SUBBAND ANALYSIS OF LINEAR PREDICTION RESIDUAL FOR THE ESTIMATION OF GLOTTAL CLOSURE INSTANTS

Vikram RL[p], KV Vijay Girish[+], Harshavardhan S[*], AG Ramakrishnan[+], TV Ananthapadmanabha[o]

[+]Electrical Engineering, [*]Electrical Communication Engineering, Indian Institute of Science,
[p]Victoria University of Wellington, New Zealand, [o]Voice and Speech Systems, Bangalore, India

ABSTRACT

Many state-of-the-art techniques for estimating glottal closure instants (GCIs) use linear prediction residual (LPR) in one way or another. In this paper, subband analysis of LPR is proposed to estimate the GCIs. A composite signal is derived as the sum of the envelopes of the subband components of the LPR signal. Appropriately chosen peaks of the composite signal are the GCI candidates. The temporal locations of the candidates are refined using the LPR to obtain the GCIs, which are validated against the GCIs obtained from the electroglottograph signal, recorded simultaneously. The robustness is studied using additive white, babble and vehicle noises for different signal to noise ratios. The proposed method is evaluated using six different databases and compared with three state-of-the-art LPR based methods. The results show that the performance of the proposed method is comparable to the best of the LPR based techniques for clean as well as noisy speech.

Index Terms— glottal closure instant, GCI, subbands, composite signal, Hamming filter, HBE, HBEBEST, LPR.

1. INTRODUCTION

The instant at which the resonances or formants of the vocal tract are significantly excited within each glottal cycle is referred to as the epoch or the glottal closure instant (GCI). Automated detection of such instants serves a variety of applications such as pitch and duration modification, speaking rate modification, pitch normalization, speech coding/compression, and speaker normalization [1],[2].

GCI estimation techniques can be broadly classified into three categories: (i) techniques which use the linear prediction residual (LPR), (ii) techniques which estimate GCIs directly from the speech signal [3] [4], and (iii) techniques which use the voice source [5] or integrated LPR [6].

Since this paper proposes a method based on LPR, we restrict our review to only those techniques that primarily employ LPR for GCI detection. An early study [7] estimated epochs as the significant peaks of the Hilbert envelope of the filtered LPR. Some methods use the center of gravity concept [8] and Gabor filtering [9] of the Hilbert envelope of the LPR for pre-processing, in an attempt to improve the performance. In [10], GCIs are estimated as the positive zero crossings of the phase slope function of the LPR. This is improved by the DYPSA [11].

In SEDREAMS [4], the search for GCI is narrowed down to a short interval starting at the minimum of a windowed mean based signal, and then picking local maxima from the LPR within the interval. This algorithm requires a priori average pitch period information for assigning the window length. In [12], an evaluation of five state-of-the-art GCI detection algorithms is presented using six different databases.

GCI estimation from a limited bandwidth or wavelet decomposition of the speech signal have been reported in [13],[14],[15]. In an earlier study, [16], we reported on the use of subbands of speech signal for GCI estimation. In an exploratory study, not reported, we have found that estimation of GCIs from the envelope of LPR gives a temporal accuracy (to 0.25 ms) of 9.7%, which increases to 29.5% if lowpass filtered LPR (0 to 2000 Hz) is used. These studies have led us to investigate the use of the subbands of LPR. This is motivated by the observation that the influence of formants is minimized in the LPR, there are a large number of harmonics in the short-time spectrum of LPR and that only a few harmonics must suffice to determine the GCIs, since pitch is a low frequency datum.

The contributions of this paper are: (a) a novel subband approach to GCI estimation using LPR; and (b) experimental evidence to show the efficacy of the proposed methods in the presence of various types of additive noises at different SNR's.

2. SUBBAND ANALYSIS OF LPR

Figure 1 shows the block diagram of the proposed approach, the details of which are given below.

