Fast Prosody Modification using Instants of Significant Excitation

S. R. M. Prasanna[1], *D. Govind*[1], *K. S. Rao*[2], *B. Yegnanarayana*[3]

[1]Department of Electronics and Communication Engineering, IIT Guwahati, Guwahati, India
[2]School of Information Technology, IIT Kharagpur, Kharagpur, India
[3]IIIT Hyderabad, Hyderabad, India

{prasanna, dgovind}@iitg.ernet.in, ksrao@sit.iitkgp.ernet.in, yegna@iiit.ac.in

Abstract

The objective of this work is to propose a fast method for prosody modification using the instants of significant excitation. The proposed method is significantly faster than the existing method based on finding the instants using group-delay and using the LP residual for incorporating the desired prosody features. This is achieved by (i) using the zero frequency filtering (ZFF) method for finding the instants of significant excitation instead of group-delay, and (ii) direct manipulation of the speech waveform rather than the Linear Prediction (LP) residual. Subjective studies indicate that the modified speech is of good quality with minimum distortion.

Index Terms: prosody modification, instants of significant excitation, zero frequency filtering, waveform modification

1. Introduction

The objective of prosody modification is to incorporate the desired prosody features by changing the pitch and duration of an utterance [1]. Such a process is useful in text-to-speech synthesis, voice conversion, expressive speech synthesis and speech rate modification [2, 3]. The two important factors in a prosody modification method are speed and perceptual quality. Fewer computations can help achieve fast prosody modification. Since speech is modified in a non-linear fashion, either in the time domain or in the frequency domain, perceptual distortion is inherent in the modified signal. But the perceptual distortion should be minimum while preserving the modified prosody features for perception. A method for prosody modification using the instants of significant excitation was proposed in [3]. The instants of significant excitation correspond to the instants around the glottal closures and the glottal openings in the case of voiced speech, and burst and frication in the case of unvoiced speech [4]. The present work focuses on improving the computation speed of prosody modification using the instants of significant excitation.

There are several methods in the literature for prosody modification [5]. These methods may be broadly grouped into time domain and frequency domain methods. Methods for prosody modification using the instants of significant excitation operate in the time domain, and hence the related time domain methods are briefly reviewed. Approaches like overlap and add (OLA), synchronous overlap and add (SOLA) and pitch synchronous overlap and add (PSOLA) operate directly on the speech signal waveform. OLA and SOLA approaches are limited to time-scale modification only, whereas the PSOLA can be applied to both time and pitch-scale modification. Most of these methods generally produce some spectral and phase distortions due to manipulation of the speech signal directly. This distortion can be reduced by using the knowledge of the instants of significant excitation, and operating on the linear prediction (LP) residual [3].

There are two main tasks in prosody modification using the instants of significant excitation, namely, finding the instants and modification of the prosody. The method for finding the instants of significant excitation based on the group-delay (GD) analysis is given in [4]. Apart from the GD method, there are several other methods for finding the instants of significant excitation. Among these, the method based on zero frequency filtering (ZFF) employs simple signal processing steps, and operates directly on the speech signal, thus reducing the number of computations significantly [6]. Performance of the ZFF method both in terms of accuracy of detection of the instants as well as location of the instants is significantly better compared to the GD method. Due to the accuracy of the detected instants, speech waveform itself can be used for prosody modification without degrading the perceptual quality.

The rest of the paper is organized as follows: Section 2 describes the significance of ZFF method for prosody modification using the instants of significant excitation. The fast prosody modification method using the instants of significant excitation is described in Section 3. Experimental results are presented in Section 4. Section 5 gives the summary, conclusions and scope for future work.

2. ZFF Method for Prosody Modification

Computational efficiency is achieved in the proposed prosody modification method due to the ZFF method for finding the instants of significant excitation and also due to the modification of the speech waveform directly. To understand this, the GD method, the ZFF method and the prosody modification using the LP residual are briefly described in this section.

