Duration modification using Glottal Closure Instants and Vowel Onset Points

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

Abstract

This paper proposes a method for duration (time scale) modification using Glottal Closure Instants (GCI, also known as instants of significant excitation) and Vowel Onset Points (VOP). In general, most of the time scale modification methods attempt to vary the duration of speech segments uniformly over all regions. But it is observed that consonant regions and transition regions between a consonant and the following vowel, and between two consonant regions do not vary appreciably with speaking rate. The proposed method implements the duration modification without changing the durations of the transition and consonant regions. Vowel onset points are used to identify the transition and consonant regions. A VOP is the instant at which the onset of the vowel takes place, which corresponds to the transition from a consonant to the following vowel in most cases. The VOPs are computed using the Hilbert envelope of Linear Prediction (LP) residual. 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 the onset of burst, in the case of nonvoiced speech. Manipulation of duration is achieved by modifying the duration of the LP residual with the help of instants of significant excitation as pitch markers. 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 found to be natural. Performance of the proposed method is compared with the method, where the duration of speech is modified uniformly over all regions. Samples of speech signals for different modification factors is available for listening at *http://sit.iitkgp.ernet.in/∼ksrao/result.html*

K. Sreenivasa Rao is with the School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India. E-mail: ksrao@iitkgp.ac.in

B. Yegnanarayana is with the International Institute of Information Technology (IIIT), Gachibowli, Hyderabad - 500032, Andhra Pradesh, India. Email: yegna@iiit.ac.in

Keywords

Instants of significant excitation, Group delay function, Hilbert envelope, Linear prediction residual, Vowel onset point, Time scale modification, Duration modification.

I. Introduction

THE purpose of Time Scale Modification (TSM) is to change the rate of speech while preserving the characteristics of the original speech such as formant structure, pitch periods, etc. There are various applications of time scale modification. For example, time scale compression can be used at the input side of the speech coder and transmission, followed by time scale expansion at the receiver to get back the original speech. In some other applications, time scale expansion can be used to enhance the intelligibility of rapid or degraded speech [1]. Time scale compression is also useful in message playback systems for fast scanning of recorded messages [2].

There are number of approaches available in the literature for time scale modification. Some of them use sinusoidal model, pitch synchronous overlap and add (PSOLA) and phase vocoders [3], [4], [5]. These methods produce some spectral and phase distortions, mainly due to manipulation of the speech signal directly. In [6], the authors have proposed an epoch-based time scale modification method, where the duration of speech signal is modified using the knowledge of epochs. Here, the epoch refers to the instant of glottal closure, where significant excitation takes place in each glottal cycle. The information around the instant of glottal closure is important to preserve the naturalness of speech. Therefore, the region around the instant of glottal closure is preserved during time scale modification using the knowledge of epochs. In addition, the TSM is performed in the residual domain to reduce the spectral and phase distortions. Linear prediction pitch synchronous overlap and add (LP-PSOLA) approach also performs TSM in the residual domain similar to the epoch-based method [7].

In the method proposed in [6], the TSM is performed by modifying the entire speech signal uniformly. In general, it is observed that all speech regions do not change uniformly with changes in the speech rate. The durations of vowel and pause regions of speech are affected more by the rate of speech production, whereas the durations of consonant and transition regions do not change much [8], [9]. The objectives of this paper are to demonstrate the need for nonuniform duration modification of different regions for synthesizing speech at different speaking rates, and to describe a method to realize the nonuniform duration modification using epochs and vowel onset points (VOPs).

Attempts to incorporate nonuniform duration modification are reported in the literature [4], [10], [11]. Quatieri *et al* have developed speech adaptive TSM method, based on voicing probability derived from sinusoidal pitch estimator [4]. The voicing probability is close to unity during steady voicing, decreases during transition, and close to zero during unvoiced speech and pauses. The assumption is that changes in speaking rate compression or expansion do not take place in sounds which are not voiced, but they occur mostly in voiced sounds. Malcolm Slaney *et al* have used nonuniform time scaling along with spectral shape and pitch modification for automatically morphing one sound to another sound [10]. Olovia Donnellan *et al* have proposed a method for speech adaptive TSM, which allows slowing down speech without compromising the quality or naturalness of the slowed speech [11]. In their method, different scaling factors are applied to different types of speech segments.