2.1. Pre-processing

Pre-processing consists of two steps: (i) computation of LPR (ii) bandpass filtering to obtain subband signals. We give some details of these steps. The pre-emphasized speech signal for the m^{th} frame, $s^m[n]$ may be modeled as $s^m[n] =$

$e^m[n] * v^m[n]$, where $e^m[n]$ is the excitation signal and $v^m[n]$ is the vocal tract impulse response. An estimate of the excitation signal $e^m[n]$, referred to as LPR and denoted by $\hat{e}^m[n]$, is obtained by inverse filtering the pre-emphasized speech signal. LP coefficients are obtained using Hanning windowed pre-emphasized speech signal using an LP order of f_s (sampling frequency) in kHz+2, for frames of length 20 ms and shift of 5 ms. Only the mid 10 ms segment of LPR is retained. The LPR, $\hat{e}[n]$ of the entire signal $s[n]$ is then obtained by concatenating the LPR, $\hat{e}^m[n]$.

Experimentally we have observed that in order to extract the GCI information from the subbands of the LPR, each subband signal must cover at least 2-3 harmonics of the pitch frequency for any speaker. A bandpass filter (BPF) with steep cut-off results in the temporal spread of the impulse-like components in LPR. A narrow bandwidth BPF may miss a harmonic or enclose a single harmonic and a fraction resulting in a distorted sinusoid-like signal. Since a Hamming function is a sufficiently approximate function to a Gaussian function in discrete time, we use each prototype filter in the filterbank to be a symmetric, odd length, Hamming filter.

The LPR, $\hat{e}[n]$ is passed through a filterbank to obtain the subband signals. The output of the p^{th} subband filter is given by:

$$\hat{e}_p[n] = \hat{e}[n] * h_p[n] \tag{1}$$

where $h_p[n]$ is the impulse response of the p^{th} filter in the filterbank given by

$$h_p[n] = \begin{cases} \left(\dfrac{\sin(2\pi f_{2p} r)}{\pi r} - \dfrac{\sin(2\pi f_{1p} r)}{\pi r} \right) w[n], \\ 2(f_{2p} - f_{1p}), \textit{if } r = 0. \end{cases}$$

where $r = n - D/2$, $w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{D}\right)$, D is the order of the filter, $0 \leq n \leq D$, f_{1p} and f_{2p} are the 3 dB cutoff frequencies and $w[n]$ is the Hamming window function. $h_p[n]$ is symmetric about $D/2$.

We choose the center frequencies of the successive filters to be separated by 200 Hz, each with a bandwidth of around 700 Hz at 3 dB falloff which covers at least 2 harmonics for any adult speaker. Since voiced speech is generally lowpass, the relative SNR is poorer at higher frequencies. Hence we limit the highest center frequency to 1700 Hz. Thus there are 9 subbands in this design.

We hypothesize that the locations of the peaks of the envelope of the absolute value of each subband signal, $\hat{e}_p[n]$, approximately correspond to the candidate GCIs. The envelope of the p^{th} subband signal denoted by $C_p[n]$ is obtained by piecewise cubic Hermite interpolation [17] between the local maxima of $|\hat{e}_p[n]|$. $C_p[n]$ is referred to as the p^{th} subband hereafter.

2.2. Selection of GCI candidates

Figure 2 shows the absolute value of the fifth subband signal, $|\hat{e}_5[n]|$ and its envelope, $C_5[n]$ for a voiced speech segment.

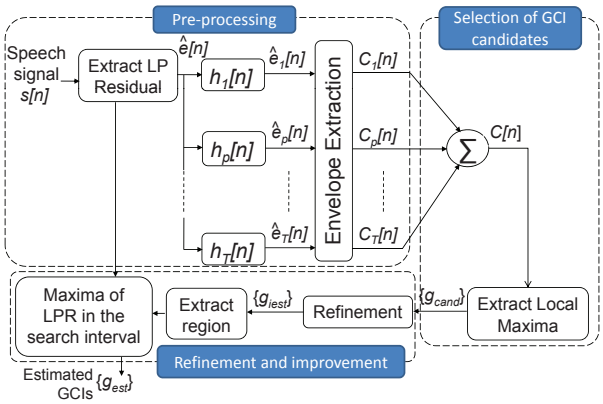


Fig. 1: Overview of the subband GCI estimation algorithm.

