Significance of Epoch Identification Accuracy for Prosody Modification

Nagaraj Adiga[1], Govind D[2] and S. R. Mahadeva Prasanna[1]
[1]Department Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Assam, India
[2]Center for Computational Engineering and Networking, Amrita Vishwa Vidyapeetham, Coimbatore, India
Email:nagaraj@iitg.ernet.in[1],govinddmenon@gmail.com[2],prasanna@iitg.ernet.in[1]

***Abstract*—Epoch refers to instant of significant excitation in speech [1]. Prosody modification is the process of manipulating the pitch and duration of speech by fixed or dynamic modification factors. In epoch based prosody modification, the prosodic features of the speech signal are modified by anchoring around the epochs location in speech. The objective of the present work is to demonstrate the significance of epoch identification accuracy for prosody modification. Epoch identification accuracy is defined as standard deviation of identification timing error between estimated epochs with the reference epochs. Initially, the epochs location of the original speech are randomly varied for arbitrary time factors and corresponding prosody modified speech is generated. The perceptual quality of the prosody modified speech is evaluated from the mean opinion scores (MOS) and objective measure. The issues in the prosody modification of telephonic speech signals are also presented.**

Index Terms: speech prosody, epoch, epoch identification accuracy, pitch, duration, telephone speech

I. Introduction

Prosody modification is the process of manipulating pitch and duration of speech without introducing spectral and temporal distortions [2]. Prosodic features of speech are the suprasegmental features which span over longer segment of speech [3], [4]. Prosody modification finds important applications such as in neutral to emotion conversion, voice conversion, text to speech synthesis, etc. [4]–[7]. For instance, in neutral to emotion speech conversion, the prosodic features of the neutral speech signals are modified to synthesize speech in the target emotion. Voice conversion is achieved by modifying the prosody and vocal tract parameters of source speaker to the target speaker [6], [8]. In unit selection speech synthesis, to improve the naturalness of synthesized speech, prosody modification is performed [7].

There are different techniques proposed in the literature for prosody modification [3], [9]–[11]. The prosody modification techniques are broadly classified into time domain and frequency domain approaches [12]. Among the time domain approaches, the pitch synchronous overlap (PSOLA) methods such as time domain PSOLA (TD-PSOLA), linear prediction PSOLA (LP-PSOLA), and epoch based approaches are popular. In frequency domain, FD-PSOLA is a well known approach predominantly used for pitch modification [9]. The Discrete cosine transform based pitch modification is also proposed in [13]. All these existing methods provide better perceptual quality in the prosody modified speech, if the analysis pitch marks for prosody modification are estimated accurately from original speech [10]. In epoch based prosody modification, the accuracies of the analysis pitch marks are ensured by estimating epochs with reduced identification timing errors [3], [10].

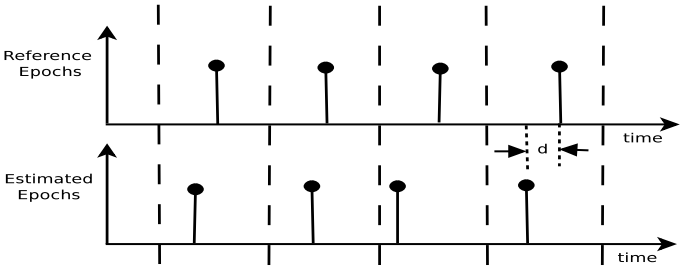

Fig. 1. Epoch location estimation showing 4 larynx cycles and its comparison with reference epochs obtained from electro-glottograph (EGG), 'd' represents the identification timing error from actual epoch location to the reference epoch

The epoch based prosody modification is performed in three steps [3], [10]. In the first step, the accurate locations of the epochs are estimated from the speech signal. Since the method based on zero frequency filtering (ZFF) of speech provides accurate estimation of epochs location, the epoch based prosody modification employs ZFF for estimating epochs [14]. Apart from the ZFF based epoch extraction, recent epoch extraction methods like DPI (dynamic plosion index of integrated LP residual) and SEDREAMS (speech event detection using residual excitation and mean based signal) are also used in the epoch based prosody modification [15] [16]. These estimated epochs location are considered as the analysis pitch marks for the prosody modification. The modified epochs location are derived according to the prosody modification scaling factors (in the second step). These modified epochs location for the given prosodic modification factors are considered as the synthesis pitch marks for the prosody modification. In the final step, the speech waveforms are reconstructed by copying residual samples of the original signal starting from the analysis epoch to the modified epochs. Finally prosody modified speech waveform is reconstructed from modified LP residual according to the prosody modification factors.

