Epoch Extraction based on Integrated Linear Prediction Residual using Plosion Index

A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, *Senior member, IEEE*

***Abstract*—Epoch is defined as the instant of significant excitation within a pitch period of voiced speech. Epoch extraction continues to attract the interest of researchers because of its significance in speech analysis. Existing high performance epoch extraction algorithms require either dynamic programming techniques or *a priori* information of the average pitch period. An algorithm without such requirements is proposed based on integrated linear prediction residual (ILPR) which resembles the voice source signal. Half wave rectified and negated ILPR (or Hilbert transform of ILPR) is used as the pre-processed signal. A new non-linear temporal measure named the plosion index (PI) has been proposed for detecting 'transients' in speech signal. An extension of PI, called the dynamic plosion index (DPI) is applied on pre-processed signal to estimate the epochs. The proposed DPI algorithm is validated using six large databases which provide simultaneous EGG recordings. Creaky and singing voice samples are also analyzed. The algorithm has been tested for its robustness in the presence of additive white and babble noise and on simulated telephone quality speech. The performance of the DPI algorithm is found to be comparable or better than five state-of-the-art techniques for the experiments considered.**

***Index Terms*—Epoch extraction, glottal closure instant, GCI detection, integrated linear prediction residual, plosion index.**

I. Introduction

Flanagan defined epoch as the instant of significant excitation within a pitch period; He remarked, "presumably if such an epoch could be determined, the pulse excitation of a synthesizer could duplicate it and preserve natural irregularities in the pitch period" [1]. Miller proposed inverse filtering technique and used it to deduce that epoch lies close to the instant of glottal closure [2]. This has motivated a large number of researchers [3]-[15] to address the problem of identification of epochs or glottal closure instants (GCIs) which has assumed a great significance. We prefer to use the term epoch since it is signal-based in contrast to GCI which is a physiological term. A critical review of the state of the art methods can be seen in [15]. Also, the importance of determining the epochs or GCIs has been covered in detail in [16]. It would be a duplication of effort to review these methods or the importance of determining the epochs or GCIs. However, a brief discussion of the key features of the five state-of-the-art techniques viz., Hilbert Envelope-based detection, the Dynamic Programming Phase Slope Algorithm (DYPSA), the Zero Frequency Resonator-based method (ZFR), the Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) and the Yet Another GCI Algorithm (YAGA) reviewed in [15] is presented.

There are two major steps in epoch extraction algorithms: (a) Pre-processing of the speech signal (b) Selection of appropriate candidates corresponding to the epochs. Motivated by the work reported in [4], some methods use center of gravity [5] and Gabor filtering [11] of Hilbert Envelope (HE) of the linear prediction residual (LPR) for pre-processing. However, a recent study [15] has shown that these approaches give the lowest scores in terms of five different performance measures considered. Smits and Yegnanarayana [7] proposed the use of a signal arrived at by computing the mean group delay (GD) of a frame of the LPR samples centered at every sample, as an alternative to the HE of LPR. Here the candidates for epochs happen to be positive-going zero-crossings. However, this method suffers from a large number insertions [10]. Hence Naylor *et al* proposed DYPSA [10] algorithm where Dynamic Programming (DP) technique, with several suitably defined cost functions, was applied on the candidates derived from GD function to select the most appropriate candidates and reduce the number of insertions. An alternative method, YAGA [14] selects zero-crossing candidates of the GD function computed on the multiscale product of voice source signal instead of the LPR. Subsequently, DP technique is applied as in DYPSA to select the most appropriate candidate and reduce insertions. The accuracy of YAGA outperforms those of other techniques. Both these techniques, DYPSA and YAGA, give lower performance in the presence of noise. The poorer performance in the presence of noise may arise due to the pre-processing. Further, techniques using DP employ parameters optimized for clean speech, which might be inappropriate for noisy speech. To alleviate these problems, ZFR [12] and SEDREAMS [13] have been proposed which operate directly on the speech signal. ZFR is based on the fact that the effect of discontinuity in excitation is present over all frequencies. Hence a two-stage double integrator (Zero Frequency Filter) has been used on the pre-emphasized speech signal for pre-processing. However, this introduces a dominant low frequency trend which is removed by a mean subtraction process repeated thrice. Positive-going zero-crossings are declared as the selected candidates. This method has a tremendous noise robustness. However, ZFR method has a relatively lower accuracy (percentage of detected epochs within ± 0.25 ms of the ground truth). SEDREAMS simplifies the pre-processing by using a mean based signal

where the mean is computed by applying a Blackman-Tukey window over an appropriately selected interval. The GCI is assumed to lie within a fuzzy interval of the pre-processed signal around the zero-crossings and the exact location is postulated to be the major discontinuity in the LPR within that interval. This method retains the advantages of noise robustness of ZFR and an improved accuracy which arises due to the use of the LPR. Although HE based methods also use the LPR, it is surprising that their reported accuracy [15] is lowest.

A recent study has shown that the frequency response of ZFR resembles that of a lowpass filter [17]. The mean based signal computed in SEDREAMS using a symmetric Blackman-Tukey window is equivalent to convolving the speech signal with the FIR filter whose frequency response is that of a low pass filter. The pre-processing of ZFR and SEDREAMS can thus be interpreted as lowpass filtering. We believe that it is the lowpass filtering which is providing the noise robustness in these methods. Lowpass filter effectively picks up the fundamental, resulting in a signal which is almost sinusoidal with the frequency equal to the pitch. For speech signals with a relatively attenuated fundamental (or stronger second harmonic) as in the case of telephone quality or high pass filtered speech, these methods result in increased number of insertions.