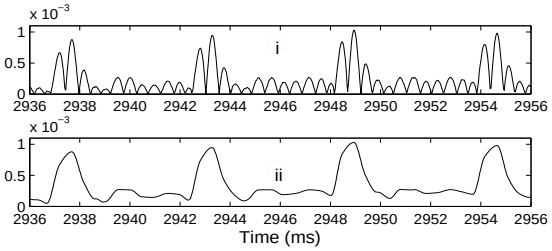


Fig. 2: (i) Absolute value of the fifth subband of LPR signal, $|\hat{e}_5[n]|$. (ii) Its envelope, $C_5[n]$).

It is observed that the subband envelopes are quasi-periodic peaky signals, with peaks near the significant excitation instants. This forms the basis for hypothesizing the peaks of subband signals as GCI candidates.

Since linear phase FIR filters are used, there is no time delay between any pair of the T subband signals, and hence we obtain a composite signal (CS), $C[n]$ as

$$C[n] = \sum_{p=1}^{T} C_p[n] \tag{2}$$

Since different subband signals are quasi-periodic, predominant peaks are obtained in the CS at the instants of significant excitation. The CS, $C[n]$ also preserves the temporal characteristics of the excitation instants required for GCI estimation. The proposed method is henceforth referred to as the Hamming Bandpass Envelope (HBE) method.

The time instants of the local maxima between successive zero crossings in the mean subtracted CS are the potential candidates for GCIs. Let the collection of these candidate GCIs be denoted by $\{g_{cand}\}$. Since the amplitude of the CS varies in voiced speech, GCI candidates may be missed if mean is subtracted over the whole $C[n]$. Hence, mean subtraction is performed framewise with a frame size of 20 ms.

2.3. Refinement and improvement of temporal accuracy

The candidate GCIs ($\{g_{cand}\}$) are refined (see Fig. 1) to obtain the set of initial GCI estimates ($\{g_{iest}\}$) by applying certain constraints on local periodicity and relative amplitudes. The local periodicity ($p[i]$) and relative amplitude ($a[i]$) are

obtained as:

$$p[i] = \sum_{j=-10}^{1} (g[i+j+1] - g[i+j])/12 \quad (3)$$

$$a[i] = (C[g[i-1]] + C[g[i+1]])/2 \quad (4)$$

where $g[i]$ is the present GCI candidate and $g[i+j]$ is the previously estimated GCI, $\{g_{iest}\}$ (for $j < 0$), or the future GCI candidate, $\{g_{cand}\}$ (for $j \geq 1$).

The conditions are defined heuristically by experimentation and a priori knowledge that GCIs occur quasi-periodically. The present GCI candidate, $g[i]$ is pruned (confirmed as spurious detection) in the refinement block, if any of the following conditions is satisfied: (i) $C[g[i]] < C[g[i-1]]$ *and* $(g[i] - g[i-1]) < 2/3(p[i])$, (ii) $(g[i] - g[i-1]) < 0.25p[i]$, (iii) $C[g[i]] < 0.1a[i]$. After this pruning step, the remaining GCI candidates form the set $\{g_{iest}\}$.

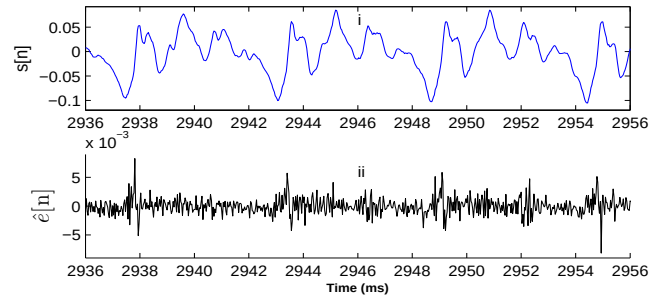
Temporal accuracy of $\{g_{iest}\}$ is improved by using the full-band LPR signal. It is known that the GCIs are concentrated near the significant local maxima (or minima in the case of polarity reversal) of the LP residual signal. A search interval 0.15 times the estimated local pitch period centered around each element of $\{g_{iest}\}$ is considered. The maxima of the LPR signal within these intervals form the set of the final estimated GCIs, $\{g_{est}\}$.