978-1-4799-4665-5/14/$31.00 ©2014 IEEE

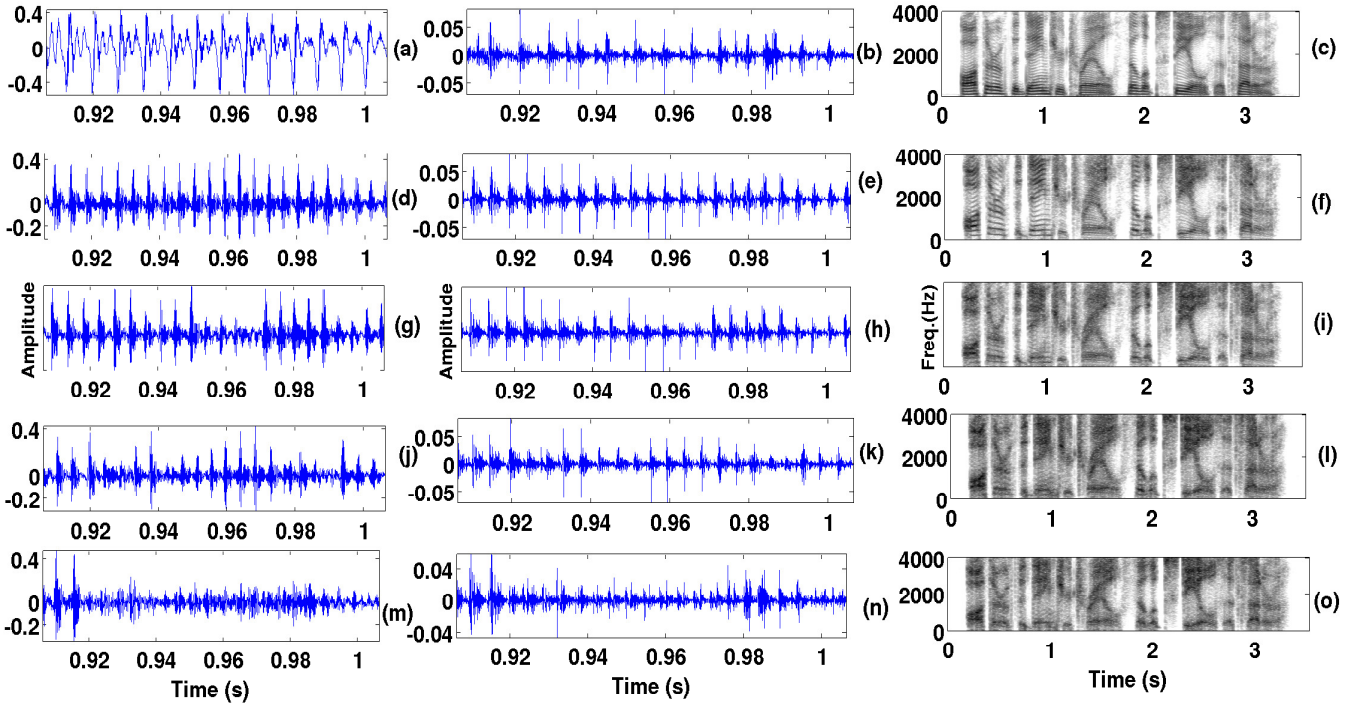

Fig. 2. Pitch modification for a pitch period scaling factor of 0.6 for randomly varied the epochs location by arbitrary time samples. (a) A voiced speech segment (b) corresponding LP residual and (c) spectrogram of original speech. ((d)-(f)) plot the corresponding segments of voiced speech, modified LP residual and spectrograms of pitch modified speech using pitch period scaling factor of 0.6. Fig ((g)-(i)), ((j)-(l)) and ((m)-(o)) show the same segments of pitch modified speech obtained by randomly varying the epochs location by 1 ms, 2 ms and 3 ms, respectively. The spectrograms are plotted for the whole utterance in each case.