In ZFR and SEDREAMS, the global average pitch period has to be known *a priori*. This average pitch period determines the window length for trend removal in ZFR and running average computation in SEDREAMS. This parameter is shown to be critical [13] in the sense that an inappropriate value degrades the performance. For a given database, with both male and female speakers and also for a database where the mean pitch period may vary over a wide range, the average pitch period has to be estimated and specified for each speaker. In addition, our experiments on synthetic vowels have shown that the location of epochs as determined by ZFR varies with open quotient. However, this problem does not arise in SEDREAMS since it relies on LPR for refining the locations of the candidates.

If the average pitch period is known *a priori* then identification of epochs becomes simpler. An approach such as assuming the current epoch to be known and choosing the immediate next epoch as the maximum value in the speech signal within $\pm 40\%$ of the assumed average pitch period, with random initialization can be adopted. When such an approach is experimented on a male speaker database (CMU Arctic BDL), it gave an identification rate of about 95% at 0 dB SNR (with additive white noise) with an accuracy of about 57%. On clean speech, it yielded about 97.5% identification rate with about 70% accuracy. Hence a bigger challenge lies in detecting epochs when the average pitch period is unknown and achieving a higher accuracy.

In this paper, we propose a non-linear signal processing algorithm for the identification of epochs with the following key features.

1) It operates on the half wave rectified integrated LP residual.
2) It selects the epochal candidates using a new non-linear temporal measure named the Plosion Index.
3) It does not assume the signal to be periodic and is independent of the energy contour.
4) It does not require *a priori* information of the average pitch period.
5) It does not depend on thresholds and cost functions.

The proposed method is validated using the entire CMU ARCTIC and APLAWD databases, which provide simultaneous recordings of speech and electroglottograph (EGG) signals. Illustrations of the performance of the algorithm on some special cases are also provided. It has been tested in the presence of additive white as well as babble noise. The results are compared with five state-of-the-art algorithms in terms of all the performance measures recently reviewed in [15]. Further, it has also been tested on simulated telephone quality speech and the performance is compared with ZFR, SEDREAMS and DYPSA.

II. Proposed method

The three major steps of the algorithm which are presented in this section are

- Obtaining the half-wave rectified and negated ILPR as the pre-processed signal
- Dealing with the effect of the phase on ILPR
- Using plosion index to determine the immediate next epoch, assuming the current epoch is known

A. Pre-processing

A two level pre-processing technique is proposed which is explained in this section.

1) Integrated linear prediction residual : Voiced speech is often modeled as the output of the vocal tract filter excited by a quasi periodic sequence of the derivative of glottal pulses. Epochal information is inherent in the derivative of glottal pulses, referred to as the voice source. Vocal tract transfer function is estimated using the popular Linear Prediction (LP) techniques [18], [19] applied on the pre-emphasized speech signal. Strictly speaking, the term voice source signal refers to the inverse filter output when the filter is tuned accurately to the formant data estimated over the closed-glottis interval. In linear prediction based inverse filtering, pre-emphasized speech signal is filtered to obtain the LP residual (LPR). Instead, if the inverse filtering is done directly on the speech signal, the resulting signal is referred to as integrated LP residual (ILPR) (see Fig. 7-1 in [20]) which closely approximates the voice source signal. In the present work, ILPR is obtained by inverse filtering the speech signal, with LP coefficients calculated on the pre-emphasized Hanning windowed speech samples using the autocorrelation method by setting the number of predictor coefficients to the sampling frequency in kHz plus four. Since the inverse filter is not tuned accurately to the formant data, we prefer to use the term ILPR instead of voice source in this paper. The locations of maximum negative peaks in ILPR are the representatives of the epochs.

The epochal information is reflected as local peaks both in LPR and ILPR. However, in LPR, it has been noted

that there are multiple bipolar peaks around the epoch [4], which makes unambiguous epoch extraction difficult [12]. This is because, pre-emphasis, a differencing operation, enhances high-frequency components. However, in ILPR, since there is no pre-emphasis, the peaks corresponding to epochs are less ambiguous. A 5-point symmetric moving averaging applied on ILPR further reduces the ambiguity. Henceforth, we refer to this smoothed version of ILPR simply as ILPR. As an illustration, Fig. 1 compares ILPR, Fig. 1(d), for a voiced speech segment shown in Fig. 1(a) with LPR, Fig. 1(b) and HE of LPR, Fig. 1(c). There are local peaks around the epochs in LPR along with undesired components. Although HE of LPR is relatively a better representation, methods based on HE [5], [11] are shown to have poorer performance. It may be observed that the negative peaks in ILPR near epochs are relatively unambiguous compared to peaks in LPR and peaks in HE of LPR.

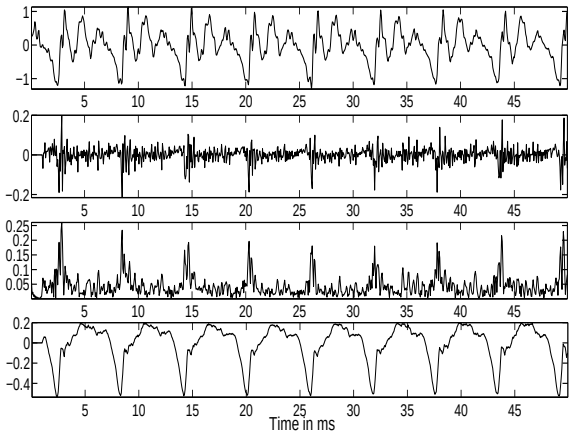

Figure 1. Illustration of manifestation of epochs in various pre-processed signals. (a) A voiced speech segment, (b) LPR, (c) Hilbert envelope of LPR, (d) ILPR.