In this paper a new method based on epochs and VOPs is proposed for nonuniform TSM. In this method the TSM is affected only in the vowel and pause regions, keeping the durations of consonant and transition regions unaltered. The different regions are identified using VOP. The TSM is performed in the residual domain. Successive samples in the linear prediction (LP) residual are less correlated compared to the samples in the speech signal. Therefore, residual manipulation to achieve the desired TSM will introduce less perceptual distortion [6]. The proposed method modifies the LP residual using epochs for preserving the naturalness.

The paper is organized as follows: In order to demonstrate the need for nonuniform TSM, in Section II the effect of speaking rate on durations of different regions of a speech utterance is examined. Detection of VOP using the Hilbert envelope of the linear prediction residual is discussed in Section III. The proposed method for duration modification is presented in Section IV. Evaluation of the proposed method using listening tests is discussed in Section V. The paper concludes with a summary and possible extensions for future work.

II. Motivation for the proposed approach

For large time scale expansion with modification factors greater than 2, the transition and consonant regions sound unnatural. Likewise in the time scale compression, with modification factors less than 0.5, some of the consonants are not perceived at all. Naturally spoken utterances at different speaking rates (fast, normal and slow) are analyzed to examine the effect of the speaking rate on the durations of different segments of speech. The analysis was performed as follows: Five sentences of Hindi were chosen for this study. These sentences were uttered by 10 cooperative subjects in a laboratory environment. Each sentence was uttered five times by each subject: (1) At normal speaking rate, (2) at slow rate, (3) at very slow rate, (4) at fast rate, and (5) at very fast rate. Here slow rate and fast rate correspond approximately to the modification factors of 1.5 and 0.75, respectively. Altogether 250 sentences were recorded using 5 different sentences uttered in five different speaking rates by 10 subjects. The sentences were analyzed manually in the transition regions, vowel regions, consonant regions and pauses. In most of the cases, the durations of the vowels and pauses are affected significantly by the speaking rate variations, whereas the durations of the consonant and transition regions remain mostly unaltered. The details of the variation in the durations of different regions for different speech rates are given in Table I. In this work, the variation in duration is measured using % deviation (D_i), given by

$$D_i = \frac{x_i - y_i}{x_i} \times 100,$$

where x_i and y_i are the durations of the speech regions under normal and varying speaking rates, respectively. The durations of the consonant, transition, vowel and pause regions of the normal speaking rate are used as reference. For each of the consonant, transition, vowel and pause regions, the deviations are computed under different speech rates. The numbers in the Table I gives the average % deviation in the durations of the regions under different speaking rates. The numbers with positive and negative signs in Table I indicate % of expansion and compression, respectively, with respect to the reference duration. From this study, the following observations can be made: (1) Durations of vowels and pauses are affected significantly by the changes in the speaking rate. (2) The durations of the consonant and transition regions do not change much with speaking rate. (3) The variation in the duration of the vowel segment depends on the category of the vowel and its preceding consonant. (4) The variation in the duration of the pause region depends on the length of the adjacent words (# syllables in a word) and the number of words present in the utterance. These observations clearly demonstrate the need for nonuniform TSM. To apply nonuniform TSM, it is necessary to identify the different regions in speech. The vowel onset point is used as an anchor point to determine the consonant, transition and vowel regions, as explained in the next section. VOP together with glottal closure instants are used to implement the nonuniform TSM as described in Section IV.

TABLE I

Percentage deviation in the durations of speech regions under different speaking rates.

