Sub-band Envelope Approach to Obtain Instants of Significant Excitation in Speech

Vikram Ramesh Lakkavalli, K V Vijay Girish, A G Ramakrishnan
Medical Intelligence and Language Engineering (MILE) Laboratory
Department of Electrical Engineering
Indian Institute of Science, Bangalore, 560012, INDIA
Email: vikram.ckm@gmail.com, kv@ee.iisc.ernet.in, ramkiag@ee.iisc.ernet.in

Abstract—**In this paper, we propose a new sub-band approach to estimate the glottal activity. The method is based on the spectral harmonicity and the sub-band temporal properties of voiced speech. We propose a method to represent glottal excitation signal using sub-band temporal envelope. Instants of maximum glottal excitation or Glottal Closure Instants (GCI) are extracted from the estimated glottal excitation pattern and the result is compared with a standard GCI computation method, DYPSA [1]. The performance of the algorithm is also compared for the noisy signal and it is shown that the proposed method is less variant to GCI estimation under noisy conditions compared to DYPSA. The algorithm is evaluated on the CMU-ARCTIC database.**

Index Terms—**glottal closure instant, epoch, GCI, DYPSA, CMU-ARCTIC.**

I. Introduction

Estimating the excitation pattern of the vocal tract helps us to understand the interaction between the vocal tract and the source in speech production. One such representation of source signal is the electro-glottograph (EGG) signal, which indicates the area of contact between the vibrating vocal folds. Thus, it is a representation of the variation of air pressure below the glottis. Vocal tract excitation is maximum when the glottis is closed abruptly and this excitation is represented by one of the peaks in the speech signal. Instant of maximum excitation is used in many applications including speech coding, speech modification, synthesis, and duration modification. To extract the instants of maximum excitation in speech signal, properties of the glottal closure instant (GCI) have been used, such as singularity property [3], and phase slope of the linear prediction residual [1]. In our approach, excitation pattern is used to estimate the GCI's. The human speech production mechanism is shown in Fig. 1.

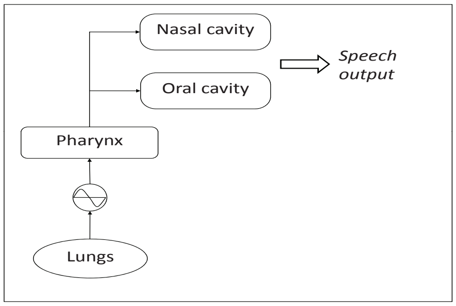


Fig. 1. Simplistic view of speech production model

Production of speech may be viewed from different perspectives. Source filter model proposed by G.Fant [10] is one such model, which assumes that the speech signal can be assumed to be generated from a source signal exciting a linear filter, where source signal is the glottal excitation signal and filter models the vocal tract. It is known that the linear prediction (LP) parameters of the speech signal gives an approximation to vocal tract shape involved in the production of speech. Speech production may also be viewed as an AM-FM model, proposed by Maragos et.al. [8], where speech signal is viewed as a combination of modulated signals. In the source-filter model of speech production, there are two factors involved in speech production, namely, the excitation signal (source), and the vocal tract transfer function (filter). Hence, extracting one information essentially needs a reliable assumption of the other.

The earliest work on estimating Glottal Closure Instant (GCI) based on the LP residual technique is by Ananthapadmanabha et.al [2]. In this approach, it is shown that the LPC residual may provide a sub-optimal GCI information. Another method based on the phase slope information of the LP residual is discussed by Smits et.al [4], where the positive zero-crossing of the phase indicates the glottal closure instants. This is further investigated by Kounoudes et.al. [1] to propose DYPSA

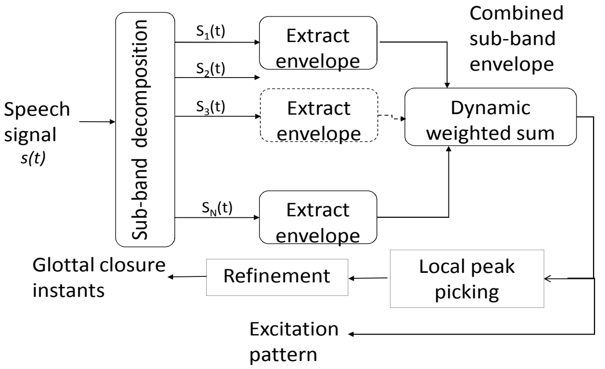


Fig. 2. GCI detection based on sub-band envelope information

algorithm. Here, dynamic programming is employed to correct the baseline phase slope based pitch mark algorithm by minimizing the pitch deviation cost and the phase slope costs.