2) Half wave rectification: It is known that volume-velocity air flow or glottal pulse reaches a peak and then decreases during the closing phase and thus has a negative slope. At or near closure, the pulse has a maximum discontinuity whose polarity is negative due to the negative slope. Thus in general, the excitation to the vocal tract primarily manifests as a large negative peak in the glottal flow derivative. Since ILPR is an approximate estimate of the glottal flow derivative, the positive going part in the ILPR contains no information about the instant of the glottal closure. Since the goal of this study is to estimate the glottal closure instant, ILPR is half-wave rectified by retaining only the negative part. Further, the rectified signal is negated.

Here, we have assumed that the speech signal to be processed is of appropriate polarity; that is the ILPR resembles the natural voice source, wherein the closure interval is relatively shorter than the opening interval resulting in a larger negative peak than the positive peak. However, if the entire speech signal is reversed in polarity due to recording conditions, then the speech signal has to be negated before epoch extraction. It has been shown that methods like ZFR and SEDREAMS too have to address this issue [21] ; That is, the type of zero-crossing (positive/negative) associated with the epochs are reversed if the polarity of the speech signal is reversed. Hence, one may use automatic methods for polarity detection such as the one proposed in [21]. In the corpora considered, there is no case of polarity reversal and hence we stick to clipping-off the positive-going part of ILPR in this work.

Figure 2 illustrates LPR, ILPR, and negated half-wave rectified ILPR (HWILPR) for a segment of voiced speech with additive babble noise at 0 dB segmental SNR. This segment is taken from Noizeus database [22] wherein the noisy signal is generated by adding noise to the speech signal filtered with a simulated telephone channel filter. The active speech level of the filtered signal is first calculated and then the noise samples are scaled and added to achieve the desired segmental SNR. This is to make the SNR independent of the silence segments which may be present in a given utterance. Such a noisy speech segment is shown in Fig. 2 (a). It is clearly seen that the HWILPR has the least ambiguity in spite of the presence of noise.

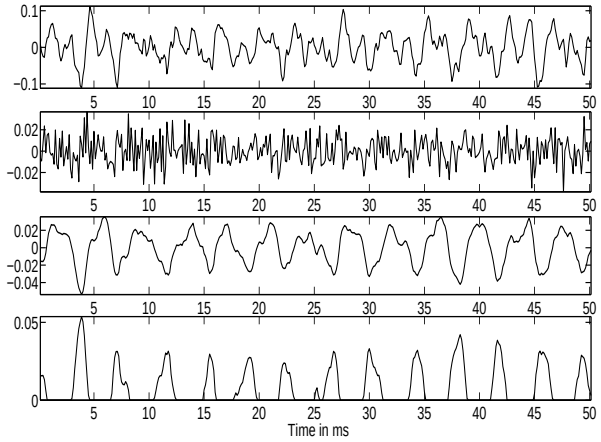

Figure 2. Illustration of the reduction of ambiguity associated with 'peaks' in ILPR corresponding to epochs through half wave rectification. (a) Voiced speech segment with additive babble noise at 0 dB segmental SNR, (b) LPR, (c) ILPR, (d) ILPR after half-wave rectification and negation (HWILPR).

3) Effect of the phase on ILPR: Inverse filtering of voiced speech using LP technique does not always compensate the phase response of the vocal tract filter exactly. It has been observed that the effect of phase angles of different formants influence the wave-shape of LPR in a complex way [4]. Hence the phase of the vocal tract filter affects the shape of the estimated ILPR as well. ILPR resembles the natural voice source signal for some speakers whereas for others the phase $(\pi/2\, radians)$ shifted version of ILPR, i.e, the Hilbert transform of ILPR (HTILPR) agrees well with the natural voice source signal. Figure 3 illustrates both ILPR and HTILPR for speakers belonging to these two categories. In Fig. 3(a), ILPR resembles the natural voice source signal but, HTILPR appears almost as a rectangular wave; this signal reaches the baseline after a prolonged closing phase and possesses an abrupt bipolar swing preceding (or following) the negative peak. These characteristics deviate from the expected natural voice source pulse shape based on the physiological considerations.

Similar observations may be noted with respect to ILPR for the speaker corresponding to Fig. 3(b), whereas HTILPR agrees well with the expected shape of the natural voice source. From this, it appears that either the ILPR or the HTILPR has to be used as pre-processed signal depending on the speaker. Although the choice of ILPR or HTILPR does not affect the identification rate of epochs, it is useful for an accurate location of epochs [1]. This behaviour of the ILPR needs a deeper investigation and is beyond the scope of the present paper. Henceforth we refer to half-wave rectified version of the appropriate signal (ILPR or HTILPR) as HWILPR for simplicity of notation.

Further, it is seen that when the ILPR is the appropriate choice, the maximum amplitude at the negative peak in ILPR is greater than that in HTILPR and vice-versa when the choice is HTILPR for a given cycle. Based on this, the choice of appropriate signal to be used for detecting the epochs for a given utterance is determined automatically using an algorithm described in the Appendix.