The objective of the present work is to demonstrate the effect of epoch identification accuracy on the perceptual quality of the prosody modified speech. Epoch identification accuracy(σ) is defined as standard deviation of identification timing error 'd' between estimated epochs with the reference epochs obtained from elctro-glottograph (EGG) [17]. Fig 1 shows epoch location estimated for 4 larynx cycles and its comparison with reference epochs obtained from EGG, 'd' represents the identification timing error from actual epoch location to the reference epoch. The effect of epoch identification accuracy is demonstrated by performing the prosody modification by preserving original epochs sequence and by randomly varying the epoch location within the pitch period of the original speech. In the first case, epochs location of the original speech are kept intact and epochs location are altered by random time factors in the second case for the prosody modification. The present work is motivated by our previous work on the significance of preserving perceptually relevant samples around epochs for source modeling [18]. Finally, to discuss the issue of epoch identification accuracy in practical scenario, epoch based prosody modification is performed on telephonic speech signals and compared with clean speech case. Due to poor epoch identification accuracy in telephone speech, perceptual quality of prosody modification is found to be in the lower side.

The present work is organized as follows: Section II discusses the procedure for checking the effect of epoch identification accuracy for prosody modification. Section III gives a detailed description of the subjective and objective evaluations. The issues in the epoch based prosody modification of telephonic speech signals are discussed in Section IV. Finally, Section V summarizes the work with scope for future work.

II. Procedure for finding the effect of Epoch Identification Accuracy

In epoch based prosody modification, prosody modification is performed on LP residual. The LP residual of the given original speech is modified according to pitch and duration scaling factors. Initially, the epochs location are extracted from speech using ZFF method. The epoch intervals are computed by finding the difference between successive epochs location. Based on the pitch or duration modification factors, the modified epochs location are derived. The modified LP residual sequence is then constructed by copying 40% of residual samples in the epoch interval to the modified epochs location and remaining 60% residual samples present in the epoch interval are resampled according to the modified epoch interval obtained from modified pitch or duration factor. Finally, the prosody modified speech is reconstructed by exciting the

original LP coefficients with the prosody modified sequence of LP residual. Significance of epoch identification accuracy is verified by randomly varying the original epochs location and synthesizing the prosody modified speech for different pitch or duration factors corresponding to modified epochs location.

Fig 2 plots longer segments of speech, LP residual, modified epochs location and spectrograms of original ((a)-(d)), pitch modified speech using the original epochs location ((e)-(h)) and pitch modified speech obtained by randomly varying epochs location by 1 ms ((g)-(i)), 2 ms ((j)-(l)) and 3 ms ((m)-(o)), respectively. In the figure speech and residual signals are plotted for around 10 ms region, and spectrogram plot is taken for entire speech utterance, as we can not take the spectrogram for short 10 ms frame. The sets of plots ((g)-(i),(j)-(l)and (m)-(o)) are generated by randomly varying original epochs location by 1 ms, 2 ms and 3 ms respectively, before deriving the modified epochs location for pitch modification. Fig 2(f) shows the spectrogram of the pitch modified speech for minimum spectral and temporal distortions. The evidence of the pitch modification can be seen as the modified pitch and harmonics as compared to the spectrogram of the original speech shown in Fig 2(c). An increase in the levels of spectral distortions can be seen as the random variations in epochs location increase in steps of 1 ms, from the plots of Fig 2((i),(l) and (o)), for example around 1 s region of Fig 2((i),(l) and (o)) smearing of pitch and harmonic information can be observed in the spectrogram. Detailed analysis of spectral distortion is reported in the Section III-B. The temporal distortions can be observed as the random variations in the amplitude envelops of pitch modified residual signals in each case Fig 2((h),(k) and (n). The temporal distortions can also be observed in the waveform samples of the pitch modified speech in each case Fig 2((g),(j) and (m)). The same observations are valid for epochs based duration modification also. Hence, from the visual comparisons, it can be concluded that the epochs identification accuracy plays an important role in the quality of the prosody modified speech.