2.1. GD method for determining instants of significant excitation

The GD method for determining the instants of significant excitation involves the following steps [4]:

- Difference input speech signal $x(n) = s(n) - s(n-1)$.
- Perform LP analysis [7]
  - Estimation of LP Coefficients (LPCs)
  - Computing the LP residual ($e(n)$) by inverse filtering of speech using LPCs
- GD analysis of LP residual by considering in blocks of about 1-2 pitch periods length *with shift of every sample*
  - Fourier transform of the LP residual and its time weighted version, i.e.,

$E(\omega) = FT[e(n)] = E_R + jE_I$,
and $F(\omega) = FT[ne(n)] = F_R + jF_I$
– Compute the group delay function ($\tau(\omega)$) using $\tau(\omega) = -\phi'(\omega) = \frac{E_R F_R + E_I F_I}{E_R^2 + E_I^2}$

- Median (5 point) filter $\tau(\omega)$ to remove isolated peaks in $\tau(\omega)$
- Average the group delay function over a block length to obtain the *phaseslope function* ($\bar{\tau}(\omega)$)
- Instants of significant excitation correspond to the positive zero-crossings in the phaseslope function.

2.2. ZFF Method for determining instants of significant excitation

The ZFF method for determining the instants of significant excitation involves the following steps [6]:

- Difference input speech signal $x(n) = s(n) - s(n-1)$
- Compute the output of cascade of two ideal digital resonators at 0 Hz i.e.,
  $y(n) = \sum_{k=1}^{4} a_k y(n-k) + x(n)$,
  where $a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1$
- Remove the trend i.e.,
  $\hat{y}(n) = y(n) - \bar{y}(n)$,
  where $\bar{y}(n) = \frac{1}{2N+1} \sum_{n=-N}^{N} y(n)$.
  Here $2N+1$ corresponds to the size of window used for computing the local mean, which is typically the average pitch period computed over a long segment of speech.
- The trend removed signal $\hat{y}(n)$ is termed as *zero frequency filtered (ZFF)* signal.
- The instants of significant excitation correspond to the positive zero-crossings in the ZFF signal.

2.3. Computational time for determining instants of significant excitation

The first observation is that the ZFF method works directly on the speech signal, and hence does not need LP analysis as in the GD method. The ZFF does not need GD analysis which is a computation intensive process. Finally, the ZFF method does not employ block processing for every sample shift to determine the instants. A speech signal (about 3 sec duration) for the text *"Don't ask me to carry an oily rag like that"* taken from the TIMIT database is used for determining the instants of significant excitation using both the GD and ZFF methods. Both the Matlab programs were run on the same computer. The time taken for determining the instants was about 4.83 secs in the case of the GD method, and it was only about 15.6 msec in the case of the ZFF method, demonstrating the computational efficiency of the ZFF method for finding the instants of significant excitation.

2.4. Prosody modification using instants of significant excitation by GD method

There are four main steps involved in the prosody manipulation as outlined in [3].

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

The reason for residual modification is due to its relatively less correlated samples compared to speech samples. But due to availability of accurate locations of the instants of significant excitation, the speech waveform itself can be used directly for prosody modification without producing significant distortion. This is because the presence of missing or spurious instants add to perceptual distortion. The ZFF method gives significantly fewer missing or spurious instants compared to the GD method [6] .

2.5. Comparison of ZFF and GD methods for determining instants of significant excitation

The CMU-Arctic database having simultaneous recordings of speech and EGG signals [8] was used to evaluate the GD and ZFF methods for determining the instants of significant excitation. This database consists of 1132 phonetically balanced English sentences, spoken by two male and one female talkers. The duration of the speech utterance is about 3 sec. For each speaker 100 sentences are randomly selected forming a set of 300 sentences. The reference locations of the instants of significant excitation are extracted from the voiced segments of the EGG signals by finding the peaks in the differenced EGG signal. The performance of detection of the instants was evaluated only in the voiced segments which contains a total of 42065 instants of significant excitation.