Speech	% Deviation from reference duration			
region	Slow	Very slow	Fast	Very fast
Vowel	+73	+137	-34	-79
Pause	+58	+103	-46	-93
Consonant	+3	+4	-1	-2
Transition	+5	+7	-3	-4

III. Detection of Vowel Onset Points

Vowel onset points are detected using the Hilbert envelope of the LP residual [12]. The speech signal is sampled at 8 kHz and preemphasized before performing LP analysis. The LP residual is computed using 10^{th} order LP analysis, with a frame size of 20 ms and a frame shift of 5 ms. The Hilbert envelope of the LP residual is then computed. The VOP evidence is obtained from the Hilbert envelope of the LP residual by convolving it with a Gabor filter. A Gabor filter with spatial spread of the Gaussian $\sigma = 100$, the frequency of modulating sinusoid $\omega = 0.0114$, and a filter length $n = 800$, is considered [13]. The Gabor filter used in this study is shown in Fig. 1. In the VOP evidence plot the peaks are located using a peak picking algorithm. Spurious peaks are eliminated using the characteristics of the shape of the VOP evidence plot, namely, between two true VOPs there exists a negative region of sufficient strength due to vowel region. For each peak, a check for the presence of such a negative region with respect to the next peak is made to eliminate spurious peaks [12].

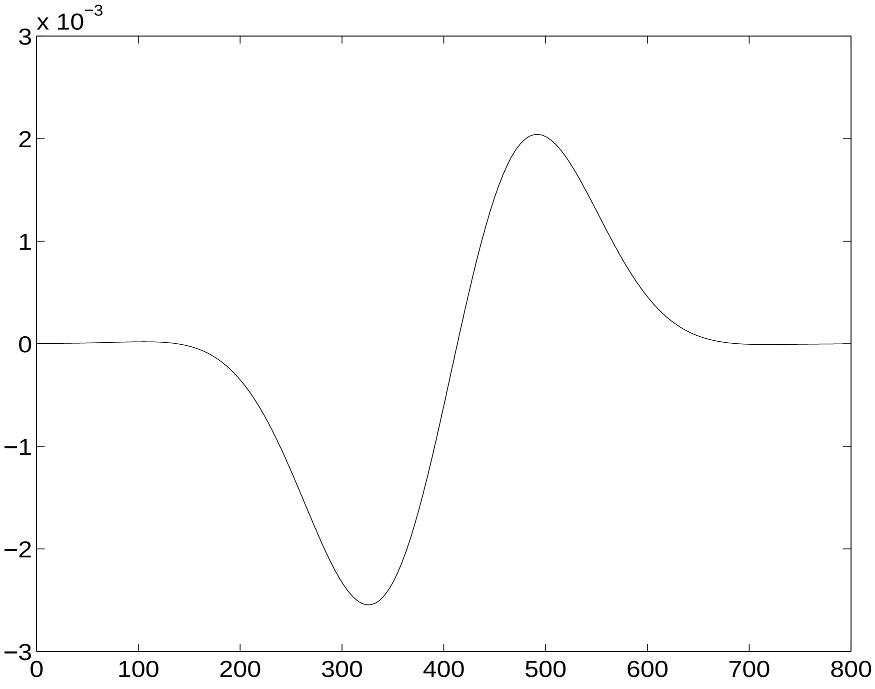

Fig. 1. Gabor window for $\sigma = 100$, $\omega = 0.0114$ and $n = 800$.

The above procedure is illustrated for the Hindi sentence "*antarAshtriyA bassevA pichale mahIne shuru huyi thi*". In this sentence there are 16 VOP events, as marked (manually) in Fig. 2(a). The Hilbert envelope and the VOP evidence plots are shown in Figs. 2(b) and (c), respectively. The output of the peak picking algorithm is given in Fig. 2(d). The hypothesized VOPs after eliminating the spurious ones are shown in Fig. 2(e). The procedure for detecting the VOPs in speech signals is summarized in Table II. The method ensures the VOP detection accuracy of about 88% [12]. The method may miss some of the VOPs, but it does not generate spurious VOPs.

We can interpret the vowel onset point as the junction point between the consonant and vowel of a CV unit. At this point the characteristics of the consonant segment will be terminated, and the characteristics of the vowel segment will be originated. Hence, it is important to locate this point (i.e., VOP) accurately, for performing the nonuniform TSM. The detailed explanation about the significance of VOP for speech analysis

TABLE II

Steps for detection of VOP events.