Figure 3 shows the GCIs estimated using the HBE method from the subband envelopes of the LPR of the clean speech signal. It can be observed that while the LPR signal is noisy, peaks of the individual subbands are approximately aligned with the GCIs and hence by adding them, the peaks are retained in the CS and lie closer to the reference GCIs. Also, few spurious peaks in the individual subbands are nullified in the CS due to the averaging. The final estimated GCIs are obtained within the rectangular search interval as the peaks of the LPR signal.

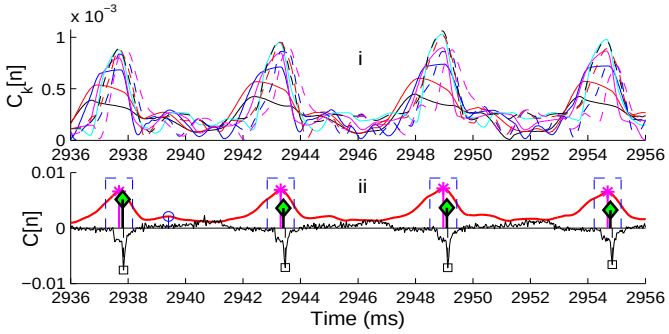
3. EXPERIMENTAL DESIGN AND RESULTS

3.1. Databases, noises and ground truth for GCIs

We have used six databases [12], each containing speech and the corresponding EGG signals. To test the robustness of the HBE method, we use speech with additive white noise and real world noises, namely babble and vehicle noise taken from the Noisex database [18]. The ground truth for GCIs is obtained from the dEGG signal [19]. The intrinsic time delay between the microphone recording and the EGG signal has been accounted for. We use the following performance measures [12] : identification rate (IDR), miss rate (MR), false alarm rate (FAR), standard deviation of error (SDE) ambiguously referred to as IDA in [12], and accuracy to 0.25 ms (Acc.25).



(a) A segment of clean speech signal and its LPR



(b) Hamming Bandpass Envelope (HBE) method applied on clean speech

Fig. 3: GCI estimation from a segment of clean voiced speech from SLT database using HBE method. (b)(i) *solid curves* are the first five subbands; *dashed curves* are the next four subbands. (ii) *thick solid curve* is $C[n]$; *circle markers* denote the spurious detections; *starred markers* denote the $\{g_{iest}\}$; *dashed rectangles* denote the search interval around $\{g_{iest}\}$; *diamond markers* denote the $\{g_{est}\}$; *thin solid curve* is the differentiated electroglottograph (dEGG) signal; *square markers* denote the reference GCIs from the dEGG signal.

3.2. Analysis of individual subbands for GCI estimation

Table 1 lists the performance of GCI estimation from individual subband envelopes (HBE 1-9) as well as the composite signal (HBE) on the APLAWD database.

Table 1: Comparison of GCI estimation performance using individual subbands, HBE and HBEBEST on clean speech from APLAWD database.

Method	IDR (%)	MR (%)	FAR (%)	Acc.25 (%)
HBE 1	81.40	18.34	0.26	34.79
HBE 2	92.69	6.28	1.02	27.19
HBE 3	94.02	4.93	1.04	29.43
HBE 4	92.94	5.60	1.46	63.86
HBE 5	91.95	6.45	1.60	70.71
HBE 6	92.11	6.38	1.51	74.85
HBE 7	91.27	7.12	1.61	75.08
HBE 8	90.36	7.98	1.66	76.13
HBE 9	89.80	8.47	1.73	75.07
HBE	93.10	6.29	0.60	81.45
HBEBEST	98.85	0.56	0.59	89.20

Table 2: Performance comparison of GCI estimation techniques on clean speech for six different databases with respect to IDR, MR, FAR, Acc.25 in % and SDE in ms. (bold entries show the best performing method)

