Prosody Modification Using Instants of Significant Excitation

K. Sreenivasa Rao and B. Yegnanarayana, *Senior Member, IEEE*

***Abstract*—Prosody modification involves changing the pitch and duration of speech without affecting the message and naturalness. This paper proposes a method for prosody (pitch and duration) modification using the instants of significant excitation of the vocal tract system during the production of speech. The instants of significant excitation correspond to the instants of glottal closure (epochs) in the case of voiced speech, and to some random excitations like onset of burst in the case of nonvoiced speech. Instants of significant excitation are computed from the linear prediction (LP) residual of speech signals by using the property of average group-delay of minimum phase signals. The modification of pitch and duration is achieved by manipulating the LP residual with the help of the knowledge of the instants of significant excitation. The modified residual is used to excite the time-varying filter, whose parameters are derived from the original speech signal. Perceptual quality of the synthesized speech is good and is without any significant distortion. The proposed method is evaluated using waveforms, spectrograms, and listening tests. The performance of the method is compared with linear prediction pitch synchronous overlap and add (LP-PSOLA) method, which is another method for prosody manipulation based on the modification of the LP residual. The original and the synthesized speech signals obtained by the proposed method and by the LP-PSOLA method are available for listening at http://speech.cs.iitm.ernet.in/Main/result/prosody.html.**

***Index Terms*—Duration, excitation source, instants of significant excitation (epochs), linear prediction pitch synchronous overlap and add (LP-PSOLA), LP residual, pitch period, prosody modification.**

I. Introduction

THE OBJECTIVE of prosody modification is to alter the pitch contour and durations of the sound units of speech without affecting the shapes of the short-time spectral envelopes [1]. Prosody modification is useful in a variety of applications related to speech communication [2]–[4]. For instance, in a text-to-speech (TTS) system, it is necessary to modify the durations and pitch contours of the basic units and words in order to incorporate the relevant suprasegmental knowledge in the utterance corresponding to the sequence of these units [5]. Time-scale (duration) expansion is used to slow down rapid or degraded speech to increase the intelligibility [6]. Time-scale compression is used in message playback systems for fast scanning of the recorded messages [6]. Frequency-scale modification is often performed to transmit speech over limited bandwidth communication channels, or to place speech in a desired bandwidth as an aid to the hearing impaired [7]. While pitch-scale modification is useful for a TTS system, formant modification techniques are also used to compensate for the defects in the vocal tract and for voice conversion [3], [8], [9].

In this paper, a method for prosody (pitch and duration) modification is proposed using the knowledge of the instants of significant excitation. The instants of significant excitation refer to the instants of glottal closure in the voiced region and to some random excitations like the onset of burst in the case of nonvoiced regions [10], [11]. The instants of significant excitation are also termed as *epochs*. The proposed method does not distinguish between voiced and nonvoiced regions in the implementation of the desired prosody modification. The method also does not involve estimation of any specific speech parameters like fundamental frequency (F_0). Since the modification is done in the excitation component of the signal, there will be no discontinuities perceived in the synthesized speech, even when large values of prosody modification factors are used.

Several approaches are available in the literature for prosody modification [1], [4], [6], [12]–[18]. Approaches like overlap and add (OLA), synchronous overlap and add (SOLA), and pitch synchronous overlap and add (PSOLA) operate directly on the waveform (time domain) to incorporate the desired prosody information [4], [12], [13]. In some of the approaches for prosody modification, the speech signal is represented in a parametric form, as in the harmonic plus noise model (HNM), speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT), and sinusoidal modeling [1], [14]–[17], [19]. Pitch modification based on discrete cosine transform (DCT) incorporates the required pitch modification by modifying the LP residual [18]. Some approaches use phase vocoders for time-scale modification [6].

We review briefly some approaches which are closely related to the method proposed in this paper. In the overlap and add approach for time-scale modification, the speech signal is split into short (about two to three pitch periods long) segments using overlapping analysis windows [12]. Each segment is multiplied with a Hann window. For synthesis, the windowed segments are overlapped and added. Based on the desired time-scale modification, some of the windowed segments are either replicated or omitted. In these cases, the information about the pitch markers was not used for splitting the speech signal into short segments. Hence, the periodicity due to pitch was not preserved well after the time-scale modification. The SOLA approach allows flexible positioning of the windowed segments by searching for the

Manuscript received May 1, 2004; revised April 25, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thierry Dutoit.

K. S. Rao is with the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India (e-mail: ksrao@iitg.ernet.in).

B. Yegnanarayana is with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India (e-mail: yegna@cs.iitm.ernet.in).

Digital Object Identifier 10.1109/TSA.2005.858051

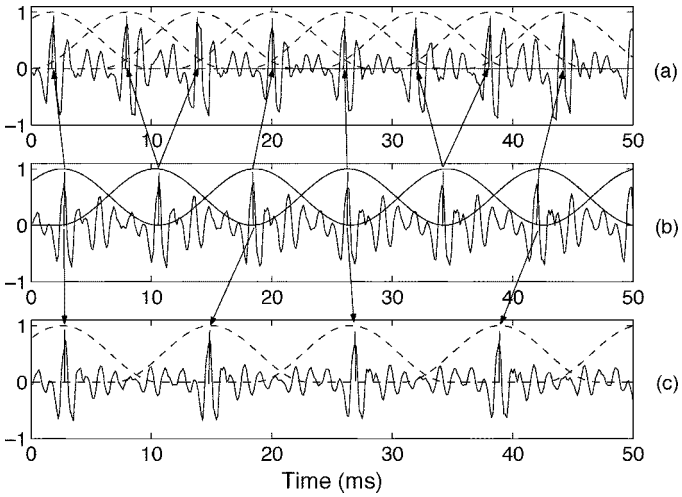


Fig. 1. Pitch period modification using TD-PSOLA method. (a) Modified speech signal using TD-PSOLA method for pitch period modification factor $\alpha = 0.75$. (b) A segment of voiced speech (original). (c) Modified speech signal using TD-PSOLA method for pitch period modification factor $\alpha = 1.5$.

placement of the analysis window in such a way that the overlapped regions have maximum correlation [13].