1. Preemphasize input speech.
2. Compute LP residual with 10^{th} order LP analysis, with a frame size of 20 ms and shift of 5 ms.
3. Compute Hilbert envelope of the LP residual.
4. Obtain the VOP event evidence plot from Hilbert envelope by convolving it with the Gabor filter.
5. Identify the peaks in the VOP event evidence plot using peak picking algorithm.
6. For each peak, if there is no negative region with reference to next peak, then eliminate such a peak as spurious.
7. Eliminate peaks which are at a distance less than 50 ms with respect to their next peak.
8. Hypothesize remaining peaks as the VOP events.

and its detection methods is given in [14], [12]. In Indian languages most of the characters are of the type CV or CCV (where C refers to consonant and V refers to vowel). The region to the left of the VOP is considered as the consonant region, and to the right of the VOP as the vowel region. In the vowel portion, a small region following the VOP is treated as transition region [15]. After determining the vowel onset point, 30 ms to the left of the VOP is marked as the consonant region, and 30 ms to the right of the VOP is marked as the transition region. In the proposed modification method, the durations of these regions are fixed.

IV. Proposed method for duration modification

The time-varying vocal tract system parameters (i.e., Linear Prediction Coefficients (LPCs)) and the corresponding LP residual signal are derived from the speech signal by LP analysis [16]. The instants of significant excitation (GCIs) are computed from the LP residual using group delay analysis [17]. The instants of significant excitation can be derived from the LP residual signal as follows: Around each sample a 10 ms segment of the LP residual signal is considered, and the group delay function $\tau(\omega)$ is computed using [18]

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

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ are the Fourier transforms of the windowed residual signal $x(n)$ and $nx(n)$, respectively. The group delay function is smoothed using a 3-point median