Database	Method	IDR	MR	FAR	SDE	Acc.25
BDL	HE	97.04	1.93	1.03	0.58	46.24
	DYPSA	95.54	2.12	2.34	0.42	83.74
	SEDREAMS	98.08	0.77	1.15	**0.31**	**89.35**
	HBE	**99.02**	**0.55**	**0.43**	0.45	87.83
JMK	HE	93.01	3.94	3.05	0.90	38.66
	DYPSA	98.26	0.88	0.86	0.46	77.26
	SEDREAMS	**99.29**	**0.25**	**0.46**	**0.42**	**80.78**
	HBE	98.09	0.83	1.09	0.71	77.93
SLT	HE	96.16	2.83	1.01	0.56	52.46
	DYPSA	97.18	1.41	1.41	0.44	72.17
	SEDREAMS	**99.15**	**0.12**	0.73	**0.30**	**81.35**
	HBE	98.98	0.39	**0.63**	0.38	75.51

Database	Method	IDR	MR	FAR	SDE	Acc.25
RAB	HE	92.08	2.55	5.37	0.78	38.67
	DYPSA	82.33	1.87	15.80	0.46	86.76
	SEDREAMS	**98.87**	**0.63**	**0.50**	**0.37**	91.26
	HBE	97.54	0.68	1.78	0.86	**91.98**
KED	HE	94.73	1.75	3.52	0.56	65.81
	DYPSA	97.24	1.56	1.20	0.34	89.46
	SEDREAMS	98.65	0.67	0.68	**0.33**	94.65
	HBE	**99.51**	**0.36**	**0.13**	**0.33**	**96.20**
APLAWD	HE	91.74	5.64	2.62	0.73	54.20
	DYPSA	96.12	2.24	1.64	0.59	77.82
	SEDREAMS	**98.67**	**0.82**	**0.51**	**0.45**	**85.15**
	HBE	93.10	6.29	0.60	0.59	81.45

Table 3: Evaluation of HBE method on the combined databases with additive noises at different SNRs

	Clean	White noise				Babble noise				Vehicle noise			
SNR (dB)		-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
IDR (%)	98.05	86.64	90.50	93.07	94.94	78.99	84.97	89.87	93.12	88.69	92.57	95.10	96.61
MR (%)	1.26	7.92	5.46	3.93	2.80	12.68	8.65	5.53	3.59	6.91	4.31	2.80	1.94
FAR (%)	0.69	5.44	4.04	3.01	2.26	8.34	6.39	4.60	3.29	4.39	3.12	2.10	1.45
SDE (ms)	0.53	1.28	1.11	0.96	0.84	1.50	1.27	1.07	0.91	1.07	0.89	0.75	0.65
Acc.25 (%)	81.90	30.55	40.32	50.50	59.81	34.14	46.10	56.49	64.90	62.95	70.63	75.76	78.82

The CS HBE is better in terms of IDR, MR and FAR than most individual subbands, while Acc.25 of CS HBE is clearly better than all individual subband envelopes. IDR for all the individual subbands is above 89% except for HBE 1, proving the claim that significant GCI information is present in all the subbands.

It is possible that dynamically selecting the optimal subband component for each frame may give better accuracy for GCIs than the summed subband components. So, to arrive at the best possible (ideal) performance by dynamically selecting the optimal subband, we pick that subband whose estimate of GCI is nearest to the GCI reference obtained from the EGG signal. This approach is named as HBEBEST. Although HBEBEST is as yet to be practically realized, it denotes the ideal result obtainable using the proposed approach. Table 1 lists the results for HBEBEST for APLAWD database, which clearly indicates the potential of the proposed subband approach.

3.3. Performance comparison

The performance of the HBE method is compared with three LPR based methods, Hilbert envelope (HE) [7], DYPSA [11] and SEDREAMS [4] on six databases . The results for HE, DYPSA and SEDREAMS have been taken from [12].