The following measures were defined to evaluate the performance of GD and ZFF methods [9].

- *Larynx cycle:* The range of samples $(1/2)(l_{r-1} + l_r) \leq n \leq (1/2)(l_r + l_{r+1})$, given a reference instant of significant excitation at sample l_r with preceding and succeeding reference instants of significant excitation at samples l_{r-1} and l_{r+1}, respectively.
- *Identification rate (IDR):* The percentage of larynx cycles for which exactly one instant of significant excitation is detected.
- *Miss rate (MR):* The percentage of larynx cycles for which no epoch is detected.
- *False alarm rate (FAR):* The percentage of larynx cycles for which more than one epoch is detected.
- *Identification error ζ:* The timing error between the reference and detected instants of significant excitation in larynx cycles for which exactly one instant of significant excitation was detected.
- *Identification accuracy σ (IDA):* The standard deviation of the identification error ζ. Small values of σ indicate high accuracy of identification.

Table 1 shows the performance result for the ZFF and GD methods. The ZFF method performs significantly better compared to the GD method in all the four factors. This improved performance will have effect on the perceptual quality of the prosody modified speech.

Table 1: *Performance of ZFF and GD methods for determining instants of significant excitation.*

Method	IDR (%)	MR (%)	FAR (%)	IDA (msec)
GD Method	94.48	4.07	1.45	0.45
ZFF Method	99.67	0.02	0.31	0.26

2.6. Computational efficiency of the proposed fast prosody modification method

For analyzing the computational efficiency, the speech signal considered in Section 2.3 is used for prosody modification (pitch period by 0.66, and duration by 2) using (i) epochs from the GD method and residual modification (EGD-RM), (ii) epochs from the ZFF method and residual modification (EZFF-RM), and (iii) epochs from the ZFF method and speech waveform modification (EZFF-SM). The time taken for each method

is tabulated in Table 2. The time for the proposed prosody modification (EZFF-SM) is significantly lower than the other two methods.

Table 2: *Computational time for prosody modification.*

Method	Time for prosody modification
EGD-RM	6.22 sec
EZFF-RM	1.78 sec
EZFF-SM	0.93 sec

3. Fast prosody modification using instants of significant excitation

There are three main steps involved in the proposed fast prosody manipulation.

1. Deriving the instants of significant excitation (epochs) from the speech signal by the ZFF method.
2. Deriving a modified (new) epoch sequence according to the desired prosody (pitch and duration).
3. Deriving a modified speech signal from the modified epoch sequence.

The prosody modification involves deriving a new speech signal by incorporating the desired modification in the pitch period and duration 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 belong to a voiced segment or unvoiced segment. The methods for creating the new epoch sequence for the desired modification are same as in the case of residual modification discussed in [3].

After obtaining the modified epochs, the next step is to derive the speech signal. For this, the original epochs closest to the modified epochs are determined. The speech samples around the original epoch are placed starting from the corresponding new epoch. Since the value of the desired epoch interval is different from the value of the corresponding original epoch interval, it is necessary to either delete some speech samples or append some new speech samples to fill the new epoch interval. Deletion of required number of speech samples is made in the tail portion of the selected speech samples. Insertion of required number of speech samples is achieved by suitably resampling about 10% of the tail portion of the selected speech samples and appending them to the end. Changes in the spectral features are visible for large modification factors of pitch period and duration as can be seen in the narrowband spectrograms for pitch period modification by 2.0 given in Fig. 1. The degradation seems to be more for residual modification compared to waveform modification.