While the OLA and SOLA approaches are limited to time-scale modification, the PSOLA approach can be applied to both time and pitch-scale modification [2], [4]. There are several versions of the PSOLA algorithm [4], [20]. The time-domain version, called TD-PSOLA, is most commonly used due to its computational efficiency [21]. The basic method consists of deriving pitch synchronous analysis segments, using pitch markers [2], [22]. The pitch markers can be determined by using a pitch extraction algorithm [21]. Analysis windows are typically of length two or four pitch periods, and are centered around the pitch marker.

Manipulation of the pitch is achieved by changing the time intervals between the pitch markers. Fig. 1 depicts the modification of pitch using TD-PSOLA approach. The modification of duration is achieved by either repeating or omitting the speech segments. The TD-PSOLA suffers from spectral and phase distortions due to direct manipulation of the speech signal. Other variations of PSOLA, namely, frequency domain PSOLA (FD-PSOLA) and linear prediction PSOLA (LP-PSOLA), are theoretically more suitable for pitch-scale modification, because they provide independent control over the spectral envelope for synthesis. The FD-PSOLA is used only for pitch-scale modification. The LP-PSOLA is used for both pitch and duration modification using the principle of residual excited vocoders [4].

In HNM, the speech signals are represented as a time-varying harmonic component plus a modulated noise component [14], [15]. The decomposition of speech into these two components allows for more flexible and natural-sounding modification of the prosody parameters of the speech signal. However, estimation of relevant parameters such as the fundamental frequency (F_0), the maximum voiced frequency, and synthesis time instants involve complex computations. Moreover, the method requires some post processing to reduce the interframe incoherence problem (phase mismatch between frames from different acoustic units) [15]. In the STRAIGHT approach, the speech signals are manipulated based on pitch-adaptive spectral smoothing and instantaneous-frequency-based F_0 extraction [16], [17]. In this approach, the speech signal is represented using a sequence of F_0 values and the pitch synchronous spectral envelope. The speech parameters are adjusted according to the desired modification either for speech rate modification or F_0 modification. This approach offers greater flexibility than the HNM approach for parameter manipulations without introducing the artificial timbre, specific to synthetic speech signals [17]. However, this approach requires estimation of instantaneous F_0 and smoothing of the F_0 trajectory.

A different approach for prosody modification is adopted in sinusoidal modeling [1]. In the sinusoidal modeling, the speech signal is characterized by amplitudes, frequencies, and phases of the component sine waves. These parameters are estimated from the short time Fourier transform of speech. For a given frequency track, a cubic function is used to interpolate the phase as per the desired prosody parameters. This modified phase function is applied to a sine wave generator. The outputs of each of the sine wave generators is amplitude modulated, and is added to similar outputs of the sine wave generators for the other frequency tracks to produce the desired prosody modification. Problems arise when changing the pitch by a large scale factor. In particular, hoarseness was perceived in the reconstruction, when F_0 was increased [23].

The discrete cosine transform (DCT) based approach performs the pitch modification in the residual domain [18]. For a given speech signal, the LP residual is extracted using the LP analysis. Pitch markers are computed using the autocorrelation function of the residual. The residual in each pitch period is accessed using the pitch markers. The residual is interpolated using DCT and inverse DCT to provide the desired pitch modification.

The digital phase vocoder relates the amplitudes and frequencies of the outputs of a digital filterbank to the properties of excitation and vocal tract. The refined phase vocoder proposed by Portnoff takes advantage of the computational efficiency of the fast Fourier transform for implementation. However phase distortions due to pitch modification occur in the synthetic speech signal with objectionable reverberant quality [6].

Methods for prosody modification generally produce some spectral and phase distortions. This is mainly due to manipulation of the speech signal directly. The distortions are reduced to a large extent by operating on the residual obtained from the linear prediction analysis. In this paper, we propose a method for prosody modification which operates on the linear prediction residual using the knowledge of the instants of significant excitation as pitch markers. The instants of significant excitation are computed using group-delay analysis [11]. The group-delay-based method is robust, and it gives accurate epoch locations even under some mild degradations due to background noise and reverberation [10]. The region around the instant of glottal closure correspond to the significant part of excitation, in the sense that the strength of excitation is maximum in that region of the pitch period. Therefore, we attempt to retain that region during pitch period modification. Since methods like LP-PSOLA also use residual manipulation for prosody modification, we compare the performance of our method with the results obtained by LP-PSOLA. An important feature of the proposed method is that the instants of significant excitation in both the voiced and nonvoiced regions are treated alike.

The basic principle of the proposed method for prosody modification is presented in Section II. Modification of pitch period is discussed in Section III, and the modification of the duration is discussed in Section IV. After obtaining the modified (new) epoch sequence according to the desired prosody information, the next step is to modify the LP residual according to the new sequence of epochs. The process of modification of the LP residual is discussed in Section V, and the process of generating the synthesized speech by exciting the time varying all-pole filter with the modified residual is discussed in Section VI. The performance of the proposed prosody manipulation method is compared with the LP-PSOLA method in Section VII. In Section VIII, a summary and possible extensions of this work are given.

II. Proposed Method for Prosody Modification

The proposed method for prosody manipulation makes use of the properties of the excitation source information for prosody modification. The residual signal in the LP analysis is used as an excitation signal [24]. The successive samples in the LP residual are less correlated compared to the samples in the speech signal. The residual signal is manipulated by using resampler either for increasing or decreasing the number of samples required for the desired prosody modification. The residual manipulation is likely to introduce less distortion in the speech signal synthesized using the modified LP residual and LP coefficients (LPCs). The time varying vocal tract system characteristics are represented by the LPCs for each analysis frame. Since the LPCs carry the information about the short-time spectral envelope, they are not altered in the proposed method for prosody modification. LP analysis is carried out over short segments (analysis frames) of speech data to derive the LP coefficients and the LP residual for the speech signal [24].

There are four main steps involved in the prosody manipulation.

- Step 1) Deriving the instants of significant excitation (epochs) from the LP residual signal.
- Step 2) Deriving a modified (new) epoch sequence according to the desired prosody (pitch and duration).
- Step 3) Deriving a modified LP residual signal from the modified epoch sequence.
- Step 4) Synthesizing speech using the modified LP residual and the LPCs.