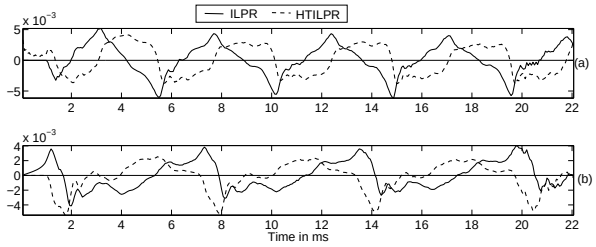

Figure 3. Illustration of the effect of phase-shift on ILPR for two different speakers. (a) ILPR and HTILPR for a speaker. (b) ILPR and HTILPR for another speaker. For the case shown in (a), ILPR resembles the natural voice source signal and for that shown in (b), HTILPR resembles the natural voice source signal.

B. Temporal features

1) Plosion Index : The goal now is to identify the instants corresponding to epochs, in the pre-processed signal, HWILPR. For this, we adopt a time domain measure which detects the location of transients.

Transients may be defined as impulse-like events occurring in a signal. Stop bursts are typical examples of such transients occurring in speech signal. It is important to detect such discontinuities in continuous speech for performing burst detection, voicing onset detection, landmark detection etc. We have proposed a point measure named plosion index (PI) for detecting such transients. Intuitively, for a signal with a transient (characterized by a significant change in local energy), the ratio of the peak amplitude in the transient to the average of absolute values over an interval of interest excluding the instant of the peak, may be expected to be very high. In order to capture the intrinsic nature of a transient-like signal, we define the temporal measure PI at an instant of interest n_0 for any signal $s[n]$ as

$$PI(n_0, m_1, m_2) = \frac{|s(n_0)|}{s_{avg}\,(n_0, m_1, m_2)} \tag{1}$$

$$where \quad s_{avg}\,(n_0, m_1, m_2) = \frac{\sum_{i=n_0-(m_1+m_2)}^{i=n_0-m_1-1} |s(i)|}{m_2} \tag{2}$$

when m_1 and m_2 are the number of samples corresponding to appropriately chosen intervals preceding n_0 and

$$s_{avg}\,(n_0, m_1, m_2) = \frac{\sum_{i=n_0+m_1+1}^{i=n_0+m_1+m_2} |s(i)|}{m_2} \tag{3}$$

when m_1 and m_2 are the number of samples corresponding to appropriately chosen intervals following n_0.

The values to be chosen for m_1 and m_2 depend on the specific application. PI is a dimensionless measure since it is a ratio and is independent of the recording level.

To illustrate the usefulness of PI for the purpose of detecting the transients (stop bursts), we consider a segment of speech signal consisting of a fricative followed by a stop followed by a vowel. Fig. 4 illustrates the PI computed (using Eq. 2 for s_{avg}) at every sample, n_0 with m_1 and m_2 corresponding to intervals of 6 and 16 ms respectively. PI is relatively high (above 500) around the stop burst (160 ms) and low elsewhere. Hence, an appropriately chosen threshold on PI can detect a transient. The concept of PI has been applied and validated for the detection of bursts associated with stops and affricates and reported recently [23].

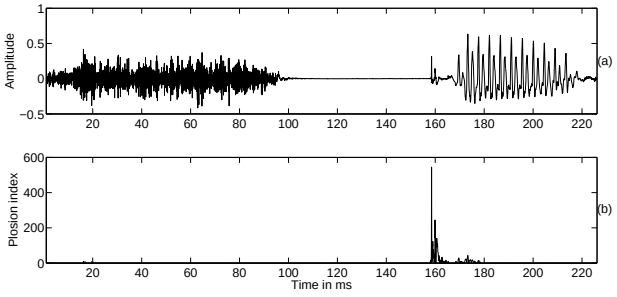

Figure 4. Illustration of the use of plosion index (PI) to capture transients. (a) A segment of speech signal with a fricative followed by a stop followed by a vowel, (b) corresponding PI.

The definition of plosion index may remind a reader of the measure crest factor or peak-to-average ratio existing in the literature [24]. However, crest factor is defined as an index for an entire signal, where both the peak and the average values are derived from the complete signal. In contrast, PI is an instant measure. Also, PI is a function of two parameters (m_1,m_2) at any given instant.

[1]Phase shift causes the negative peak in ILPR to get shifted only by a few samples, typically of order of 1 ms and hence does not affect the identification rate.

2) Dynamic plosion index: In order to measure inter-epoch interval, we define an extended temporal feature named the dynamic plosion index (DPI). DPI is PI computed as a function of varying m_2 for a given n_0 and m_1 using Eq. 3 for s_{avg}. Assuming that the lowest pitch to be extracted is 65 Hz which corresponds to a pitch period of approximately 15 ms, m_2 is varied over a range corresponding to an interval of 0 to 15 ms. DPI is a vector of dimension $1 \times N$ where N is the extent of variation of m_2. In the present context, m_2 is to the right of current epoch n_0 which is assumed to be known and the variable m_1 is chosen to be -2. The problem is to identify the immediate next epoch. We shall see later how to initialize the process for the current epoch.

DPI computed for HWILPR (Fig.5 (a)) of a voiced segment of duration 18 ms is depicted in Fig.5 (b). There are four pitch peaks in HWILPR. As m_2 increases past the reference instant, marked as n_0 in Fig. 5(a) , DPI gradually increases, reaches a peak and then decreases when m_2 begins to include the signal corresponding to next cycle. It attains a local minimum around the peak in HWILPR which is close to the next epoch. Similar behavior is repeated for the subsequent cycles.