4. Experimental Results and Discussions

Performance of the proposed prosody modification (EZFF-SM) method is compared with EGD-RM, EZFF-RM, and also with the PSOLA method operating on the speech waveform called time domain (TD)-PSOLA . The TD-PSOLA method performs pitch and time-scale modifications of the speech waveform using pitch markers as anchor points. Perceptual evaluation was carried out by conducting subjective tests with 10 research scholars. Two sentences of Indian English accent (1 male and 1 female) and two sentences of American English accent (1 male and 1 female) are used for prosody modification. For each sentence the pitch period was modified by three factors: 0.66, 1.5 and 2. Similarly, the duration was modified by factors: 0.5, 1.5 and 2.5. After the required modification using the three methods, the filenames were coded to avoid bias toward a specific method. The tests were conducted by playing the speech signals through headphones. In the test, the subjects were asked to judge the naturalness, distortion and quality of the speech for various modification factors on a five-point scale given in Table 3.

The Mean Opinion Score (MOS) for each of the pitch period and duration modification factors are given in Tables 4 and 5, respectively. For moderate modification factor of 1.5, all the methods seem to provide at least fair quality speech, and among these the proposed method provides the best possible speech quality. For all the modification factors, the scores for the methods based on the knowledge of the instants of significant excitation are better than the TD-PSOLA. This demonstrates the significance of the instants of significant excitation for prosody modification. The TD-PSOLA method uses pitch markers computed by conventional pitch extraction methods like autocorrelation analysis. From the speech production and perception point of view, most of the speech signal characteristics are present in the samples around the instants of significant excitation. Thus preserving these samples using the knowledge of the instants of significant excitation results in better speech quality. Table 6 gives the level of significance obtained by student-t distribution. The values given in the table indicate the significance of the difference in MOS of the proposed EZFF-SM from other methods for pitch and duration modifications[10]. From Table 6, it can be seen that difference in MOS score for TD PSOLA and EZFF-SM is significant as the level of confidence is more than 99%Àlso it is to be noted that the level of significance of difference between MOS scores of EZFF-SM and EZFF-RM is less than 80% both in the case of pitch and duration modification. Lower values of confidence level between the MOS of EZFF-SM and EZFF-RM indicates nearly the same perceptual qualities of speech produced from both the approaches. The difference between speech produced by ZFF based direct waveform prosody modification and GD based residual prosody modification is also significant as given in the Table 6.

It is interesting to note that the degradation in perception quality is less in duration modification compared to pitch period modification. This is because in duration modification the waveform in each pitch period is preserved. Only the number of pitch cycles are either reduced or increased depending on the modification factor. On the other hand, in pitch period modification the waveform is either truncated or stretched depending on the modification factor. Thus the degradation will be more for large pitch period modification factors. The residual modification seems to introduce higher distortion compared to waveform modification for large pitch modification factors as can be seem from column four in Table 4.

Table 3: *Ranking used for judging the quality and distortion of the speech signal for 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

5. Summary and Conclusions

In this work a method for fast prosody modification using the instants of significant excitation is developed. The existing method of prosody modification using the instants of signifi-

