Significance of Glottal Activity Detection for Duration Modification

D. Govind[1], *S. R. Mahadeva Prasanna*[1] and *B. Yegnanarayana*[2]

[1]Electro-Medical and Speech Technology Laboratory
Indian Institute of Technology Guwahati, Assam, India
[2]International Institute of Information Technology Hyderabad, A.P, India
govinddmenon@gmail.com, prasanna@iitg.ernet.in, yegna@iiit.ac.in

Abstract

The objective of the present work is to demonstrate the significance of glottal activity (GA) detection for duration modification. The accurate GA regions of the speech are derived using zero frequency filtered signal (ZFFS) obtained from zero frequency filtering (ZFF) of speech. The duration of the speech is modified according to the desired scaling factors from the epochs estimated using ZFF method. Initially, the duration modified speech is synthesized using the existing epoch based fast duration modification method by processing all the epochs present in the original speech. The final duration modified speech is derived by retaining the duration modified speech samples of the GA regions and original speech samples in the non-GA regions. The improved perceptual quality of the duration modified speech is confirmed from the waveforms, spectrograms and subjective evaluations.

Index Terms: Duration modification, epochs, zero frequency filtering, glottal activity detection

1. Introduction

Duration modification is the process of modifying the speech rate according to desired modification factors without affecting the pitch, spectral and speaker characteristics of the original speech [1, 2]. Duration modification of speech finds important applications in modifying the speech rate of slow or fast speech there by improving the intelligibility and naturalness [3]. Duration modification can also be used as tool for incorporating duration features in neutral to target emotion conversion systems [4]. Other application include the fast scanning of the recorded messages in the message play back systems [5].

The main objective in duration modification is to modify the speech rate according to desired modification factors with minimum perceptual distortion. There are several methods discussed in the literature to achieve duration modification. Among these the popular methods are Overlap add (OLA), synchronous overlap add (SOLA) and pitch synchronous overlap add (PSOLA) and epoch based methods [6]. In PSOLA, speech is divided into several analysis segments, which are centered around the analysis pitch marks of the speech and having length equal to two to three pitch periods. These analysis segments are placed centering around the synthesis pitch marks and overlap added to obtain the duration modified speech. The accuracy of the PSOLA method depends mainly on the accuracy of estimated pitch marks. For further improving the perceptual quality and reducing computational complexity for duration modification, epoch based methods are developed.

Epochs are the instants of glottal closure in case of voiced speech and onset of burst and frication in case of unvoiced speech [7, 8]. In epoch based methods, the duration modification is anchored around the accurate epochs location estimated from speech. Duration modification is achieved by repeating or deleting the epoch intervals which are the intervals between successive epoch intervals. Epoch based duration modification proposed in [5] uses the epochs estimated from LP residual using group delay (GD) method and duration modification achieved by modifying the LP residual using GD epochs as anchor points. As the residual samples are less correlated than the speech samples, the duration modification performed on the LP residual gives less perceptual distortion. The zero frequency filtering (ZFF) based epoch estimation proposed in [7] provided a simple, accurate and computationally efficient way of estimating epochs from speech signals. Due to the accuracy in finding epoch locations using ZFF method, the duration modification directly on the speech samples itself provided less perceptual distortion in duration modified speech compared to GD based residual modification. Using these accurate and computationally effective epoch locations from ZFF method and modification of speech waveform samples, the work in [8] achieved a fast duration modification.

In the epoch based approaches, the duration of all the epoch intervals that belongs to both voiced and unvoiced region are uniformly modified. However, duration modification of the unvoiced region causes unnaturalness for higher duration modification factors [2]. Therefore duration modification is not recommended for the unvoiced region. [2] demonstrated improvement in the naturalness of the duration modified speech by performing duration modification only in the vowel and consonant to vowel transition regions. Alternatively, in the present work the improved duration modification is achieved by accurately finding glottal activity (GA) regions from the original speech signals and performing uniform duration modification only to detected GA regions.

The organization of the paper is as follows: Section 2 explains the ZFF based epoch estimation and GA detection. Section 3 describes the epoch based duration modification. The proposed algorithm for duration modification using GA detection is given in Section 4. The subjective studies performed to evaluate the significance of GA detection in duration modification is described in Section 5 and finally Section 6 concludes with scope for future work.

2. Zero Frequency Filtering Method for Epoch Estimation and Glottal Activity Detection

This section describes the ZFF based epoch estimation algorithm and GA detection as described in [7] and [9, 10], respec-

tively.