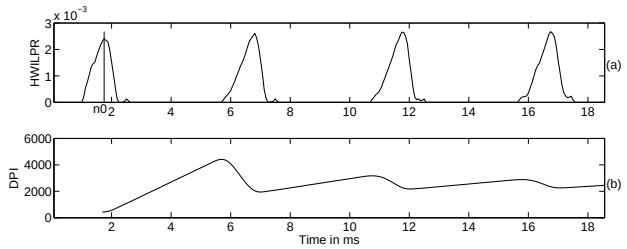

Figure 5. Illustration of determination of next epoch given the current epoch using DPI. (a) HWILPR of a voiced segment, (b) DPI computed with reference to n_0 on the signal shown in Fig. 5(a) .

C. Epoch Extraction

1) Initialization : As mentioned earlier, the problem is posed as that of determining the next epoch given the current epoch. This requires a knowledge of the current epoch. It has been found that the proposed method is insensitive to the initialization for the very first cycle which may be done arbitrarily. Subsequently, the estimated epoch location is used for initialization for the next cycle.

2) Determination of successive epochs: Having known the current epoch, the next epoch is detected as follows.

- DPI of HWILPR is computed with the current epoch as n_0.
- The peaks and valleys in the DPI are computed by detecting the positive and negative zero-crossings in its derivative, respectively.
- As noted previously, each peak-valley pair in HWILPR corresponds to a cycle. The absolute difference in the values of DPI at each peak-valley pair is computed.
- It is evident from Fig. 5(a) and Fig. 5(b), that the peak-valley pair with the largest difference corresponds to the immediate next cycle. The time instant corresponding to such a valley is noted.
- Thus, the instant of peak in HWILPR within ± 2 ms of the valley determined in the previous step, is hypothesized as the estimate of the immediate next epoch.
- The above procedure is repeated over the entire speech signal irrespective of voiced/un-voiced regions.

The proposed algorithm is henceforth referred to as DPI algorithm. Figure 6 shows a segment of voiced speech along with the corresponding DEGG signal (whose negative peaks are considered the ground truth) and the epoch locations estimated using the DPI algorithm. It may be seen that the epochs are correctly determined irrespective of the signal energy and also their locations nearly coincide with the negative peaks in DEGG signal.

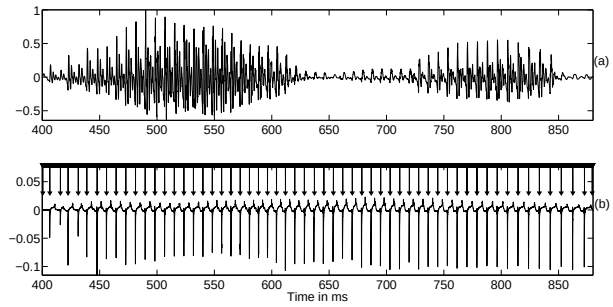

Figure 6. Illustration of the epochs estimated by the proposed algorithm. (a) A segment of voiced speech, (b) Estimated epoch locations (top trace), DEGG signal (bottom trace).

III. Evaluation

A. Databases considered and the performance measures

1) Comparison with DEGG signal: It has been shown that negative peaks in DEGG signal are very close to the instants of glottal closure [25], [26]. Epoch extraction techniques are often validated by considering negative peaks in DEGG signal as the ground truth. The DPI algorithm is validated only on the voiced segments of any given utterance since epochs are meaningful only for voiced segments. Voiced-Unvoiced decision is made by applying a negative threshold on DEGG signal. A previous study [27] has used 1/6 times the peak-to-peak value of DEGG as the threshold for V-UV decision. However, we use a worse case choice of $(1/9)$ times the maximum negative value of the DEGG signal for a given utterance so that even low energy voiced segments are captured. The compensation for the delay between the EGG signal and the acoustic signal captured by the microphone is done manually for each speaker and is assumed to be constant for all the utterances of the speaker in the database.

2) Databases: Six large databases containing speech and simultaneous EGG recordings are used for validation. The first five are from CMU ARCTIC databases. The first three contain 1132 phonetically balanced sentences. Each of these are single speaker databases corresponding to BDL-US male, JMK-Canadian male and SLT-US female. The fourth database contains non-sense words containing all phone-to-phone transitions in English uttered by a male speaker (RAB-UK male). The fifth database contains 452 sentences used in TIMIT

databases uttered by a male speaker (KED-US male). These are available in public domain in Festvox webpage [28]. APLAWD [29] is the sixth database consisting of five English sentences repeated five times by five male and five female speakers. It has been mentioned in [15], that the all pass equalization filter used in this database for correcting low frequency phase distortion has no effect on GCI detection. This database has been obtained from the author of [13]. Table I lists the number of true epoch candidates (obtained from the DEGG signal) in each of these databases.

Table I
SUMMARY OF DATABASES USED FOR VALIDATION.

Name of the Database	No. of epochs (duration in min)
BDL (1-Male)	218802 (54)
SLT (1-Female)	338875 (55)
JMK (1-Male)	152510 (54)
KED (1-Male)	64072 (20)
RAB (1-Male)	67176 (29)
APLAWD (5-Males, 5-Females)	114430 (20)
Total number of true epochs	955865 (232)

3) Performance measures: We employ the same standard performance measures, identification rate (IDR), miss rate (MR), false alarm rate (FAR), identification accuracy (IDA) and accuracy to ± 0.25 ms, as those described in many of the recent studies [10], [12], [15] which are illustrated in Fig. 6 of [15]. The first three of the above are collectively called the reliability measures and the others are called the accuracy measures.