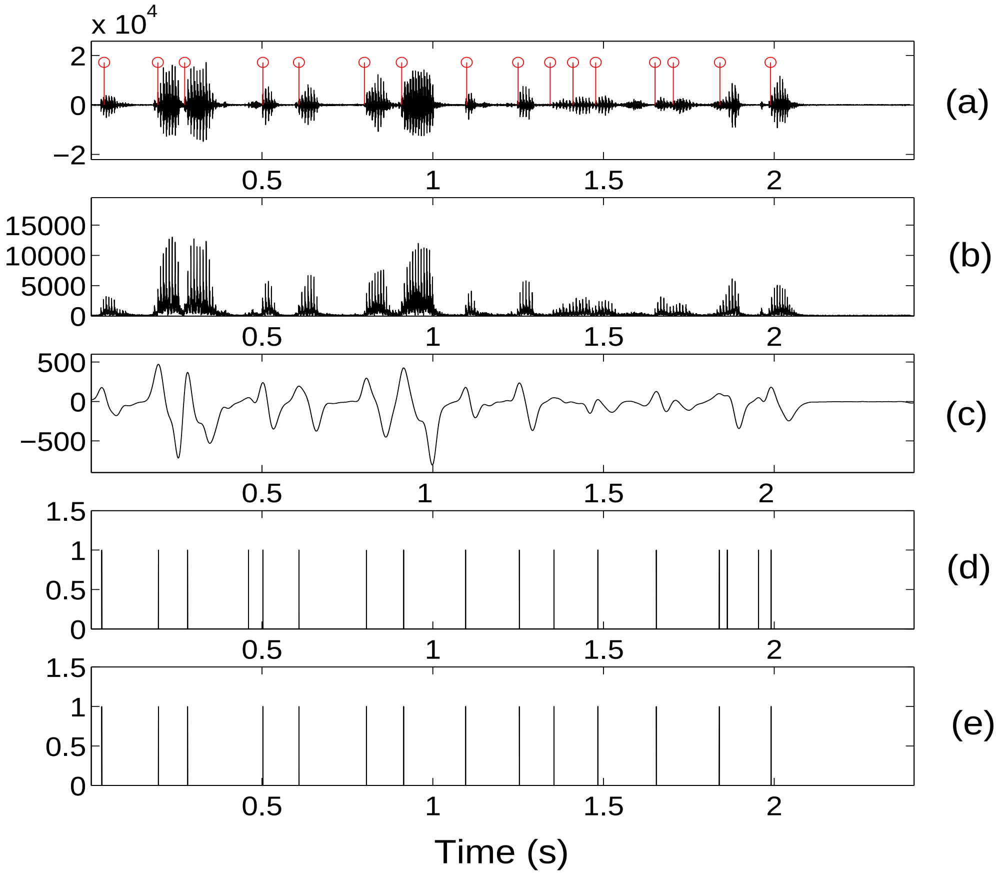

Fig. 2. For the Hindi sentence *antarAshtriyA bassevA pichale mahIne shuru huyi thi:* (a) Waveform with manually marked VOP events, (b) Hilbert envelope of LP residual, (c) VOP evidences, (d) Output of peak picking algorithm, and (e) Hypothesized VOP events after eliminating some spurious peaks.

filter to remove any discontinuities in the group delay function. The negative of the average of the smoothed group delay function is called *phase slope* [17]. The phase slope value is computed at each sampling instant to obtain the *phase slope function*. If the instant of significant excitation within a frame is at the midpoint of the frame, then the phase slope is zero. Therefore the positive zero-crossings of the phase slope function correspond to the instants of significant excitation. Fig. 3 shows a segment of voiced speech, the LP residual, the phase slope function and the instants of significant excitation.

There are four main steps involved in the epoch-based time scale modification method (uniform duration modification) [6]: (1) Deriving the instants of significant excitation (epochs) from the LP residual signal. (2) Deriving a modified (new) epoch sequence according to the desired duration modification factor. (3) Deriving a modified LP residual signal from the modified epoch sequence. (4) Synthesizing speech using the modified

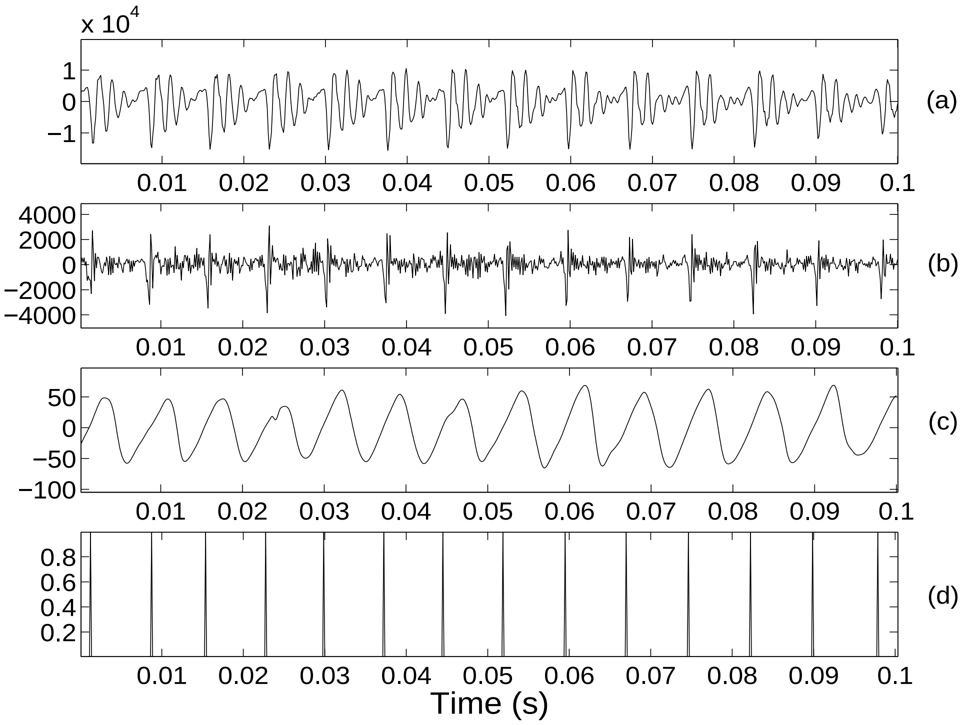

Fig. 3. (a) A segment of voiced speech, (b) LP residual, (c) Phase slope function and (d) Instants of significant excitation (GCIs).