Wavelet analysis has also been employed for the detection of GCI which is based on its singularity detection property, as GCI's are associated with singularity. The method in [3] does not yield good results for soft glottal closures such as in the cases of voice onsets and offsets. In this method, the lines of maximum amplitudes in each wavelet band is tracked dynamically to arrive at the GCI. Also, this method makes a fundamental assumption that the speech signal has predominantly negative peaks, which is equivalent to making the assumption on the polarity of the pitch mark. Sub-band analysis of speech to find pitch frequency (F_0) is discussed in [5] and [6], both using the auditory models of speech perception.

In this paper, we derive a representation of the excitation pattern of vocal tract using sub-band motivated processing. To validate our claim, GCI is extracted from the estimated excitation pattern and the result is compared with the baseline GCI obtained from the EGG signal and with the DYPSA algorithm. In order to test the robustness of the algorithm, DYPSA and the proposed method are also tested on noisy data. All the experiments are carried out on the CMU-ARCTIC database.

II. Proposed method

First, we show that the peaks of the sub-band envelope (SBE) information represent the maximum excitation instants.

Consider $v(t)$ to represent the vocal tract transfer function, and $e(t)$, the excitation signal. Speech signal $s(t)$ may be written as $s(t) = e(t) * v(t)$. Let $s_k(t)$ be the filtered speech signal around a centre frequency w_k which may be written as,

$$s_k(t) = e(t) * v(t) * h_k(t) \tag{1}$$

where, $h_k(t)$ is the impulse response of the filter selecting the speech signal around the frequency w_k, and $*$ indicates the convolution operation. Since $e(t)$ is considered to be a sequence of impulses placed at the excitation instants; the speech signal is harmonic in $w_0 = 2\pi/T$. Considering the speech signal in k^{th} band, we write (1) as,

$$s_k(t) = e(t) * v_k(t); \qquad v_k(t) = v(t) * h_k(t) \tag{2}$$

And, in the frequency domain, we may write,

$$S_k(w) = E(w)V_k(w) \tag{3}$$

Since $e(t)$ is assumed to be a sequence of impulses, that is, $e(t) = \delta(t - rT), -\infty \leq r \leq \infty$,

$$S_k(w) = \{\sum_r \delta(w - rw_0)\}V_k(w) \tag{4}$$

Here, the excitation pulses are assumed to be placed at regular interval of T for ease of analysis. Now considering only the harmonics of the excitation signal in the k^{th} band (assuming $2K+1$ harmonics, and $w_k \approx mw_0$), we have,

$$\begin{aligned} e_k(t) = exp(-j(m-K)w_0t) + ... + exp(-j(m-1)w_0t) + \\ exp(-jmw_0t) + exp(-j(m+1)w_0t) + ... + \\ exp(-j(m+K)w_0t) \end{aligned} \tag{5}$$

$$\begin{aligned} e_k(t) = exp(-jmw_0t)(1+2(cos(w_0t)+cos(2w_0t)+... \\ + cos(Kw_0t))) \end{aligned} \tag{6}$$

The envelope is defined by the term $1 + 2(cos(w_0t) + cos(2w_0t) + ... + cos(Kw_0t))$, and it is easy to notice that the excitation envelope has local maxima at $t = rT; -\infty \leq r \leq \infty$. Now consider the weighting introduced by the vocal tract on the envelope. The envelope may be approximated by

$$\begin{aligned} C_k(t) \approx a_0 + 2(a_1cos(w_0t) + a_2cos(2w_0t) + ... + \\ a_Kcos(Kw_0t)) \end{aligned} \tag{7}$$

$a_i \geq 0$. Extracting the envelope information from each band of the signal, we have a representation of the excitation signal in each band. The source excitation

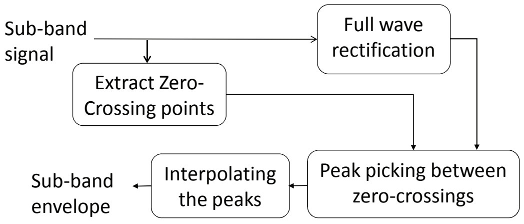


Fig. 3. Extracting the envelope from each sub-band

pattern of speech is computed as the sum of individual excitation patterns obtained from each sub-band.

$$C(t) = \sum_{k=1}^{N} C_k(t) \tag{8}$$

The algorithm is explained through a block diagram shown in Fig. 2. Speech is decomposed into sub-bands and the envelope information in each band is obtained. Sub-band envelope is extracted by considering the peak values between successive zero-crossings in the sub-band speech signal. These points are interpolated using cubic spline interpolation to obtain a smoothed sub-band temporal envelope. Extraction of sub-band temporal envelope is shown as a block diagram in Fig 3.