B. Results on Clean Speech

The results of the DPI algorithm validated on clean speech using the above performance measures, are given in Table II. Also given are the results of algorithms with maximum and minimum performances amongst those compared (HE based, DYPSA, YAGA, ZFR, SEDREAMS) in a recent review paper [15]. In Table II, algorithms SEDREAMS and DYPSA have been abbreviated as SED and DYP, respectively. The fifth performance measure is considerably low for HE based method on all databases. Hence, while comparing that measure, we present the maximum and minimum amongst the remaining four algorithms along with the results of DPI algorithm.

The reliability measure, IDR, of DPI algorithm is the highest for all CMU ARCTIC databases. For the APLAWD database, it it is slightly less than the best, but well above the lowest reported. Irrespective of the database, DPI algorithm's IDR is more than 97%. As far as the standard deviation of timing error (IDA) and accuracy to ± 0.25 ms are concerned, it is observed that DPI algorithm outperforms all other algorithms for all databases. For the speakers JMK, KED and RAB, the choice for the pre-processed signal happens to be HTILPR and for others the choice is ILPR. In case we use ILPR for JMK, the accuracy to 0.25 ms would fall down significantly to about 75% from 88%.

Table III summarizes the IDR and accuracy performance measures for all the algorithms averaged over all the databases. For DPI algorithm, IDR averaged over all six databases is 99.13% which is the highest amongst all the algorithms compared. Figure 7 is a normalized histogram of the timing error of the DPI algorithm averaged over all the six databases. Identified epochs which lie within ± 0.25 ms of the ground truth is 90.77% which is the highest. IDA is 0.23 ms for DPI algorithm which is the least amongst all the algorithms.

These results may be due to the fact that DPI algorithm uses an appropriate choice of ILPR or HTILPR for pre-processed signal and rectification. The performance measures of YAGA algorithm is close to that of DPI algorithm, which may be explained by the fact that YAGA also uses the estimated voice source signal. The methods which use LPR or voice source for refinement give a better accuracy. This shows that epochs can be more precisely detected in these representations. In summary, the performance of the DPI algorithm is comparable to the best amongst the state-of-the-art algorithms, without the need for average pitch information and dynamic programming.

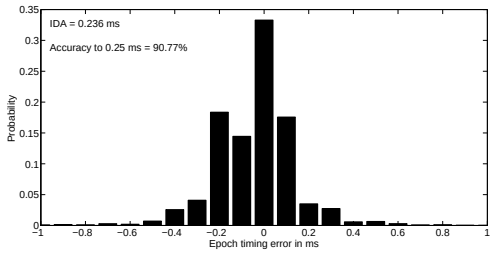

Figure 7. Normalized histogram of epoch timing error made by the DPI algorithm over all databases.

C. Demonstration of the efficacy on some special cases

In this section, we illustrate the efficacy of the DPI algorithm on typical examples of some special cases.

1) Voice-bar and nasal murmur: To demonstrate the fact that the DPI algorithm is independent of the energy contour, we consider a segment of speech taken from the utterance "will **Robin w**ear a yellow lilly" from KED database. This segment shown in Fig. 8 consists of a strong vowel followed by a weak-voice bar of a voiced-stop consonant followed by a vowel and a nasal. It also depicts the detected epochs and the DEGG signal. It is clear from Fig. 8 that the DPI algorithm detects epochs irrespective of the energy contour and even during the low-level voiced segments.

2) Creaky voice segment: Since the DPI algorithm does not make quasi-periodicity assumption, it has been applied on an arbitrarily chosen segment of creaky voiced speech taken from Voqual03 database [30] (BrianCreak3.wav). Two important distinctions of creaky voiced speech from normal speech are (i) irregular periodicities with long pitch periods (ii) presence of secondary and tertiary excitations which may arise due to ventricular incursion [31]. Figure 9 shows a segment of speech of creaky voice, along with the corresponding DEGG signal and the determined epochs. Locations of primary excitations are shown by solid lines and those of secondary excitations are shown by dashed lines. It may be seen from DEGG that there are irregular periodicities throughout the entire segment.

Table II
SUMMARY OF PERFORMANCE OF THE PROPOSED ALGORITHM ON CLEAN SPEECH ON SIX DATABASES AND COMPARISON WITH OTHER METHODS

Database	IDR %	MR %	FAR %	IDA in ms	Accuracy to±0.25ms %
BDL	**DPI - 99.11** YAGA - 98.43 HE - 97.04	**DPI - 0.15** YAGA - 0.39 DYP - 2.12	**DPI - 0.75** ZFR - 0.98 DYP - 2.34	**DPI - 0.21** YAGA - 0.29 HE - 0.58	**DPI - 92.17** YAGA - 90.31 ZFR - 80.93
SLT	**DPI - 99.47** ZFR - 99.26 HE - 96.16	**SED - 0.12** DPI - 0.22 HE - 2.38	**DPI - 0.31** ZFR - 0.59 DYP - 1.41	**DPI - 0.19** ZFR - 0.22 HE - 0.56	**DPI - 89.29** YAGA - 86.16 SED - 81.35
JMK	**DPI - 99.45** SED - 99.29 HE - 93.01	**DPI - 0.16** SED - 0.25 HE - 3.94	**DPI - 0.39** ZFR - 0.40 HE - 3.05	**DPI - 0.24** YAGA - 0.40 HE - 0.90	**DPI - 88.53** SED - 81.05 ZFR - 41.62
KED	**DPI - 99.64** SED - 98.65 ZFR - 87.36	**DPI - 0.08** YAGA - 0.63 ZFR - 7.90	**DPI - 0.02** SED - 0.68 ZFR - 4.74	**DPI - 0.17** SED - 0.33 HE - 0.56	**DPI - 98.59** YAGA - 95.14 ZFR - 46.82
RAB	**DPI - 98.96** SED - 98.87 DYP - 82.33	**DPI - 0.01** SED - 0.63 ZFR - 6.31	**DPI - 1.03** SED - 0.50 DYP - 15.80	**DPI - 0.27** SED - 0.37 HE - 0.78	**DPI - 94.01** SED - 91.26 ZFR - 55.87
APLAWD	**ZFR - 98.89** DPI - 97.17 HE - 91.74	**YAGA - 0.52** DPI - 1.99 HE - 5.64	**SED - 0.51** DPI - 0.84 HE - 2.62	**DPI - 0.34** SED - 0.45 HE - 0.73	**DPI - 89.13** YAGA - 85.51 ZFR - 57.87

