, namely 0.8–1.3, there is only minor distortion in some cases, and a minor degradation in the speaker identity in around 30% of the cases, as shown by Table 1. Since the residual signal is not impulsive, the temporal structure of the utterance may not be preserved faithfully for pitch modification factors very much greater or less than 1. This is clear from the results shown in Fig. 4(b) and (c). However, for factors close to one, the resultant degradation in the intelligibility of the pitch modified signal is minimal.

The algorithm depends on pitch marking, as does PSOLA. In our text-to-speech work for the Kannada language, the basic units used for concatenation are polyphones, consisting of CV, VC, VCV, VCCV and VCCCV, where V refers to a vowel and C refers to a consonant. The total number of basic units are around 23000 and all these units have been pitch marked offline. Thus, less problems are expected due to errors in pitch marking. In order to further test the sensitivity of the proposed method with respect to pitch marking, we applied our algorithm on a few diplophonic, creaky and breathy voices and found that the technique results in intelligible outputs for such utterances too.

In pitch lowering, our method results in loss of bandwidth, and this can potentially cause distortions in the speech. However, we address this problem as follows. The recorded speech (or units) is of 8 kHz bandwidth and is sampled at 16 kHz. Thus the loss of bandwidth for pitch change factors down to 0.8 does not significantly affect the intelligibility, speaker identity and distortion level of the modified speech. Thus, one needs to have a high bandwidth speech and therefore, a high sampling rate to start with. Tables 1 and 2, which compare the performance of PSOLA and LP-based method (with and without pole modification) based on perceptual tests, show that the resulting distortion in our method is never high.

Obtaining time varying pitch modification with TD-PSOLA is a difficult task, because it involves shifting of the overlapped segment by different number of samples for different pitch synchronous frames. This is cumbersome, since the windowing effect needs to be compensated for differing lengths of overlap. However, in our case, since the DCT-IDCT operation is individually performed for each pitch synchronous frame, no additional complexity is introduced when the pitch change factor is time varying. Further, as already explained, the method proposed by Gimenez de los Galanes et al. (1995) requires redesigning the interpolation and decimation filters for varying factors of pitch change.

4. Conclusion

The proposed algorithm is simple and elegant. It directly follows from the basic source-filter 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. The proposed scheme maintains the relative pitch contour of the original signal, without any additional processing or precautions to be taken. The same basic scheme is valid for both constant and time-varying pitch modification factors. In the case of female speech, when the pitch is modified by the direct LP method even by small factors, gain difference occurs due to the peaky nature of the LPC spectrum. In such cases, the 3dB-bandwidth of the peak in LPC spectrum is adaptively increased to position the new harmonic peak so as to minimize the gain variation. This whole process widens the LPC spectral peaks and the modified algorithm is not sensitive to pitch marking errors (Ansari, 1997).

Acknowledgements

We thank the Ministry of Communication and Information Technology, Government of India for funding part of this research under the project titled “Algorithms for Kannada speech synthesis”. The authors are also grateful to the contributions of the reviewers in significantly enhancing the presentation of our results, and the discussion.

References

Abe, M., 1996. Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System, Progress in Speech Synthesis. Springer, New York, 1996.

Ahmed, N., Rao, K.R., 1975. Orthogonal Transforms for Digital Signal Processing. Springer, New York.

Ansari, R., 1997. Inverse filter approach to pitch modification: application to concatenative synthesis of female speech. In: Proc. Int. Conf. on Acoust. Speech and Signal Process., pp. 1623–1626.

Anssi, R., 1999. Pitch modification and quantization for offline speech coding, M.S. Thesis, Tampere University of Technology, May.

Baastian Kleijn, W., Paliwal, K.L., 1995. Speech Coding and Synthesis. Elsevier B.V.

Charpentier, F., Stella, M., 1986. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: Proc. Int. Conf. on Acoust., Speech and Signal Process., pp. 2015–2018.

Edgington, M., Lowry, A., 1996. Residual-based speech modification algorithms for TTS synthesis. ICSLP 96, Philadelphia, USA.

George, E.B., Smith, M.J.T., 1992. Analysis-by-synthesis/overlap -add sinusoidal modeling applied to the analysis and synthesis of musical tones. J. Audio Eng. Soc. 40 (6), 497–516.

Gimenez de los Galanes, F.M., Savoji, M., Pardo, J.M., 1995. Speech synthesis system based on a variable decimation/interpolation factor. In: Proc. Int. Conf. on Acoust. Speech and Signal Process., pp. 636–639.

Liberman, M., 1994. Computer speech synthesis: its status and prospects, Voice Communication between Humans and Machines. National Academy of Sciences.

Makhoul, J., 1975. Linear prediction: a review. Proc. IEEE 63, 561–580.

McAulay, R.J., Quatieri, T.F., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust., Speech, Signal Process. 34 (4), 744–754.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453–468.

Portnoff, M.R., 1981. Time-scale modification of speech based on short-time Fourier analysis. IEEE Trans. Acoust., Speech, Signal Process. 30, 374–390.

Quatieri, T.F., McAulay, R.J., 1986. Speech transformation based on a sinusoidal representation. IEEE Trans. Acoust., Speech, Signal Process. 34 (6), 1449–1464.

Rabiner, L.R., Schafer, R.W., 1975. Digital Processing of Speech Signals. Prentice-Hall, Inc., Englewood Cliffs, New Jersy 07632.

Rao, K.R., Yip, P., 1990. Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, New York.

Roe, D.B., Wilpon, J.G. (Eds.), 1994. Voice Communication between Humans and Machines. National Academy of Sciences.

Syrdal, A., Bennet, R., Greenspan, S., 1995. Applied Speech Technology. CRC Press, Boca Raton, FL.

Vergin, R., O'Shaughnessy, D., Farhat, A., 1997. Time domain technique for pitch modification and robust voice transformation. In: Proc. ICASSP 97, Vol. II of V, Speech Processing, pp. 947–950.