In this section, we will briefly discuss the method of extracting the instants of significant excitation (or epochs) from the LP residual [10], [11]. Methods for pitch period modification and duration modification are described in Sections III and IV, respectively. The group-delay analysis is used to derive the instants of significant excitation from the LP residual [10], [11]. The analysis involves computation of the average slope of the unwrapped phase spectrum (i.e., average group-delay) for each frame. If $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of the windowed signal $x(n)$ and $nx(n)$, respectively, then the group-delay function $\tau(\omega)$ is given by the derivative of the phase function $\phi(\omega)$ of $X(\omega)$, and is given by [10], [25]

$$\tau(\omega) = -\phi'(\omega) = \frac{X_R Y_R + X_I Y_I}{X_R^2 + X_I^2}$$

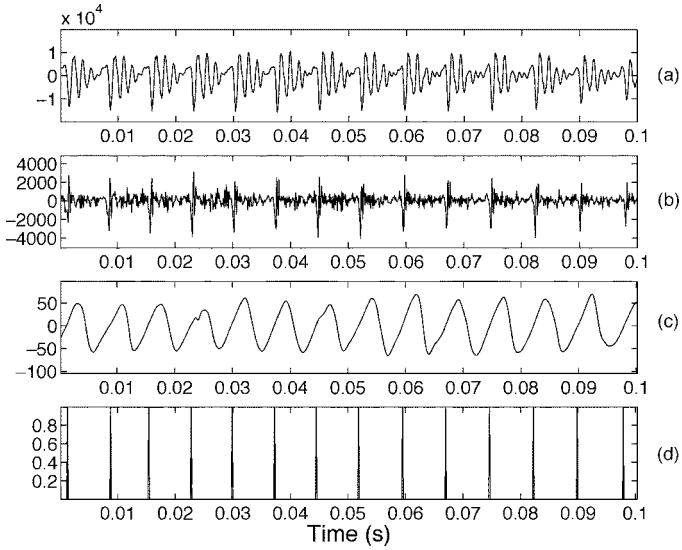


Fig. 2. (a) Segment of voiced speech and its (b) LP residual. (c) Phase slope function. (d) Instants of significant excitation.

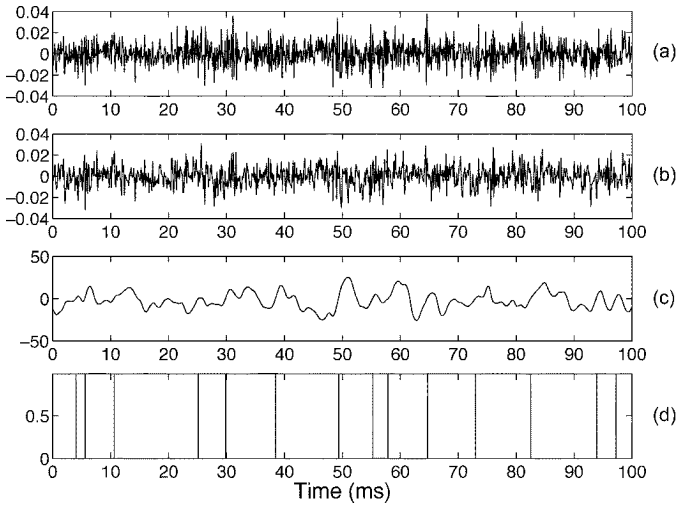


Fig. 3. (a) Segment of nonvoiced speech and its (b) LP residual. (c) Phase slope function. (d) Instants of significant excitation.

where $X_R + jX_I = X(\omega)$, and $Y_R + jY_I = Y(\omega)$. Any isolated sharp peaks in $\tau(\omega)$ are removed by using a three-point median filtering. Note that all the Fourier transforms are implemented using the discrete Fourier transform. The average value $\bar{\tau}$ of the smoothed $\tau(\omega)$ is the value of the *phase slope function* for the time instant corresponding to the center of the windowed signal $x(n)$. The phase slope function is computed by shifting the analysis window by one sample at a time. The instants of positive zero-crossings of the phase slope function correspond to the instants of significant excitation. Figs. 2 and 3 illustrate the results of extraction of the instants of significant excitation for voiced and nonvoiced speech segments, respectively. For generating these figures, a tenth-order LP analysis is performed using a frame size of 20 ms and a frame shift of 5 ms. Throughout this study, a signal sampled at 8 kHz is used. The signal in the analysis frame is multiplied with a Hamming window to generate a windowed signal. Note that for nonvoiced speech, the epochs occur at random instants, whereas for voiced speech the epochs occur in the regions of the glottal closure, where the LP

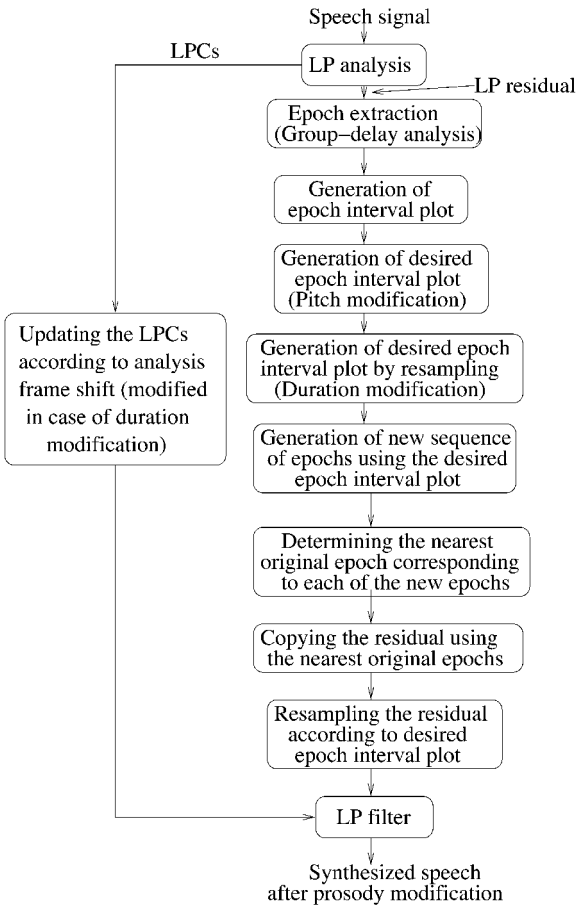


Fig. 4. Block diagram for prosody modification.

residual error is large. The time interval between two successive epochs correspond to the pitch period for voiced speech. With each epoch, we associate three parameters, namely, time instant, epoch interval, and LP residual. We call these *epoch parameters*.