2.1. Epoch Estimation

For finding the epoch locations from speech, the difference speech signal is passed through cascade of two ideal zero frequency resonators (ZFR). Since the filtering through an ideal zero frequency resonator is equivalent to double integration and the output of two ZFR in cascade is equivalent to four time successive integration [7], it is given by

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \quad (1)$$

where a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1

and the difference speech signal $x(n)$ is given by,

$$x(n) = s(n) - s(n-1) \quad (2)$$

The trend in the ZFR output is removed by local mean subtraction using a window having length equal to the average pitch period of the entire utterance. The following equation describes the local mean subtraction operation

$$\hat{y}(n) = y(n) - \frac{1}{(2N+1)} \sum_{n=-N}^{N} y(n) \quad (3)$$

where $2N+1$ corresponds to the average pitch period computed over a longer segment of speech.

The trend removed signal $\hat{y}(n)$ is known as the zero frequency filtered signal (ZFFS). The positive zero crossings of the ZFFS give the epoch locations.

2.2. Glottal Activity Detection

The regions of glottal activity (GA) are characterized by the vibration of the vocal folds [9]. The GA can be determined from the strength of excitation which is the rate of glottal closure during each glottal cycle. Since ZFFS contains mainly the source characteristics, the strength of excitation can be computed from ZFFS. The strength of excitation can be computed as the ZFFS slope around each epoch locations. Figure 1(c) shows the strength of excitation derived from the ZFFS. From the strength of excitation of the speech, the GA regions can be determined as the region having strength of excitation greater than 1% of the maximum strength value. This threshold chosen for the GA detection is experimentally verified for different databases in [10]. GA regions derived from the strength of excitation are shown as the red color plot in Figure 1.

Figure 2 plots the strength of excitation of the GA and non-GA regions. The lower strength values in the non-GA regions (Figure 2(f)) has to be observed as compared to strength of excitation in the GA regions. Also the ZFFS segment (Figure 2(e)) in the non-GA region is more random as compared to the GA region ZFFS (Figure 2(b)).

3. Epoch based duration modification

The present work is an extension to the epoch based fast duration modification method proposed in [8]. The steps required for the duration modification as given in [8] are as follows:

1. ***Find the accurate epoch locations***
   The accurate epoch locations are estimated from speech by the ZFF method. These epochs are used as the anchor points for the duration modification.

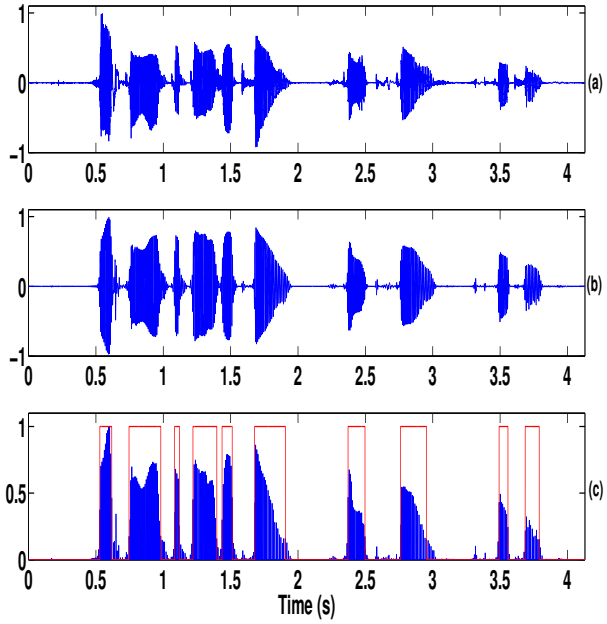

Figure 1: Glottal activity Regions in speech. (a) original speech, (b) zero frequency filtered signal, and (c) Strength of excitation. The GA regions are shows as red color plot in (c).

2. ***Derive the epoch interval plot***
   The epoch interval plot is derived by finding the intervals between successive epoch locations. This epoch interval corresponds to the instantaneous pitch period.
3. ***Resample the epoch interval plot according to the desired duration modification factor***
   To modify the duration, the epoch interval plot obtained is interpolated and resampled according to the duration modification.
4. ***Derive the modified epoch location from the resampled epoch interval plot***
   Starting from a point the new locations are derived from the modified epoch interval plot (interpolated and resampled original epoch interval). These new locations will be the modified epoch locations for the duration modification.
5. ***Waveform reconstruction***
   To reconstruct the duration modified speech, the original epoch locations that corresponds to the modified epoch locations are found first. The waveform samples in the original epoch intervals are copied to the corresponding modified epoch locations. In this way some of the epoch intervals are repeated (in case of increase in duration) and some of the epoch intervals are deleted (decrease in duration) in the duration modified speech.