III. Quality evaluation of Prosody modified speech

The effect of epoch identification accuracy on prosody modified speech is evaluated by both subjective and objective evaluation. In subjective evaluation mean opinion score (MOS) by subjects is taken, and in objective evaluation cepstral distance is used to find the perceptual quality of prosody modified speech. Details of the evaluation and results are given in the following.

A. Subjective evaluation

In the subjective evaluation, 6 different sentences are selected from CMU ARCTIC database [19] for 3 (2 male and 1 female) speakers with sampling frequency of 32 kHz. Prosody modification is performed on each speech file by modifying both pitch and duration for arbitrary prosody modification factors. The pitch and duration modified speech files for the evaluation are generated by randomly varying epochs location to 1 ms, 2 ms and 3 ms from the original epochs location. The pitch period scaling factor of 0.7 and duration scaling factor of 1.3 are used as the prosody modification factors for subjective evaluations. Fixed scaling factor is used here just for evaluation purpose and these results are consistent for any other scaling factor.

TABLE I
Ranking used for judging the quality and distortion of the speech signal for different modification factors.

Rating	Speech Quality	Justification for the ranking
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 speech files were randomized and file names were coded before presenting to the subjects for the evaluation. 17 subjects participated in the subjective evaluation and were asked to observe the perceptual distortions present in each file and give there opinion scores accordingly on a five point scale. The justification for each scale is given in Table I. A total of 72 (6*3*3+ 18 original prosody modified speech) files are used for the subjective evaluation of both pitch and duration modification. The mean of the scores obtained for all the files is calculated as the MOS. The MOS obtained is given in Table II. It has to be observed from the Table II that by varying the epochs location for different time factors, MOS of the prosody modified speech is also decreasing. For instance, a significant drop in MOS 3.78 to 3.01 is observed when the epochs are randomly varied by 1 ms from the original locations in case of pitch modification which further reduced to a MOS of 2.45 for a 3 ms deviations from the original epochs location. This degradation in the MOS indicates the dependency of the perceptual quality of the prosody modified speech on epoch identification accuracy. The consistent drop in the perceptual quality is also observed for the duration modification also. The synthesized prosody modified files for different epoch deviation factors are available for listening in the following link: http://www.iitg.ernet.in/cseweb/tts/Assamese/epochaccuracy.php

TABLE II
Mean opinion scores evaluated for synthesized speech and prosody modified speech for different epoch accuracy cases

Epoch location variation	**MOS of Prosody modified speech**	
	Pitch Modification	**Duration Modification**
0 ms (No variation)	3.78	3.32
1 ms	3.01	3.10
2 ms	2.72	2.75
3 ms	2.45	2.52
Telephone speech	1.85	1.83

B. Objective evaluation

Besides the subjective evaluation, spectral distortion of the prosody modified speech for different epoch identification

accuracy is also calculated using objective measure. Cepstral distance is used as objective measure here [20], which is calculated between prosody modified speech obtained from original epoch location and altered epoch location for different epoch identification accuracy. Cepstral distance is the Euclidean distance between cepstral coefficients of test frame with those of reference frame. After calculating cepstral distance for each frame, average cepstral distance for entire speech utterance is determined. In this evaluation, 50 randomly selected sentences from all 3 speakers of CMU ARCTIC database [19] are used for objective measure. First 13 cepstral coefficients of cepstrum obtained from each speech frame is used here for distance calculation and average cepstral distance is calculated for all the sentences. Average cepstral distance for both pitch modified and duration modified cases is given in Table III for different epoch accuracy cases. We can observe from the Table III that cepstral distance is increased from 0.73 to 0.94 for varying epoch location randomly from 1 ms to 3 ms in case of pitch modified speech, which indicates that spectral distortion is introduced for drop in epoch identification accuracy. Similar observations is also happened in the duration modified speech, which signifies the importance of epoch identification accuracy for prosody modification.