The prosody manipulation involves deriving a new excitation (LP residual) signal by incorporating the desired modification in the duration and pitch period for the utterance. This is done by first creating a new sequence of epochs from the original sequence of epochs. For this purpose, all the epochs derived from the original signal are considered, irrespective of whether they correspond to a voiced segment or a nonvoiced segment. The methods for creating the new epoch sequence for the desired prosody modification are discussed in Sections III and IV.

For each epoch in the new epoch sequence, the nearest epoch in the original epoch sequence is determined, and, thus, the corresponding epoch parameters are identified. The original LP residual is modified in the epoch intervals of the new epoch sequence, and, thus, a modified excitation (LP residual) signal is generated. The modified LP residual signal is then used to excite the time varying all-pole filter represented by the LPCs. For pitch period modification, the filter parameters (LPCs) are updated according to the frame shift used for analysis of the original signal. For duration modification, the LPCs are updated according to the modified frame shift value. Generation of the modified LP residual according to the desired pitch period and duration modification factors is described in Section V, and the speech synthesis procedure in Section VI. Fig. 4 shows the block diagram indicating various stages in prosody modification.

III. Pitch Period Modification

The objective is to generate a new epoch sequence, and then a new LP residual signal according to the desired pitch period modification factor. Note that for generating this signal, all epochs in the sequence are considered, without discriminating the voiced or nonvoiced nature of the segment to which the epoch belongs. Thus, it is not necessary to identify the voiced, unvoiced, and silence regions of speech.

Fig. 5 illustrates the prosody modification where the pitch period is reduced by a factor $\alpha = 0.75$. The original epoch interval plot is shown by the solid curve, and the desired epoch interval plot is shown by the dotted curve, which is obtained by multiplying the solid curve by α. The original epochs are marked by circles ('$\circ$') on these two curves. The epoch interval at any circle is the spacing (in number of samples) between this and the next circle. The solid and dotted curves are obtained by joining the epoch interval values.

Starting from the point A, the new epoch interval value is obtained from the dotted curve, and this value is used to mark the next new epoch B along the x-axis. The epoch interval at this instant on the dotted curve is used to generate the next new epoch C, and so on. The new epoch sequence is marked as "$\times$" along the dotted curve, and also along the x-axis. The nearest original epoch for each of these new epochs is also marked as a sequence of circles ("$\circ$") along the x-axis.

The procedure for the generation of new epoch sequence, when the pitch period is scaled up, is similar to the one used for the case of Fig. 5, except that the new epoch interval values are obtained from the scaled up plot. Note that in the above discussion for pitch period modification, the random epoch intervals in the nonvoiced regions are modified by the same pitch period modification factor. As we will see later in Section VI, this will not have any effect on the synthesized speech.

IV. Duration Modification

Generation of new epoch sequence for duration modification is illustrated in Fig. 6 for a duration increase by $\beta = 1.5$ times. For generating the desired epoch interval plot for duration modification, the original epoch interval plot (solid line in Fig. 6) is resampled according to the desired modification factor.

The desired epoch interval plot is shown by the dotted curve. The modified (new) epoch sequence is generated as follows. Starting with the point A in Fig. 6, the epoch interval value is obtained from the dotted curve, and it is used to determine the next epoch instant B. The value of the next epoch interval at B is obtained from the dotted curve, and this value is used to mark the next new epoch C. The new epochs generated by this process are marked as "$\times$" along the x-axis in Fig. 6. The new epochs are also marked ("$\times$") on the desired epoch interval plot along with the mapped original epochs ("$\circ$"). Those mapped original epochs nearest to the new epochs are shown along the x-axis by circles ("$\circ$"). In a similar manner, the new epochs are generated for the case of decrease of duration. Note that in this case also, no distinction is made for epochs in the voiced and nonvoiced regions. Fig. 7 shows the generated new epoch sequence when both the pitch period and duration are modified simultaneously.

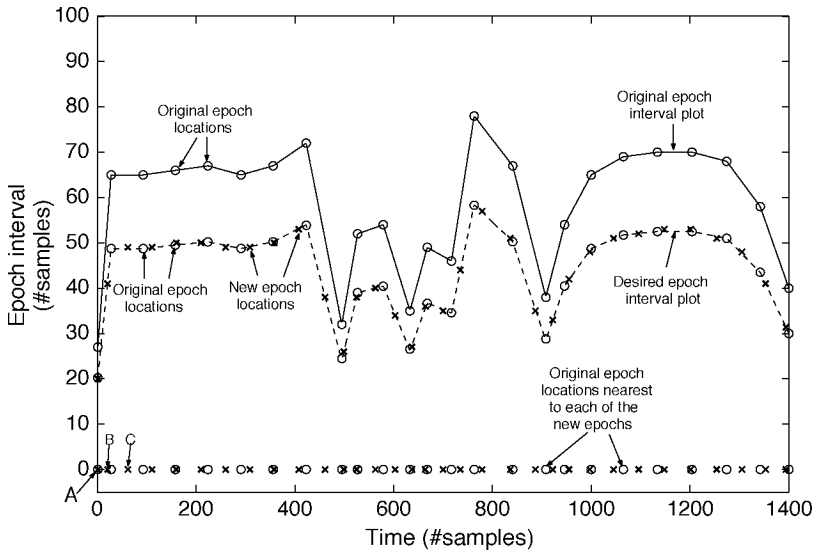


Fig. 5. Generation of new sequence of epochs for the modification of pitch period by a factor $\alpha = 0.75$.