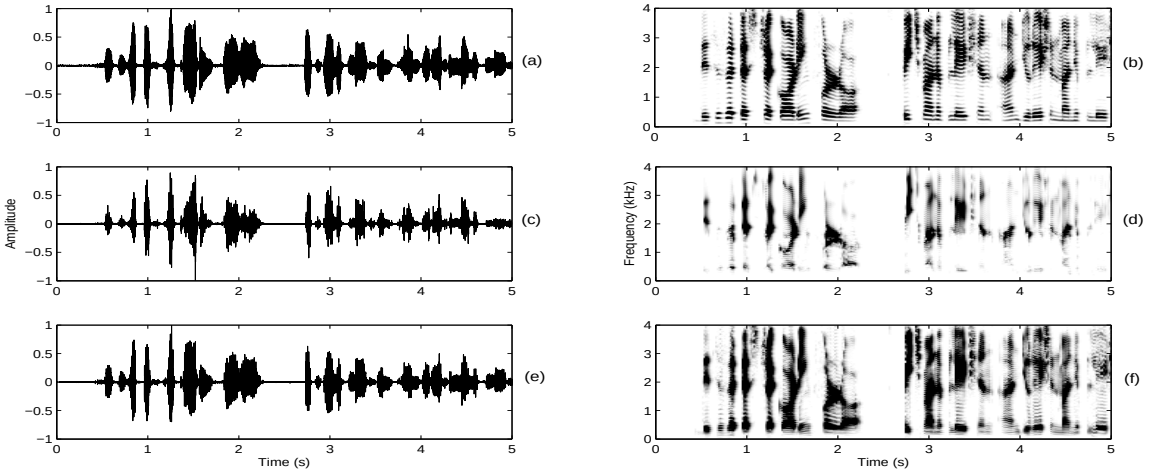

Figure 1: *Speech waveforms and their narrowband spectrograms for original speech ((a) and (b)), pitch modification by factor of 2.0 for EZFF-RM((c) and (d)) and EZFF-SM ((e) and (f)).*

Table 4: *Mean opinion scores for different pitch modification factors.*

Method	0.66	1.5	2.0
TD-PSOLA	2.31	2.97	2.39
EGD-RM	3.46	2.84	2.03
EZFF-RM	3.90	3.38	2.34
EZFF-SM	3.44	3.68	3.03

Table 5: *Mean opinion scores for different duration modification factors.*

Method	0.5	1.5	2.5
TD-PSOLA	2.75	3.17	2.05
EGD-RM	3.41	3.86	2.59
EZFF-RM	3.79	4.67	3.72
EZFF-SM	3.97	4.52	3.88

Table 6: *Comparison of significance of differences in MOS scores of different methods with EZFF-SM for pitch modification and duration modification.*

Pitch Modification			
Modification factors	EZFF-RM	EGD-RM	TD-PSOLA
0.66	<80 %	<80%	>=99%
1.5	<80 %	>=95%	>=90%
2	>=90%	>=97.5%	>90%
Duration Modification			
0.5	<80%	<80%	>=99.5%
1.5	<80%	>=95%	>=99.5%
2.5	<80%	>=95 %	>=99.5%

cant excitation uses the GD method and residual modification. This is replaced with the ZFF method and speech waveform modification. The refinement provided a computationally efficient prosody modification method. Experimental results using subjective studies show that we can even modify speech waveform and preserve the speech quality, since the instants or pitch markers are accurate.

The fast prosody modification developed in this work is computationally efficient, and hence can be explored for other applications like speech rate modification and voice conversion. For large scale modification other factors such as spectral transitions and changes in loudness in different segments need to be incorporated to reduce perceptual distortion.

6. Acknowledgements

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

7. References

[1] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification", IEEE Trans. Signal Proc., 40(3):497–510, Mar. 1992.

[2] D. G. Childers, K. Wu, D. M. Hicks and B. Yegnanarayana, "Voice conversion", Speech Commun., 8:147–158, June 1989.

[3] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation", IEEE Trans. Audio, Speech, Language Proc., 14(3):972–980, May 2007.

[4] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function", IEEE Trans. Speech, Audio Proc., 3(5):325-333, Sept. 1995.

[5] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech Commun., 16:175–205, June 1995.

[6] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals", IEEE Trans. Audio, Speech, Language Proc., 16(8):1602–1613, Nov. 2008.

[7] J. Makhoul, "Linear prediction: A tutorial review", Proc. IEEE, 63(4):561–580, Apr. 1975.

[8] J. Kominek and A. Black, "The CMU Arctic speech databases", in fifth ISCA Speech Synthesis Workshop, Pittsburgh, PA, 2004: 223–224.

[9] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm", IEEE Trans. Audio, Speech, Language Proc., 15(1):34–43, Jan. 2007.

[10] W. Chase and F. Bown, "General Statistics",Newyork : John Wiley and sons, 2000.