TABLE III
AVERAGE CEPSTRAL DISTANCES CALCULATED BETWEEN ORIGINAL PROSODY MODIFIED SPEECH AND PROSODY MODIFIED SPEECH WITH DIFFERENT EPOCH ACCURACIES

Epoch location variation clean speech	**Average cepstral distance of Prosody modified speech**	
	Pitch Modification	**Duration Modification**
1 ms	0.73	0.58
2 ms	0.85	0.71
3 ms	0.94	0.78
Telephone speech		
DYPSA	1.35	1.44
ZFF	1.47	1.49
SEDREAMS	1.46	1.42
DPI	1.24	1.35

IV. EPOCH IDENTIFICATION ACCURACY IN TELEPHONIC SPEECH QUALITY

Telephonic speech signals are band-limited signals with bandwidth ranging from 300 Hz to 3.4 kHz as compared to speech signals collected from a studio. Apart from the band limited nature of the signal, the channel also introduces degradations to the speech signal. Hence the algorithms that give the best results with clean speech show performance degradations when operated on the telephonic speech signals. In order to study the perceptual quality of the prosody modification in telephonic speech, first the epoch identification accuracy is tested and compared with that of the clean speech signals, and then followed by prosody modification in both telephonic speech and studio quality clean speech signals.

For the epochs estimation performance analysis, the phonetically balanced CMU ARCTIC database having simultaneous EGG recording, is used. There are a total of 1132 phonetically balanced utterances present for each speaker. The reference epochs for computing epoch identification accuracy are estimated from the available EGG signals for each utterance. To compute the epoch identification accuracy for telephonic speech signals, the telephonic version of CMU ARCTIC database is generated. The telephonic speech signals are generated by converting the clean speech utterances in CMU ARCTIC database using the "G.191- Software tools for speech coding standardization", freely available from ITU software library [21]. We followed exactly the same procedure to simulate the telephonic channel as given in [21].

TABLE IV
COMPARISON OF EPOCH IDENTIFICATION ACCURACY OBTAINED FROM CLEAN SPEECH AND TELEPHONIC SPEECH. THE EPOCH IDENTIFICATION ACCURACY IS ANALYZED FOR 500616 REFERENCE EPOCHS.

Method	Epoch Identification Accuracy, σ (ms	
	Clean Speech	Tel. Speech
$DYPSA$	0.36	1.71
ZFF	0.29	2.1
$SEDREAMS$	0.32	1.78
DPI	0.26	1.25

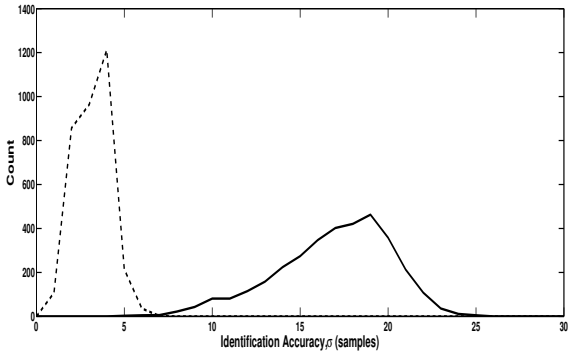

Fig. 3. Comparison of epoch identification accuracies obtained by the ZFF of clean speech and simulated telephonic speech arctic utterances. The dotted curve in plot indicate the distribution of epoch identification accuracies obtained for clean speech and continuous curve in plot indicates the corresponding epoch identification accuracy distribution for telephonic speech.

As ZFF based epoch extraction is one of the state of the art methods, the accurate locations of the epochs are estimated using the ZFF method. Table IV shows the epochs identification accuracies obtained for the clean speech and telephonic speech signals. Also epoch identification accuracies obtained from ZFF method are compared with other state-art-of algorithms like DYPSA, SEDREAMS and DPI methods [15]–[17]. The details of the ZFF, DYPSA, SEDREMS and DPI algorithms can be found in the papers by Murty et al. , Naylor et al., Drugman et al., and Prathosh et al. respectively [14]–[17]. A significant difference between the epoch estimation accuracies can be found for clean speech and simulated telephonic speech from Table IV in case of ZFF method. The same trend can also be observed for DYPSA, SEDREAMS and DPI methods too. The degradation