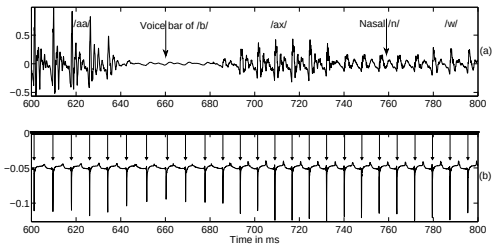

Figure 8. Demonstration of the independence of the DPI algorithm on the energy contour of the signal. (a) A segment of voiced speech comprising a strong vowel followed by a voiced stop consonant followed by a vowel and a nasal, (b) epochs determined from DPI algorithm (top trace), corresponding DEGG signal (bottom trace). DEGG has been shifted for illustrative purpose.

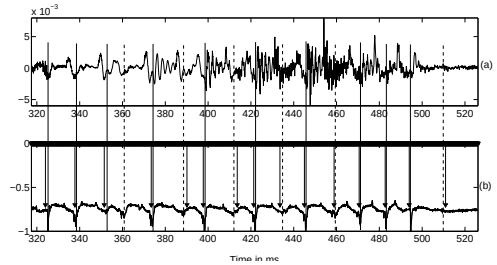

Figure 9. Demonstration of DPI algorithm on creaky voiced segment. (a) A creaky voiced segment, (b) epochs determined by DPI algorithm (top trace), DEGG signal (bottom trace). DEGG has been shifted for illustrative purpose.

DPI algorithm detects the primary epochs for this difficult case. Although there is a missed detection around 360 ms, secondary excitations around 390, 410, 430 and 460 ms have been detected. A large scale study on the performance of epoch extraction on different voice qualities (breathy, creaky, loud etc.) as reported in [32] is a problem by itself which is beyond the scope of this paper.

3) Singing voice: Since the DPI algorithm does not require *a priori* pitch information, it is expected to perform reasonably well on singing voice where the pitch spans a very large range. To ascertain this, we validate the DPI algorithm on singing voice utterances taken from Voqual03 database [30], which consists of simultaneous EGG recordings. It consists of three singers - one male and two female. The number of true epochs is 3338. We also compare the results with ZFR and SEDREAMS which require *a-priori* average pitch information for epoch detection[2]. Table IV compares the reliability performance measures for the three algorithms on singing voice. It may be seen that the large variation of pitch does not degrade the performance of the DPI algorithm whereas the performance of ZFR and SEDREAMS are relatively more affected.

IV. ROBUSTNESS ASPECTS

Some applications demand epoch extraction algorithms to be robust against various types of degradation in speech signal. In this section, we study the performance of the algorithms under two types of speech degradation namely addition of noise (white and babble) and bandwidth reduction as in telephone quality speech.

A. Noisy conditions

Two types of noise are considered in the present study, a stationary white noise and a non-stationary babble noise or cocktail party noise. White noise generated from sampling a zero mean normal distribution is added to every utterance. The variance is set in accordance with the desired global SNR. Samples corresponding to babble noise are taken from Noisex-92 database [33], scaled and added to speech signal to achieve desired global SNRs. Figures 10 and 11 depict the performance of six algorithms averaged over all databases under various SNRs for the two noise cases, respectively. Performance measures of the algorithms other than DPI are taken from the recent review paper [15].

[2]Average picth period was estimated and provided for ZFR and SEDREAMS.

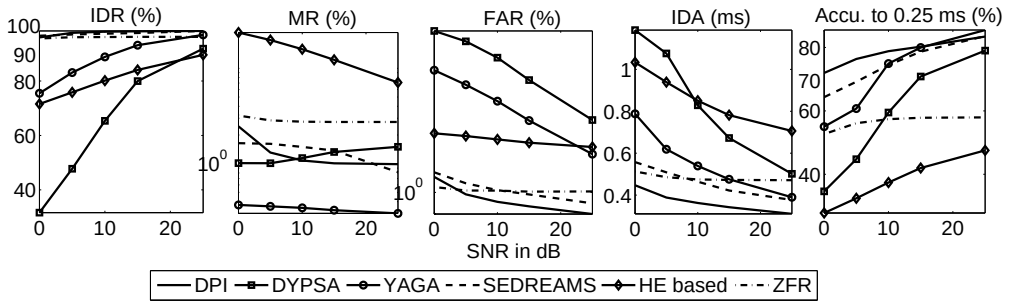

Figure 10. Performance of six different algorithms over all databases at different SNRs (0 to 25 dB) with additive white noise. The values of performance measures for algorithms other than DPI method have been taken from [15].