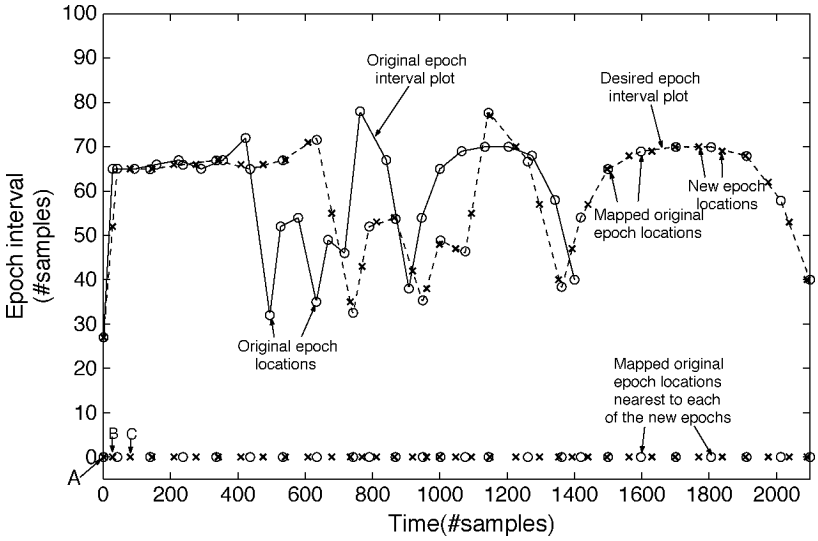


Fig. 6. Generation of new sequence of epochs for the modification of duration by a factor $\beta = 1.5$.

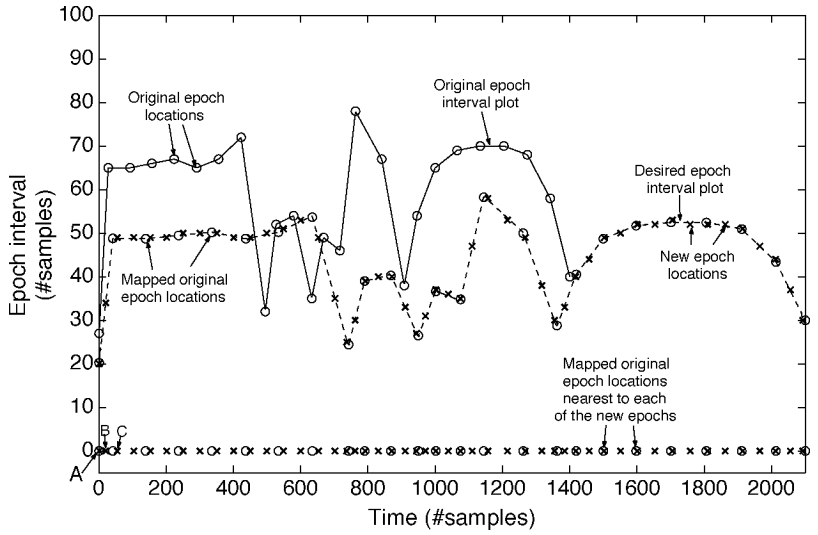


Fig. 7. Generation of new sequence of epochs for the modification of pitch period by a factor $\alpha = 0.75$ and duration by a factor $\beta = 1.5$.

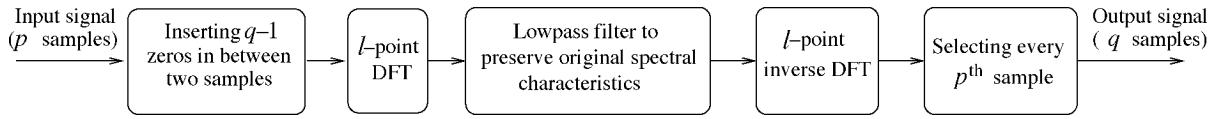


Fig. 8. Method for resampling the input signal by a factor q/p.

V. Modification of LP Residual

After obtaining the modified epoch sequence, the next step is to derive the excitation signal or LP residual. For this, the original epoch (represented by a "○") closest to the modified epoch ("×") is determined from the sequence of "○" and "×" along the desired epoch interval curve (dotted curves in each of the Figs. 5–7). As mentioned earlier, with each original epoch, i.e., the circles ("○") in the plots, there is an associated LP residual sequence of length equal to the value of the original epoch interval for that epoch. The residual samples are placed starting from the corresponding new epoch. Since the value of the desired epoch interval (M) is different from the value of the corresponding original epoch interval (N), there is a need either to delete some residual samples or append some new residual samples to fill the new epoch interval.

Increasing or decreasing the number of LP residual samples for pitch period modification can be done in two ways. In the first method, all the residual samples are used to resample them to the required number of new samples. While there is no distortion perceived in the synthetic speech, the residual samples are expanded or compressed even in the crucial region around the instant of glottal closure. In the second method, a small percentage of the residual samples within a pitch period are retained (i.e., they are not modified), and the rest of the samples are expanded and compressed depending on the pitch period modification factor. The residual samples to be retained are around the instant of glottal closure, as these samples determine the strength and quality of excitation. Thus, by retaining these samples, we will be preserving the naturalness of the original voice.

The percentage of samples to be retained around the instant of glottal closure may not be critical, but if we use a small number (say less than 10% of pitch period) of samples, then we may miss some crucial information in some pitch periods, especially when the period is small. On the other hand, if we consider large number (say about 30%) of samples, then we may include the complete glottal closure region, which will not change in proportion when the pitch period is modified.

We have examined the effect of retaining the percentage of the residual samples by subjective evaluation. We have considered three cases, namely, 0%, 20%, and 33% of the residual samples to be retained around the instant of glottal closure. No significant difference was perceived in the quality of the synthetic speech. In fact, listening tests gave nearly the same level of confidence in all the cases. We have chosen to retain 20% of the residual samples in this study.

In this study, we resample the residual samples instead of deleting or appending the samples. The first $0.2N$ (nearest integer) residual samples are retained and the remaining $(p = N - 0.2N)$ residual samples are resampled to generate $(q = M - 0.2N)$ new samples. The resampling is done as follows: Resampling is performed by inserting $q - 1$ zero value samples in between successive original residual samples. The resulting samples are appended with zeros to obtain the number of samples to the nearest power of 2, i.e., $2^m = l$, where $2^{m-1} < p * q < 2^m$. An l-point DFT is obtained on this data. The DFT is lowpass filtered to preserve the spectral characteristics of the original residual signal, and, thus, avoiding repetition of the spectrum of the original residual samples due to upsampling. An l-point inverse DFT is performed on the lowpass filtered DFT to obtain the samples in the time domain. The desired number (q) of residual samples are derived by selecting every pth sample from the new samples in time domain. The process of resampling is shown in Fig. 8.