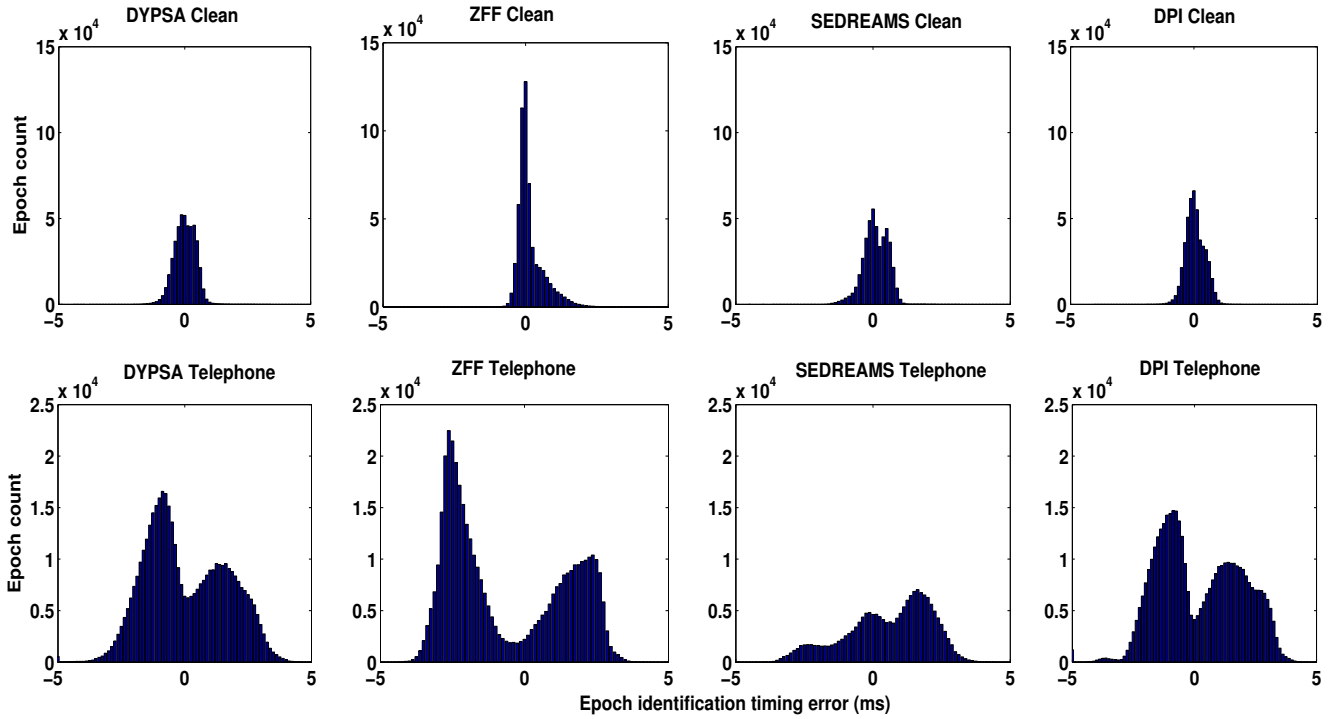

Fig. 4. Histogram of epoch identification timing error obtained by the DYPSA, ZFF, SEDREAMS and DPI methods for both clean and simulated telephonic speech for CMU ARCTIC database.

in epoch estimation accuracy is due to the severe attenuation of fundamental frequency components in telephonic speech signals as compared to clean speech signals which in turn reduces the energy of the signal at epochs location. Fig 3 also reinforces, the drastic difference between epoch accuracy distribution of clean speech and telephonic speech signals. The "dotted curve" in plot represents the histogram of the epoch accuracy distribution in clean speech and "continuous curve" in plot indicates that of the simulated telephonic speech signals in CMU ARCTIC database. The average epoch identification accuracy is around 2.1 ms in case of ZFF method and to check this drop in accuracy, histogram of identification timing error of epochs for both clean speech and telephonic speech is plotted in the Fig 4. From the histogram we can observe that identified epochs are within 0.4 ms deviation for all the methods in case of clean speech. However, in case of telephone speech, ZFF method performs poorly with timing error of 2 ms deviation from actual location, whereas, other methods performs relatively better with identification accuracy less than 2 ms with DPI method having lowest accuracy of 1.25 ms.