III. Implementation

Before starting the process, first we identify the voiced and unvoiced parts of the speech signal, and take the voiced portion for detecting pitch marks or GCI. Then, a linear phase FIR filter bank with 80 bands is designed using filter order of 64. Then the speech signal is filtered with first 10 low frequency bands since the other bands are found not to contribute much to the robustness of the GCI estimate. Then envelop of local maxima of the 10 filtered signals is taken and the unvoiced regions are assigned to zero to prevent detection of pitch in unvoiced regions. Then the envelope signal is considered frame by frame for further analysis. Transitions in each sub-band signal are then estimated, and only those bands having higher transition rate are considered to find the GCI, and this method corresponds to the dynamic weighting as indicated in Fig. 2. The processed dynamic weighted signal is the estimated excitation pattern.

On the processed dynamic weighted signal, the local maxima are found which are the contenders for the pitch marks. Now, these contenders include many extra detections other than the potential pitch marks.

The refinement of the contenders for pitch marks is now carried out by exploiting the property of local periodicity and relative amplitudes of the successive local maximas. The local pitch period is found by considering the average time-differences between consecutive maximas (which lie within the range of minimum and maximum possible pitch period) around the point of consideration.

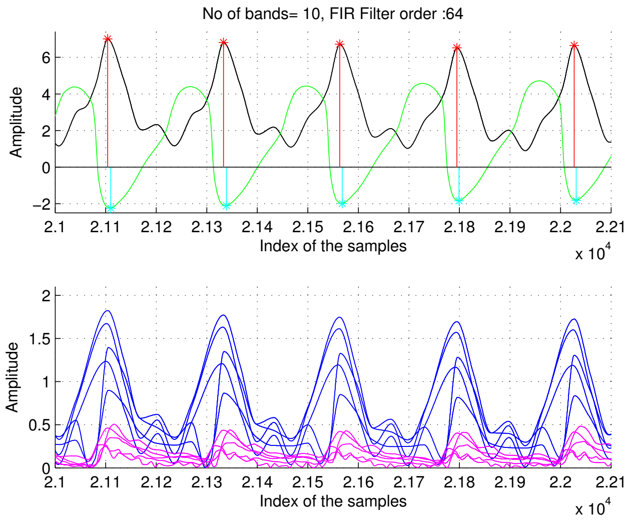


Fig. 4. Extraction of GCI from clean speech (the black curve is the Processed dynamic weighted signal; the blue curves are the envelope signals selected for addition; red peaks are the estimated GCI's; the green curve is the EGG signal; cyan peaks are the GCI's detected by EGG signal)

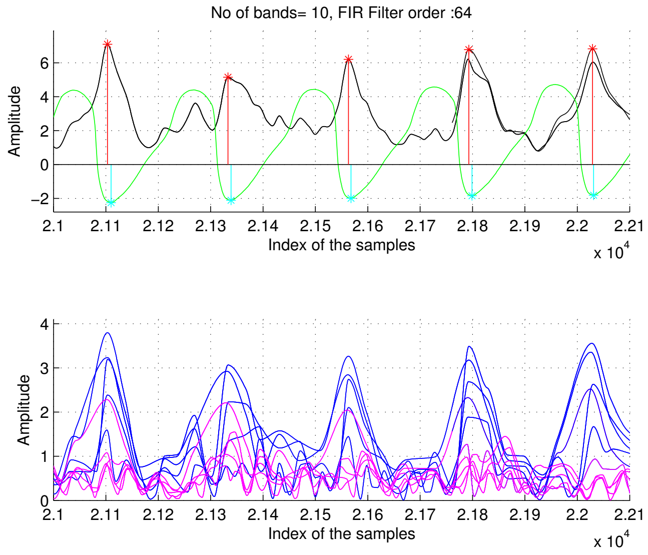


Fig. 5. Extraction of GCI from noisy signal with SNR=0 dB. Color conventions are same as Fig. 4