VI. Generating the Synthetic Signal

The modified LP residual signal is used as an excitation signal for the time varying all-pole filter. The filter coefficients are updated for every P samples, where P is the frame shift used for performing the LP analysis. In these studies, a frame shift of 5 ms and a frame size of 20 ms are used for LP analysis. Thus, the P samples correspond to 5 ms when the prosody modification does not involve any duration modification. On the other hand, if there is a duration modification by a scale factor β, then the filter coefficients (LPCs) are updated for every P samples corresponding to 5β ms.

Since the LP residual is used for incorporating the desired prosody modification, there is no significant distortion due to resampling the residual samples both in the voiced and in the nonvoiced regions. This is because there is less correlation among samples in the LP residual compared to the correlation among the signal samples.

Fig. 9 shows the speech waveforms and narrowband spectrograms for pitch period modification. It can be noted that there are no discontinuities in the synthesized speech waveforms or in their spectrograms. Most of the features (such as pitch changes and formant transitions) seem to have been preserved well. Similar characteristics were observed in the waveforms and spectrograms for duration modification. This is also verified by perceptual studies discussed in Section VII.

VII. Comparison of the Performance of Proposed Method With LP-PSOLA Method

The performance of the proposed epoch-based prosody modification method is compared with that of the LP-PSOLA method, since the latter method also performs pitch period and duration modification by manipulating the LP residual. The LP-PSOLA method performs pitch and time-scale modifications using pitch markers as anchor points. Here, we have used the instants of significant excitation (epochs) derived by the group-delay analysis as the pitch markers for performing the pitch period and duration modification. We will briefly describe the process of residual manipulation used in the LP-PSOLA method.

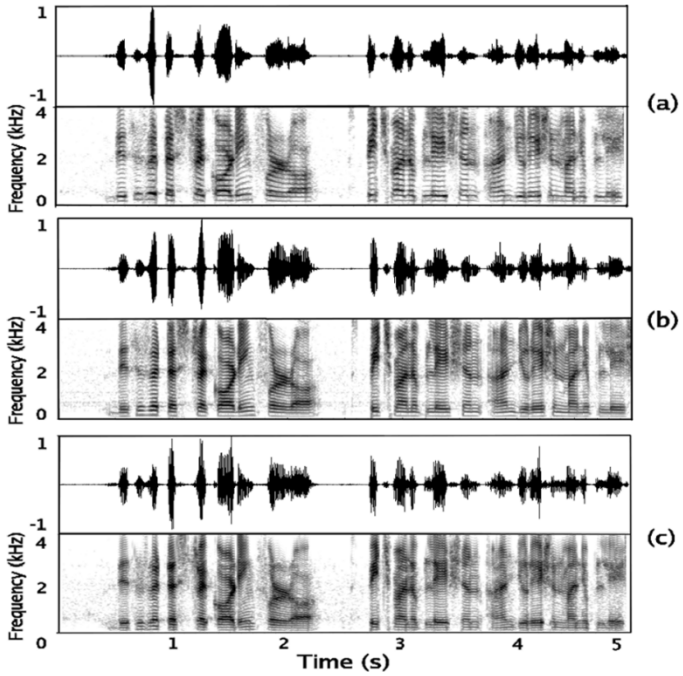


Fig. 9. Speech signal and its narrowband spectrogram for the utterance "*This is a test recording for pitch manipulation and duration manipulation*." (a) Pitch period modification factor $\alpha = 0.66$, (b) original, and (c) pitch period modification factor $\alpha = 1.33$.

The LP residual signal is divided into overlapping short-time segments by multiplying the LP residual signal by a sequence of pitch-synchronous analysis windows. A Hann window is used in the analysis, and the window is centered around the pitch marker (the instant of glottal closure) for each pitch period. The size of the window is about two pitch periods, where the pitch period is estimated as the average of two intervals of the pitch markers on either side of the current pitch marker. New pitch markers are generated from the pitch markers of the given speech signal according to the desired pitch period modification factor. In the case of pitch period modification, the residual signal is manipulated by positioning the windowed segments centered around the new pitch markers, and then adding the overlapped regions. For maintaining the length of the speech signal same before and after the pitch period modification, some of the windowed residual segments are dropped in the case of increasing pitch period, and some of the segments are replicated in the case of decreasing pitch period. Fig. 10 shows the manipulation of the LP residual for pitch period modification using the LP-PSOLA method for the illustration shown in Fig. 1. For duration modification, the interval between the pitch markers is not changed, and new time-scaled pitch markers are derived according to the desired modification factor. Finally, the duration modification is realized by deleting or replicating some of the windowed residual segments.

Perceptual evaluation was carried out by conducting subjective tests with 25 research scholars in the age group of 25–35 years. The subjects have sufficient speech knowledge for proper assessment of the speech signals, as all of them have taken a full semester course on speech technology. Four sentences were randomly chosen from the TIMIT database to perform the test [26]. The sentences were spoken by two male and two female

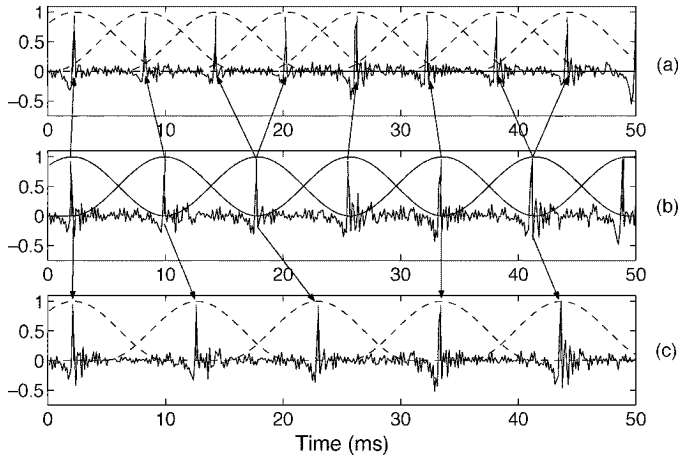


Fig. 10. Pitch period modification using LP-PSOLA method. (a) Modified residual signal using LP-PSOLA method for pitch period modification factor $\alpha = 0.75$. (b) LP residual for a segment of voiced speech. (c) Modified residual signal using LP-PSOLA method for pitch period modification factor $\alpha = 1.33$.