To further investigate the effect of this deviation in epochs, pitch and duration modifications are applied on the telephonic speech. Significant decrease in MOS score can be observed from Table II, which indicates that drop in the perceptual quality of telephonic speech for prosody modification. This sudden drop may be due to the fact that average epoch identification timing error of epochs in telephone speech is around 2 ms and where as in case of clean speech, epoch locations are distributed randomly from 0 to 3 ms. Objective measure of prosody modified speech using epoch obtained from ZFF, DYPSA, SEDREAMS and DPI methods are calculated using cepstral distance between clean speech and telephone speech. Cepstral distance for all the methods is also found to be higher as reported in Table III for both pitch and duration modification, which indicates that the increase in spectral distortion of the prosody modified speech for telephone speech in case of ZFF method. Whereas, cepstral distance obtained from DPI method is less compared to all other methods due to low epoch identification accuracy, however, still cepstral distance obtained from clean speech is least when compared DPI method.

V. Conclusion

In this work, role of accurate estimation of epochs for prosody modification is discussed. A significant drop in the perceptual quality of the prosody modified speech is observed when the epochs location of the original speech are randomly varied. The paper presented the issues in the epoch estimation performance of telephonic speech with reduced epoch identification accuracies using the available state of the art epoch estimation algorithms which work well for clean speech signals. The poor epoch identification accuracies make the telephonic speech less suitable for prosody modification applications. Hence a robust epoch estimation algorithm has to be devised to improve epoch estimation performance in telephonic speech and use it for telephone prosody modification applications.

VI. Acknowledgment

This work is part of the ongoing project on the development of Text-to-Speech Synthesis for Assamese and Manipuri languages funded by the Technology Development for Indian Languages (TDIL) Program initiated by the Department of Electronics and Information Technology (DeitY), Ministry of Communication and Information Technology, Govt. of India under the consortium mode headed by IIT Madras. Govind D is funded by DST fast-track project titled,"Analysis, processing and synthesis of emotions in speech".

References

[1] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 4, pp. 325–333, Sep.1995.

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

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

[4] D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *proc. INTERSPEECH 2011*, Aug. 2011.

[5] J. P. Cabral and L. C. Oliveira, "Emo voice: a system to generate emotions in speech," in *Proc. INTERSPEECH*, 2006, pp. 1798–1801.

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

[7] P. Taylor, *Text to Speech Synthesis*. Cambridge university press, 2009.

[8] K. S. Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Computer, Speech And Language*, vol. 24, no. 3, pp. 474–494, July 2010.

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

[10] S. R. M. Prasanna, D. Govind, K. S. Rao, and B. Yenanarayana, "Fast prosody modification using instants of significant excitation," in *Proc Speech Prosody*, May 2010.

[11] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Application of DYPSA algorithm to segmented time scale modification of speech," in *proc. EUSIPCO*, 2008.

[12] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 452–467, 1990.

[13] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification pitch using dct in the source domain," *Speech Commun.*, vol. 42, pp. 143–154, 2004.

[14] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 8, pp. 1602–1614, Nov. 2008.

[15] A. Prathosh, T. Ananthapadmanabha, and A. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 2471–2480, Dec 2013.

[16] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *In INTERSPEECH,2891-2894.*, 2009.

[17] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using DYPSA algorithm," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[18] N. Adiga and S. R. M. Prasanna, "Significance of instants of significant excitation for source modeling," in *proc. INTERSPEECH 2013*, Aug 2013, pp. 1677–1681.

[19] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *in 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[20] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *Selected Areas in Communications, IEEE Journal on*, vol. 10-5, pp. 819–829, 1992.

[21] S. King and V. Karaiskos, "The blizzard challenge 2009," in *Blizzard Challenge 2009*, 2009.