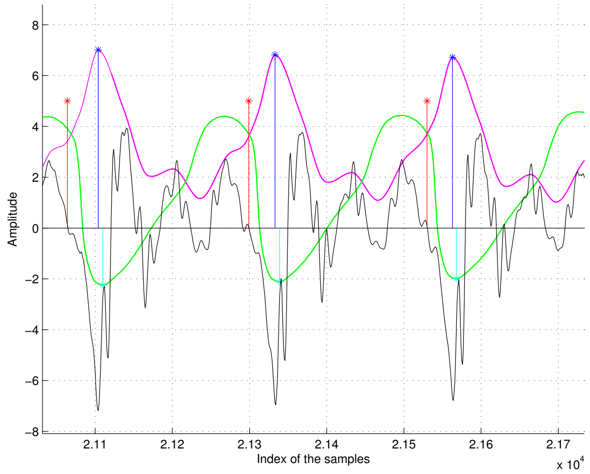


Fig. 6. Extraction of instants of minimum excitation energy from clean speech signal (The black curve is the speech signal; the magenta curve is the Processed dynamic weighted signal; blue peaks are the estimated GCI's; the green curve is the EGG signal; cyan peaks are the GCI's detected by EGG signal; red peaks are the minimum excitation points)

IV. Finding Instants of Minimum Excitation Energy in Voiced Speech

The instants of minimum excitation energy in voiced speech are important as they represent the time instants at which the glottis is completely open and the excitation energy is minimum. These instants are used in unit-concatenation for MILE-TTS synthesis system. This minimum excitation energy is useful as any concatenation at a higher excitation energy region in voiced speech is prone to degradation in naturalness of the output speech and the minimum excitation instants do not pose such challenges. Experiments on the concatenation based on the instants of minimum excitation energy is implemented in MILE-TTS [11]. A minimum excitation instant is estimated from the excitation pattern as the instant before the estimated GCI, where the derivative of the envelope is minimum, or it can also be considered as the instant of zero-crossing in speech signal occurring before the estimated GCI. The instants of minimum excitation energy and their detections are shown in Fig. 6.

V. Evaluation of GCI Accuracy

The GCI is detected from the estimate of the excitation signal using the proposed analysis of the speech signal. From Fig. 4, we may see that the peak of the estimated excitation pattern corresponds to GCI. Evaluation of the accuracy of GCI detection is carried out on the CMU-ARCTIC database. The recordings consist of the EGG signal along with the corresponding speech signal sampled at the rate of 32 kHz. First, the ground truth for glottal closure instants is collected from the recorded EGG signal. The accuracy is reported based on the deviation of the estimated GCI position with respect to the reference obtained from the EGG signal. Generally, a deviation of 1 millisecond is taken as a safe bet to consider it to be accurate. Extra detection indicates the number of extra GCIs over those detected using the EGG signal.

TABLE I
Comparison of GCI detection accuracy and Extra Detections on CMU ARCTIC database without noise

Method	**Detection accuracy in %**	**Extra detections in %**
Proposed	92.8%	1.73%
DYPSA	96.7%	2.18%

VI. Results

Table I compares the detection accuracy (deviation within 1ms duration w.r.t. GCI from EGG signal), percentage of extra detections using our SBE method and DYPSA algorithm on the clean database. It is observed from Table I that SBE method has comparable accuracy with that of DYPSA on the clean speech database. Fig. 7 compares the accuracy and extra detections of SBE and DYPSA algorithm for various values of signal to noise ratios. It is observed that our method outperforms DYPSA algorithm as the SNR decreases. Fig. 8 shows the histogram of number of estimated GCI's for the CMU ARCTIC database for deviation within 1 ms, between 1-2 ms, 2-3 ms, and above 3 ms by four bins. It is seen from Fig. 8 that when noise is added, most of the GCI's are concentrated within 64 samples or 2 ms duration using our proposed method, whereas many GCIs have deviation greater than 2 ms using DYPSA algorithm.

VII. Discussion

The proposed SBE method makes few assumptions to estimate reliable epoch information. First, it does not depend upon the explicit pitch information; however, the pitch information is estimated from the excitation pattern to prune the spurious GCIs. Second, the algorithm is simple and cost effective for real time implementation, with few filtering operations and interpolation. The proposed algorithm is compared with DYPSA for both noisy and clean speech and the results show that the SBE algorithm outperforms DYPSA for noisy speech. This shows that the algorithm is robust and may be employed in real time