Table 2 indicates that for IDR and Acc.25, the performance of HBE is almost comparable to SEDREAMS and better than HE and DYPSA for all the databases. We have observed that subband envelopes centered in the high frequency region contribute to better Acc.25 of HBE, and the averaging of subband envelopes helps in reducing the FAR. The results in Table 2 reflect the same for all the databases. Our method consistently gives accuracy of more than 75% for all the databases.

Table 3 shows the performance of our method on clean and noisy speech with white, babble and vehicle noise averaged over all databases for global SNR ranging from -5 to 10 dB in steps of 5 dB. It may be noted that the overall performance of HBE method decreases compared to the results on clean speech in Table 2. However, it is observed that the IDR, MR and FAR do not drastically degrade with noisy conditions due to the averaging of subband envelopes, which smoothens noise in any subband envelope. The Acc.25 degrades in most of the noisy cases. It is seen that the performance of the HBE method depends on the type of noise. In the case of white and babble noise, Acc.25 decreases by around 10% for a decrease in SNR by 5 dB, and does not degrade much in the case of vehicle noise. The variation in IDR with SNR is similar for white and vehicle noise.

4. CONCLUSION AND FUTURE WORK

We have shown that significant GCI information exists in each subband of speech up to 2000 Hz, and a minimum of 89% IDR (for subbands other than lowpass) can be obtained for clean speech using the HBE method. We have assumed that the pitch period does not change very rapidly while deriving constraints for the refinement algorithm, and the polarity of speech utterances are positive. Dynamic selection of the best subband using some additional knowledge may achieve robust GCI estimation closer to the HBEBEST. As an enhancement to this approach, different filterbanks may be explored with varying bandwidths and filter characteristics. Also, we will explore the effect of tuning the centre frequencies of the subband filters close to the formant frequencies of the speech signal.

5. REFERENCES

[1] B. Yegnanarayana and S. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana* , vol. 36, pp. 651-697, Oct. 2011, part 5

[2] R. Muralishankar, A. G. Ramakrishnan and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Communication*, Vol. 42/2, pp. 143-154, 2004.

[3] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, issue. 8, pp. 1602-1613, Nov. 2008.

[4] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," *Proc. Interspeech Conf.*, 2009.

[5] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 82-91, Jan 2012.

[6] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch Extraction based on Integrated Linear Prediction Residual using Plosion Index," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, Dec 2013.

[7] T. V. Ananthapadmanabha, B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis," *IEEE Trans. on ASSP*, vol. 27, no. 4, pp. 309-318, 1979.

[8] Y. M. Cheng and D. O.Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 37, no. 12, pp. 1805-1815, Dec. 1989.

[9] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group-delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762-765, Oct. 2007.

[10] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 325-333, 1995.

[11] A. Kounoudes, P. A Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I-349-I-352, 2002.

[12] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, issue. 3, pp. 994-1006, Mar. 2012.

[13] C. Prakash, N. Dhananjaya, and S. V. Gangashetty, "Detection of glottal closure instants from Bessel features using AM-FM signal," *18th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 1-4, 2011.

[14] Aicha Bouzid and Noureddine Ellouze, "Open Quotient Measurements Based on Multiscale Product of Speech Signal Wavelet Transform," *Research Letters in Signal Processing*, Volume 2007, December 2007.

[15] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals ," *IEEE Trans. Information Theory*, vol. 38 , no. 2, pp. 917-924, March 1992.

[16] Vikram Ramesh Lakkavalli, K. V. Vijay Girish, and A. G. Ramakrishnan, "Sub-band envelope approach to obtain instants of significant excitation in speech," *Proc. 18th National Conference on Communications (NCC)*, Kharagpur, India, pp. 1-5, February 3-5, 2012 .

[17] F. N. Fritsch, and R. E. Carlson, "Monotone Piecewise Cubic Interpolation," *SIAM J. Numerical Analysis*, vol. 17, pp.238-246, 1980.

[18] *Noisex-92*. [Online], Available:`http://spib.rice.edu/spib/select_noise.html`

[19] D. G. Childers, C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *Journal of the Acoustical Society of America*, vol. 97, Issue 1, pp. 505-519, Jan. 1995.