TABLE I
RANKING USED FOR JUDGING THE QUALITY AND DISTORTION OF THE SPEECH SIGNAL MODIFIED BY DIFFERENT MODIFICATION FACTORS

Rating	Speech quality	Level of distortion
1.	Unsatisfactory	Very annoying and objectionable
2.	Poor	Annoying but not objectionable
3.	Fair	Perceptible and slightly annoying
4.	Good	Just perceptible but not annoying
5.	Excellent	Imperceptible

TABLE II
MEAN OPINION SCORES AND CONFIDENCE VALUES FOR DIFFERENT PITCH PERIOD MODIFICATION FACTORS

Pitch period modification factor (α)	Mean opinion score (MOS)		Level of confidence in % for the significance of difference in MOSs
	LP-PSOLA method	Epoch-based method	
2	3.38	3.94	> 99.5
1.33	4.22	4.48	> 99.5
0.66	4.39	4.57	> 97.5
0.5	3.93	4.42	> 99.5
0.4	3.79	4.19	> 99.5

speakers. For each sentence, the pitch period was modified by factors 2, 1.33, 0.66, 0.5, and 0.4, which corresponds to pitch frequency modification factors 0.5, 0.75, 1.5, 2, and 2.5, respectively. After modification using the proposed method and the LP-PSOLA method, the file names were coded to avoid bias toward a specific method. Each of the subjects were given a pilot test about perception of speech signals for different pitch period modification factors. Once they were comfortable with judging, they were allowed to take the tests. The tests were conducted in the laboratory environment by playing the speech signals through headphones. In the test, the subjects were asked to judge the distortion and quality of the speech for various modification factors. Subjects were asked to assess the quality and distortion on a five-point scale for each of the sentences obtained by both the methods. The five-point scale for representing the quality of speech and the distortion level is given in Table I [27]. Altogether each subject had to judge 40 sentences.

The mean opinion scores (MOSs) for each of the pitch period modification factors are given in Table II. The significance of

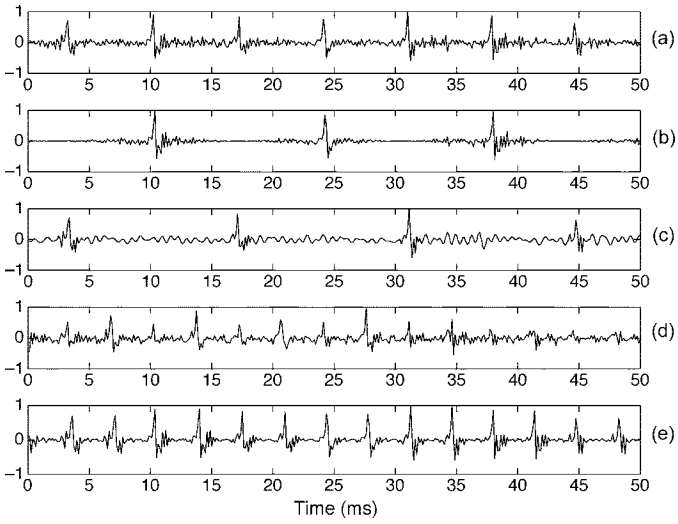


Fig. 11. (a) LP residual for a segment of voiced speech. (b) Modified LP residual signal using LP-PSOLA method for pitch period modification factor $\alpha = 2$. (c) Modified LP residual signal using epoch-based method for $\alpha = 2$. (d) Modified LP residual signal using LP-PSOLA method for $\alpha = 0.5$. (e) Modified LP residual signal using epoch-based method for $\alpha = 0.5$.

the differences in the pairs of the MOSs is tested using hypothesis testing [28]. The level of confidence for the observed differences in the sample means was obtained in each case using the sample variances and values of Student-t distribution. The level of confidence is high (>97.5%) in all cases. This indicates that the differences in the pairs of the MOSs in each case is significant. Hence, the epoch-based (proposed) method is preferred over the LP-PSOLA method by all the subjects.

The scores also indicate that the performance is low for the pitch period modification factors 2 and 0.4. For small modifications (0.66 and 1.33) both the methods seem to perform equally well. For the pitch period modification factor of 2, the performance of the LP-PSOLA method is slightly inferior to that of the epoch-based modification method. In the LP-PSOLA method increase in the pitch period upto a factor of 2 can be achieved using a segment of window length of two pitch periods. For further increase in the pitch period, the window size should include more than two pitch periods. For instance, using a window size of four pitch periods, the pitch period can be increased upto four times. However, with large window sizes, secondary excitations are introduced. For the modification at the lower end (less than 0.5), the LP-PSOLA introduces some audible distortion, whereas the epoch-based method gives lesser distortion. The LP-PSOLA method produces phase mismatches and audible distortion due to overlap and adding of the windowed residual segments, for both increasing and decreasing pitch period cases.

To visualize these distortions, the residual manipulation using the epoch-based method and the LP-PSOLA are illustrated for different pitch period modification factors. Fig. 11 shows the LP residual for a segment of voiced speech, and the residuals modified for pitch period modification factors of 2 and 0.5 using the LP-PSOLA method and the proposed epoch-based method. From the figure, we can see that, after modification, the general characteristics of the LP residual are preserved better in the epoch-based approach, compared to the results from the LP-PSOLA method. These changes in the residual are reflected in the results of the listening tests for the pitch period modification factors of 2, 0.5 and 0.4.