LP residual and the LPCs. It involves deriving a new excitation (LP residual) signal by incorporating the desired modification in the duration of the utterance. This is done by first creating a new sequence of epochs from the original sequence of epochs. Each epoch is associated with the time, pitch period, LP residual and LPCs. The new epoch sequence consists of either insertion of new epochs for time scale expansion, or deletion of some epochs for time scale compression. The residual is accessed from the original epochs, and it is modified according to the new epoch sequence. To increase the duration, some portions of the residual are replicated at specific locations. Similarly for reducing the duration, some portions of the residual are to be omitted at specific locations.

For preserving the consonant and transition regions of speech from duration modification, vowel onset points are used as anchor points to identify those regions. Vowel onset points are determined for a given speech utterance using the procedure discussed in Section III. These VOPs are known as original VOPs (i.e., VOPs corresponding to speech utterance for normal speaking rate). After applying the TSM uniformly over all regions for the desired speaking rate, the VOPs corresponding to the desired speaking rate are obtained by scaling the locations of the original VOPs by the desired modification factor. These VOPs are termed as new VOPs (i.e., VOPs corresponding to the speech utterance for the desired speaking rate). The scaled consonant and transition regions in the modified residual are identified using the new VOPs. The region to the left of the new VOP is marked as the scaled consonant region, and to the right is marked as the scaled transition region. In the proposed nonuniform TSM, to retain the original consonant and transition regions of speech utterance, the scaled consonant and transition regions of the modified LP residual are replaced with their corresponding original LP residual segments. The original consonant and transition regions of the LP residual are determined by using the original VOPs. The filter coefficients (LPCs) are updated depending on the modified LP residual. Speech for the desired duration modification can be synthesized by exciting the all-pole filter using the modified LP residual. The proposed nonuniform TSM method is summarized in Table III.

TABLE III

Steps for nonuniform TSM using glottal closure instants and VOPs.

1. Preemphasize input speech.
2. Compute LP residual with 10^{th} order LP analysis, with a frame size of 20 ms and shift of 5 ms.
3. Derive the epochs from LP residual using group delay analysis.
4. Determine the VOPs from the Hilbert envelope of the LP residual, and mark the consonant and transition regions of the speech utterance corresponding to normal speaking rate.
5. Derive the new epoch sequence according to the desired speaking rate.
6. Determine the nearest original epoch corresponding to each of the new epochs.
7. Modify the LP residual according to new epoch sequence.
8. Determine the new VOPs, and mark the scaled consonant and transition regions in the modified LP residual.
9. Replace the scaled consonant and transition regions of the modified LP residual (marked in step-8) with their corresponding original segments of the LP residual (marked in step-4).
10. Update the filter coefficients according to the modified residual.
11. Synthesize the speech by exciting the updated filter with the modified LP residual.

V. Evaluation of the proposed duration modification method

Performance of the proposed duration modification method is compared with the epoch-based duration modification method using perceptual evaluation. In the epoch-based duration modification method, duration of speech is modified uniformly over all regions, whereas in the proposed method the durations of consonant and transition speech regions are unaltered during TSM for different speech rates. Perceptual evaluation was carried out by conducting subjective tests with 25 research scholars in the age group of 21-35 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. Two sentences were chosen to perform the test. Speech signals were derived for the modification factors from 0.25 to 3 in steps of 0.25. Each of the subjects was given a pilot test about perception of speech signals for different speaking rates (slow, normal and fast). 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 perceptual distortion and quality of the speech for various modification factors. Subjects were asked to assess the quality and perceptual distortion on a 5-point scale for each of the sentences obtained by both the methods. The 5-point scale for representing the quality of speech and the distortion level is given in Table IV [19].

TABLE IV

Ranking used for judging the quality and perceptual distortion of the speech signal modified by different modification factors.