This existing epoch based duration modification algorithm process all the epochs that belong to both GA and non-GA regions equally well for the duration modification. This however causes unnaturalness in the duration modified speech by uniformly modifying the durations of the non-GA regions also. Improving the naturalness of the duration modification by uniformly modifying the duration of the GA regions and retain-

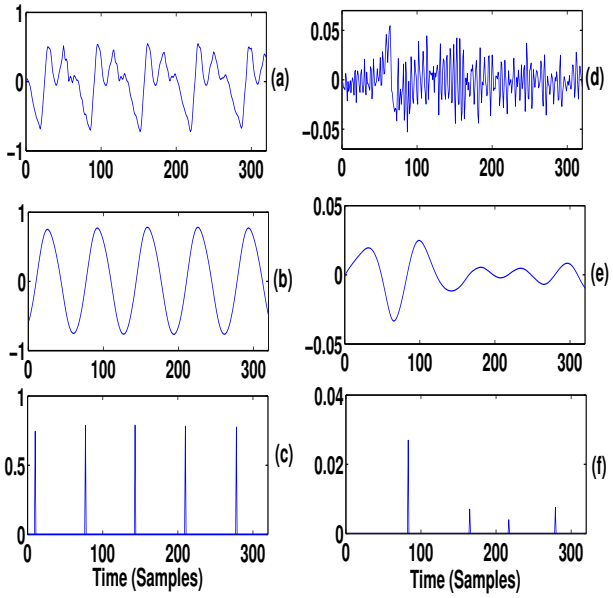

Figure 2: Strength of excitation in GA and non-GA Regions. (a) a voiced segment of speech waveform , (b) corresponding ZFFS segment and (c) Strength of excitation. (d) An unvoiced waveform segment, (e) correspomding ZFFS segment and (f) strength of excitation.

ing the original durations of the non-GA and silence regions is the motivation for the present work. The following section describes the proposed algorithm for improved duration modification.

4. Duration modification using GA detection

The steps in proposed algorithm for the improved duration modification are as follows:

- Find the epoch locations using ZFF method [7]
- Compute the strength of excitation from ZFFS and derive the GA regions [10]
- Perform the duration modification for all the epochs in both GA and non-GA regions [8]
- Identify the duration modified GA regions that corresponds to the GA regions of the original signal
- A new sequence is generated by copying successively the non GA speech segment from the original speech and duration modified GA region speech segments
- The resulting sequence will be the duration modified signal with improved naturalness

Figure 3 clearly shows how duration of each region in the speech signal is getting modified. Figure 3 illustrates a segment of speech waveform comprising of GA and non GA regions, corresponding duration modified segments without GA detection and with GA detection. The Figure demonstrates the duration modification by a factor of 2.5 (β=2.5). All the regions of the original waveform are uniformly modified by the duration modification factor in case of duration modification without GA

Table 1: 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

detection as shown in Figure 3(b). Uniform duration modification of initial silence regions also can be observed from Figure 3(b). Figure 3(c) indicates the non uniform duration modification of the segment given in Figure 3(a) by modifying only the GA regions and keeping the duration of the non GA regions intact. From the Figure 3(c) it has to be observed that duration of the non GA silence region (up to 0.5) is same as that of the original speech.

Figure 3((d)-(f)) also show the narrowband spectrogram of the original and duration modified speech segments. It has to be observed from the spectrograms that there are no spectral discontinuities in the duration modified speech from both the methods. Also it has to be noted that the duration of the non GA regions having no spectral significance (silence and short pause regions) as indicated in Figure 3(d) are modified by β in Figure 3(e). Figure 3(f) shows no spectral discontinuities by keeping the original duration of the non GA regions by GA detection.

The following section describes the subjective evaluations performed for comparing the perceptual distortions and naturalness of duration modified speech signals using both the methods for various duration modification factors.

5. Subjective evaluations

Four phonetically balanced sentences from 3 speakers (2 male and 1 female) of Arctic database are used for the study. The files initially sampled at 32 kHz are down sampled to 8 kHz and used for synthesizing the duration modified speech for various modification factors. 15 research scholars of EMST lab participated in the subjective evaluation. The duration modified speech files using both the methods along with the original speech files are presented to the subjects for the evaluations. The speech files were randomized and file names were coded before presenting to the subjects for the evaluation. A pilot test was give to each subject before the evaluation. The subjects were asked to observe the naturalness and perceptual distortions present in each file and give there opinion scores accordingly. The description of each of the scores are given in Table 1.