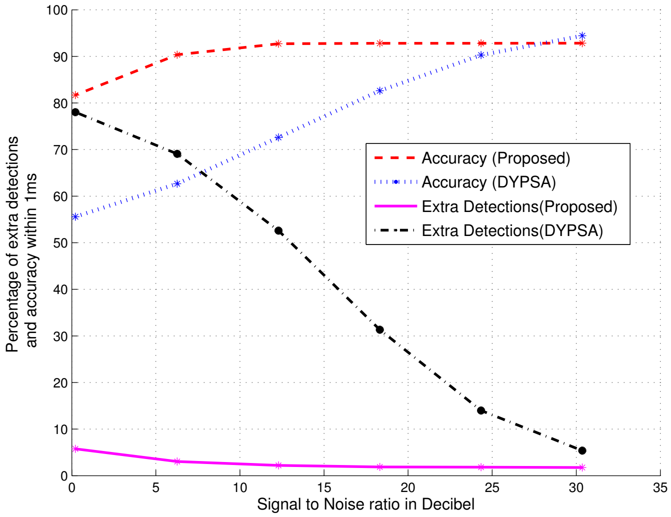
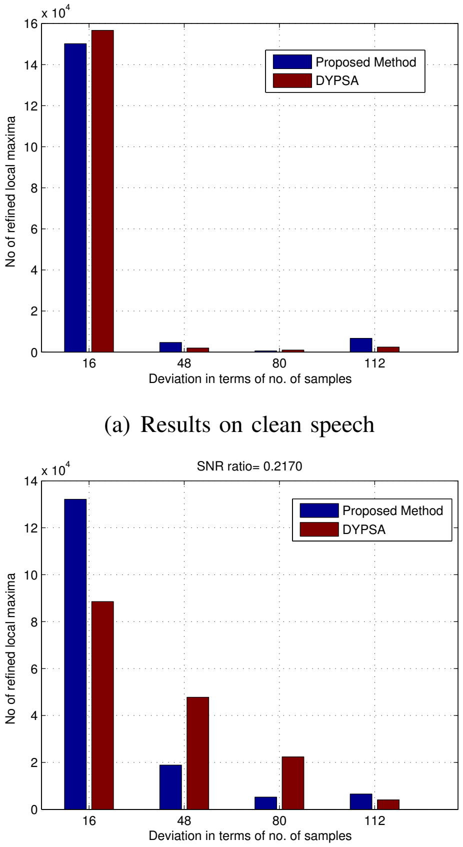


Fig. 7. Accuracy and number of extra detections as a function of SNR in dB



(a) Results on clean speech

(b) Results on noisy speech with SNR=0 dB

Fig. 8. Histograms showing the no of detected GCIs vs the deviation from those detected from EGG. 32 samples are equivalent to 1 ms

scenario. Also, the SBE algorithm gives us the flexibility to estimate the instant of minimum excitation energy which is not discussed here. The algorithm is employed for pitch synchronous unit concatenation [11] in MILE-TTS.

VIII. Conclusion

We have proposed a new method to estimate the glottal closure instants. The method estimates the glottal excitation pattern to arrive at the glottal closure instants. The excitation pattern obtained also gives a handle to estimate instants of minimum excitation, which find application in speech unit concatenation. The results of the proposed method are promising and the GCI estimation is robust to noise.

References

[1] 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)*, 2007, pp. I-349-I-352.

[2] T.V. Ananthapadmanabha, B. Yegnarayana, “Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis,” *IEEE Trans. on ASSP*, vol. 27, no. 4, 1979, pp. 309–318.

[3] N. Sturmel, C. d’Alessandro, Francois Rigaud, “Glottal Closure Instant Detection using Lines of Maximum Amplitudes of the Wavelet Transform,” *Proc. Intl. Conf. on Audio and Speech Signal Processing, ICASSP*, 2009, pp. 4517–4520.

[4] 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, 1995, pp. 325-333.

[5] K. Gopalan,“Pitch Estimation using a Modulation Model of Speech,” *ICSP* 2000, pp. 786–791.

[6] S.C. Sekhar, S. Pilli, L. C, and T.V. Sreenivas, “Novel Auditory Motivated Subband Temporal Envelope Based Fundamental Frequency Estimation Algorithm,” *14th European Signal Processing Conference (EUSIPCO 2006)*, Florence, Italy, September 4-8, 2006.

[7] M.D. Plumpe, T.F. Quatieri, and D. a Reynolds, ”Modeling of the glottal flow derivative waveform with application to speaker identification,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, 1999, pp. 569-586.

[8] A. Potamianos and P. Maragos, ”Speech analysis and synthesis using an AMFM modulation model,” *Speech Communication*, vol. 28, July 1999, pp. 195-209.

[9] D.G. Childers and C.K. Lee, ”Vocal quality factors: analysis, synthesis, and perception,” *The Journal of the Acoustical Society of America*, vol. 90, Nov. 1991, pp. 2394-410.

[10] G. Fant, “Acoustic Theory of Speech Production,” The Hague, The Netherlands: Mouton, 1960.

[11] V.R. Lakkavalli, Arulmozhi. P, and A.G. Ramakrishnan, ”Continuity Metric for Unit Selection based Text-to-Speech Synthesis,” *IEEE International Conference On Signal Processing and Communications*, 2010.