Rating	Speech quality	Level of perceptual 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

The Mean Opinion Scores (MOSs) for each of the modification factors are given in Table V. The level of confidence is computed for the difference of each pair of MOSs [20]. Results of perceptual evaluation show that for lower modification rates both the methods produce good quality speech. When the modification factors are say around 0.5 for compression and 2 for expansion, the epoch-based method produces perceptual distortion, whereas the proposed method gives intelligible and better quality speech. A sample of the speech signals used for subjective tests is available for listening at *http://sit.iitkgp.ernet.in/∼ksrao/result.html*

TABLE V

Mean opinion scores and confidence values for different modification factors.

Duration modification factor	Mean opinion score		Level of confidence in % for the significance of difference in MOSs
	Epoch-based method	Proposed method	
0.25	2.32	3.27	> 99.5
0.50	3.41	3.93	> 99.5
0.75	4.37	4.49	> 95
1.25	4.67	4.72	> 90
1.50	4.62	4.70	> 90
1.75	4.46	4.62	> 95
2.00	4.21	4.43	> 97.5
2.25	4.03	4.27	> 99
2.50	3.68	4.02	> 99.5
2.75	3.32	3.76	> 99.5
3.00	3.07	3.57	> 99.5

VI. Summary and Conclusions

A new method is proposed for time scale modification to produce high quality synthesized speech. The method is based on processing the LP residual using the knowledge of the instants of significant excitation for maintaining the original pitch information, and the VOPs for identifying the transition and consonant regions. The method provides flexibility in time scale manipulation over large range of values of the modification factor. The effectiveness of the proposed method depends mainly on the accuracy in detecting the glottal closure instants and the locations of the vowel onset points, because the residual manipulation is performed using the epochs and VOPs as anchor points. Subjective tests indicate that the performance of the proposed method is superior compared to the epoch-based time scale modification method. By exploiting the duration characteristics of various speech units with respect to their linguistic context and production constraints at different speaking rates, further improvement may be achieved in the quality of synthesized speech.

References

[1] H. G. Ilk and S. Guler, "Adaptive time scale modification of speech for graceful degrading voice quality in congested networks for voip applications," *Signal Processing*, vol. 86, pp. 127–139, 2006.

[2] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, pp. 374–390, June. 1981.

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

[4] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Processing*, vol. 40, pp. 497–510, Mar. 1992.

[5] J. di Marino and Y. Laprie, "Supression of phasiness for time-scale modifications of speech signals based on a shape invarience property," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Saltlake city, Utah, USA), May 2001.

[6] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Speech and Audio Processing*, vol. 14, pp. 972–980, May 2006.

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

[8] J. L. Flanagan, *Speech analysis synthesis and prception.* Berlin: Springer-verlag, 1972.

[9] K. N. Stevens, *Acoustic Phonetics.* Cambridge, MA, USA: MIT Press, 1999.

[10] M. Slaney, M. Covell, and B. Lassiter, "Automatic audio morphing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Atlanta, GA, USA), May 1996.

[11] O. Donnellan, E. Jung, and E. Coyle, "Speech-adaptive time-scale modification for computer assisted language-learning," in *Proceedings of The 3rd IEEE International Conference on Advanced Learning Technologies (ICALT03)*, (Aix-en-Provence, France), 2003.

[12] S. R. M. Prasanna and J. M. Zachariah, "Detection of vowel onset point in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Orlando, Florida, USA), May 2002.

[13] D. Gabor, "Theory of communication," *J. IEE*, vol. 93, no. 2, pp. 429–457, 1946.

[14] S. R. M. Prasanna, *Event-Based Analysis of Speech.* PhD thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, Mar. 2004.

[15] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Extraction of fixed dimension patterns from varying duration segments of consonant-vowel utterances," in *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing*, (Chennai, India), pp. 159–164, Jan. 2004.

[16] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[17] R. Smits and B. Yegnanarayana, “Determination of instants of significant excitation in speech using group delay function,” *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 325–333, Sept. 1995.

[18] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Englewood Cliffs, New Jersey, USA: Prentice Hall, 1975.

[19] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals.* New York, USA: Macmilan Publishing Company, 1993.

[20] R. V. Hogg and J. Ledolter, *Engineering Statistics.* 866 Third Avenue, New York, USA: Macmillan Publishing Company, 1987.