There were a total of 36 ($4\times4\times2$ + 4 original files) files used for the evaluation. The mean of the scores obtained for all the files for a given duration modification factor is calculated as the mean opinion score (MOS). The MOS obtained for all the modification factors and for both the methods are given in Table 2. We can observe from the Table 2 that there is a significant improvement in MOS scores for the duration modification using GA detection as compared to duration modification considering all the epochs. Also it has to be noted that the improvement is more significant for the extreme modification factors like 0.5 and 2.5.

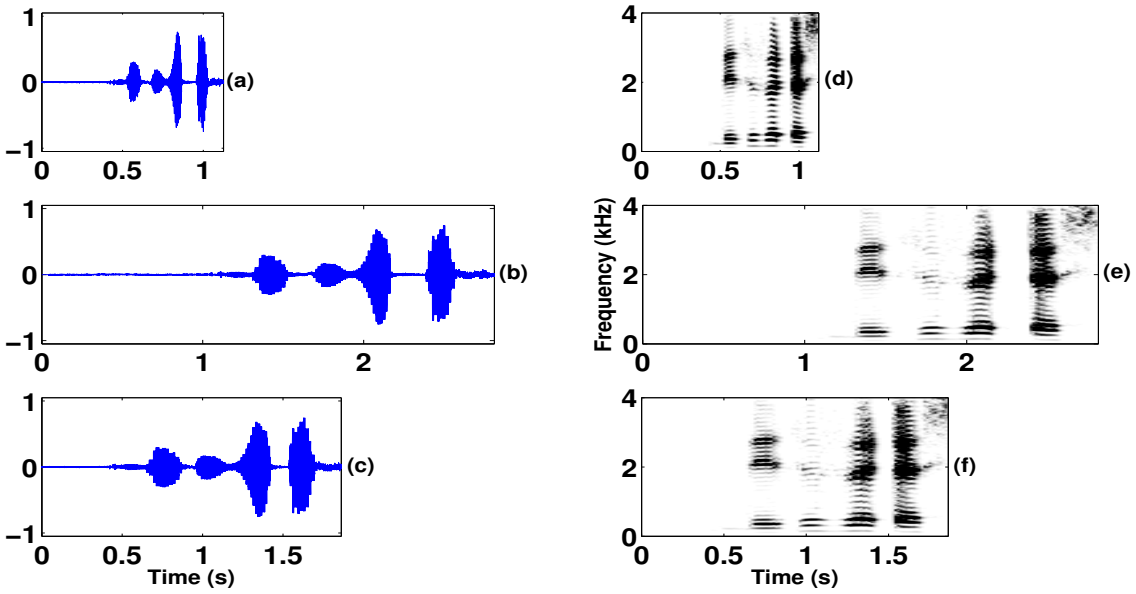

Figure 3: Duration modification by β=2.5. (a) A longer segment of the speech waveform, (b) corresponding duration modified segment with GA detection, (c) duration modified segment with GA detection and ((d)-(f)) shows the corresponding narrowband spectrograms.

Table 2: Mean opinion scores for different duration and pitch modification factors.

Mod.Factors	All Epochs	GA Region Epochs
0.5	3.12	3.93
0.7	4.15	4.45
1.5	3.90	4.1
2.5	2.75	3.55

6. Summary and conclusion

In this paper we have proposed an improved duration modification method by GA detection. The accuracy of the epochs and GA regions derived using ZFF method are exploited in the proposed duration modification. As the proposed duration modification applies on fast duration modification in the GA regions detected using ZFF method, the method is still simple and computationally effective. The subjective study shows the effectiveness of the algorithm in improving the naturalness over the duration modification achieved using the ZFF based fast duration modification method. This approach also reinforces the accurate GA detection using the ZFF method.

This improved duration modification algorithm can be used for effectively incorporating emotion specific duration features for neutral to emotional speech conversion systems. The duration information of different sound units from a large database has to be studied for various linguistic contexts and use these information for duration modification to get good quality synthesized speech.

7. Acknowledgements

This work is a part of ongoing UKIERI project (2007-2011) titled, Study of *Source Features for Speech Synthesis and Speaker Recognition* between IIT Guwahati, IIIT Hyderabad and CSTR, University of Edinburgh, UK.

8. References

[1] 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.

[2] K. S. Rao and B. Yegananarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Commun.*, vol. 51, no. 12, pp. 1263–1269, Dec. 2009.

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

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] K. S. R. Murty and B. Yegnanarayana, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, June 2009.

[10] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.