TABLE III
MEAN OPINION SCORES AND CONFIDENCE VALUES FOR DIFFERENT DURATION MODIFICATION FACTORS

Duration modification factor (β)	Mean opinion score (MOS)		Level of confidence in % for the significance of difference in MOSs
	LP-PSOLA method	Epoch-based method	
0.5	3.78	3.63	> 90
0.75	4.57	4.63	< 90
1.5	4.65	4.71	< 90
2	4.12	4.37	99.5
2.5	3.96	4.07	< 90

For evaluating the proposed method for duration modification, a similar approach was followed as in the case of pitch period modification. The mean opinion scores for different duration modification factors are given in Table III. The significance of differences in the pairs of MOSs is tested using hypothesis testing in this case also. In this case, the differences in the pairs of MOSs is not significant for most of the duration modification factors, as the percentage confidence was only about 90%. The scores show that both the methods seem to perform equally well for duration modification.

VIII. SUMMARY AND CONCLUSION

In this paper, we have proposed a flexible method for manipulating the prosody (pitch and duration) parameters of a speech utterance. The method uses the features of source of excitation of the vocal tract system. The linear prediction residual was used to represent the excitation information. The prosody manipulation was performed by extracting the instants of significant excitation (epochs) from the LP residual, and generating a new epoch sequence according to the desired prosody modification. A modified LP residual was generated using the knowledge of the new epoch sequence. In generating this residual, the perceptually significant portion (20% of the region around the instant of glottal closure) was retained, and the remaining 80% of the residual samples were used to generate the required number of samples in the modified residual. It is interesting to note that the epochs in both the voiced and nonvoiced regions are treated alike, thus, avoiding a separate voiced, unvoiced and silence (V/UV/S) decision making. Also, since the manipulation was performed on the residual signal, distortions were not perceived. This is because the residual samples are less correlated than the signal samples. This feature also helps in realizing prosody modification by large modification factors. The modification procedure is similar for both pitch period and for duration.

Since the prosody modification is done on the residual, the spectral features are not modified. Thus, there are no spectral distortions. But there will be some degradation in the naturalness of the synthesized speech when large pitch period modification factors are involved. This is because, in natural speech, low pitch periods also correspond to some extent increased formant frequencies, for example, for female and children voices. Thus, for large changes in the scale factors, it is essential to incorporate the corresponding changes in the formant locations. These changes can be incorporated by modifying the LPCs.

REFERENCES

[1] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar. 1992.

[2] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 453–467, Dec. 1990.

[3] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Commun.*, vol. 8, pp. 147–158, Jun. 1989.

[4] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, Feb. 1995.

[5] B. Yegnanarayana, S. Rajendran, V. R. Ramachandran, and A. S. M. Kumar, "Significance of knowledge sources for TTS system for Indian languages," in *Proc. SADHANA Academy Engineering Sciences*, vol. 19, Feb. 1994, pp. 147–169.

[6] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 374–390, Jun. 1981.

[7] M. R. Schroeder, J. L. Flanagan, and E. A. Lundry, "Bandwidth compression of speech by analytic-signal rooting," *Proc. IEEE*, vol. 55, no. 3, pp. 396–401, Mar. 1967.

[8] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 737–793, Sep. 1987.

[9] M. Narendranadh, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Commun.*, vol. 16, pp. 206–216, Feb. 1995.

[10] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant excitation from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 609–619, Nov. 1999.

[11] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[12] R. Crochiere, "A weighted overlap-add method of short time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 1, pp. 99–102, Feb. 1980.

[13] S. Roucos and A. Wilgus, "High quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, Mar. 1985, pp. 493–496.

[14] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Minneapolis, MN, Apr. 1993, pp. 550–553.

[15] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[16] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Munich, Germany, 1997, pp. 1303–1306.

[17] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[18] R. MuraliSankar, A. G. Ramakrishnan, A. K. Rohitprasad, and M. Anoop, "DCT-based pitch modification," in *Proc. SPCOM 6th Biennial Conf.*, Bangalore, India, Jul. 2001, pp. 114–117.

[19] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 1–11, Jan. 1998.

[20] S. Lemmetty, "Review of speech synthesis technology," M.S. thesis, Dept. Elect. Commun. Eng., Helsinki Univ. Technol., Espoo, Finland, 1999.

[21] R. Kortekaas and A. Kohlrausch, "Psychoacoustical evaluation of the pitch synchronous overlap-and-add speech waveform manipulation technique using single formant stimuli," *J. Acoustic Soc. Amer.*, vol. 101, no. 4, pp. 2202–2213, 1997.

[22] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA techniques," *Speech Commun.*, vol. 11, pp. 175–187, 1992.

[23] Y. Jiang and P. Murphy, "Production based pitch modification of voiced speech," in *Proc. Int. Conf. Spoken Language Processing*, Denver, CO, Sep. 2002, pp. 2073–2076.

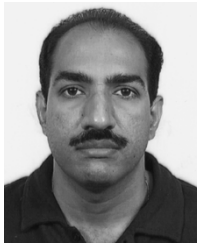
[24] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[25] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[26] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, Feb. 1986, pp. 93–99.

[27] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[28] R. V. Hogg and J. Ledolter, *Engineering Statistics*. New York: Macmillan, 1987.



K. Sreenivasa Rao was born in India, in 1969. He received the B.Tech. degree in electronics and communication engineering from Rayapati Venkata Rangarao (RVR) College of Engineering, Nagarjuna University, India, in 1990, the M.E. degree in communication systems from Peelamedu Sominayudu Govindasaminayudu (PSG) College of Technology, Bharathiar University, India, in 1993, and the Ph.D degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, India, in 2005.

He worked as a Project Officer in the Department of Computer Science and Engineering, Indian Institute of Technology Madras, from February 2005 to September 2005. Since October 2005, he has been an Assistant Professor in the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India. His research interests are speech signal processing and neural networks.

B. Yegnanarayana (SM'84) was born in India in 1944. He received the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978 in the Department of Electrical Communication Engineering, Indian Institute of Science. From 1978 to 1980, he was a Visiting Associate Professor of computer science at Carnegie Mellon University, Pittsburgh, PA. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India. He was the Chairman of the department from 1985 to 1989. His current research interests are in signal processing, speech, vision, neural networks, and man–machine interfaces. He has published papers in reviewed journals in these areas.

Dr. Yegnanarayana is a Fellow of the Indian National Science Academy, Indian National Academy of Engineering, and Indian Academy of Sciences. He is an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.