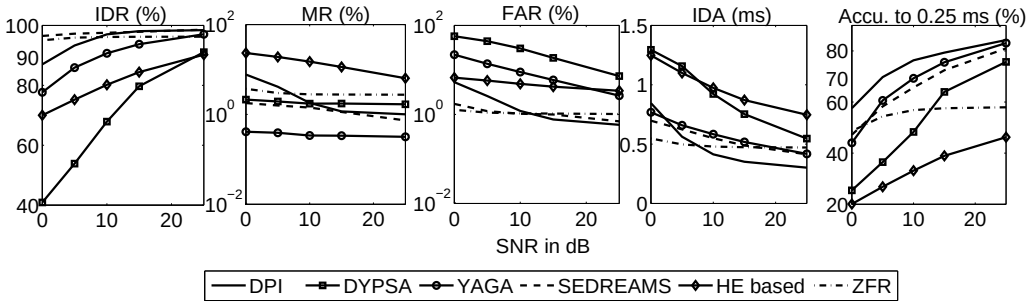

Figure 11. Performance of six different algorithms over all databases at different SNRs (0 to 25 dB) with additive babble noise. The values of performance measures for algorithms other than DPI method have been taken from [15].

In the case of white noise, it may be seen that the IDR of the DPI algorithm is almost unchanged and comparable to ZFR and SEDREAMS and is 96% at 0 dB SNR. The superiority of the accuracy performance of the DPI algorithm is retained even at 0 dB SNR. It has the lowest IDA at all SNRs and better accuracy below 15 dB SNR. Even at 0 dB SNR, almost 72% of the determined epochs are within 0.25 ms of the ground truth.

In the case of babble noise, IDR of the DPI algorithm degrades gradually below 10 dB SNR from about 97% and reaches 87% at 0 dB SNR. This lowering of performance below 10 dB may be due to the model dependence. However, it is better than other model based techniques such as DYPSA, YAGA and HE. IDA of the DPI algorithm is lowest above 10 dB SNR, and becomes slightly higher than ZFR and SEDREAMS, below 10 dB. Accuracy to 0.25 ms of DPI algorithm remains highest at all SNRs.

In summary, DPI algorithm is highly robust against white noise in terms of every performance measure considered and offers the highest accuracy at all SNRs. For babble noise, the performance is comparable to the best in the literature till 10 dB SNR below which it slightly degrades. However, the accuracy performance is superior for all SNRs.

B. Telephone quality speech

To examine the robustness of epoch extraction algorithms against bandwidth degradation as in telephone quality speech, we validate the performance of four algorithms viz., DPI, DYPSA, ZFR and SEDREAMS on simulated telephone quality speech, using the same performance measures as defined in the earlier section. Since a large database consisting of actual telephone channel speech with simultaneous EGG recordings is not available, we use simulated data.

Table III
PERFORMANCE MEASURES AVERAGED OVER ALL DATABASES FOR VARIOUS ALGORITHMS.

Method	IDR in %	IDA in ms	Accuracy to 0.25 ms (%)
DPI	**99.13**	**0.23**	**90.77**
YAGA	98.38	0.34	83.40
SEDREAMS	98.81	0.34	80.80
ZFR	96.37	0.42	57.90
DYPSA	95.11	0.44	71.90
HE BASED	94.60	0.67	39.70

Table IV
PERFORMANCE OF THREE ALGORITHMS ON SINGING VOICE.

Algorithm	IDR (%)	MR (%)	FA (%)
DPI	**94.1**	5.04	**0.5**
ZFR	88.2	0.01	11.8
SEDREAMS	83.3	3.03	13.63

Telephone channel can be approximated by a bandpass filter (BPF) between 300 and 3400 Hz. We designed a BPF and used it to simulate the telephone quality speech. The magnitude response of the filter is defined in the frequency domain using a raised cosine function between 0 and 300 Hz, unity between 300 and 3400 Hz and again a raised cosine function from the folding frequency up to 3400 Hz. The speech signal is down-sampled to 8 kHz and the frequency domain implementation of BPF gives simulated telephone quality speech which is then used as input for the epoch extraction algorithms.

The algorithms are evaluated on three databases namely BDL, KED and SLT, which cover male speakers of two

Table V
RESULTS OF VARIOUS ALGORITHMS ON SIMULATED TELEPHONE QUALITY SPEECH

Database	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to 0.25 ms (%)
BDL	DPI	**97.69**	0.04	**1.87**	**0.20**	**93.09**
	ZFR	42.05	0.01	57.95	0.28	35.21
	SED	83.72	0.02	16.26	0.31	72.80
	DYP	95.07	0.05	4.37	0.35	85.82
KED	DPI	93.44	0.04	6.14	**0.26**	**93.88**
	ZFR	30.19	0.07	69.75	0.97	6.12
	SED	78.70	0.01	21.29	0.36	79.55
	DYP	**98.12**	0.04	**0.14**	0.28	86.24
SLT	DPI	98.66	1.01	**0.03**	0.28	**87.60**
	ZFR	**99.28**	0	0.07	**0.19**	78.94
	SED	99.18	0	0.08	0.33	78.56
	DYP	96.15	0.09	2.92	0.41	72.56

different accents and one female speaker, respectively. The results are presented in Table V.

The performance of ZFR and SEDREAMS degrade severely for male speakers since the resulting zero-frequency resonator output and mean-based signal, are not sinusoidal. As every zero-crossing is deemed as an epoch candidate, false alarms significantly increase. The degradation in the performance of SEDREAMS is less than that of ZFR. This is due to the fact that effective lowpass filter of ZFR is steeper than that used in SEDREAMS (frequency response of Blackman-Tukey window). Further, the performance is much worse in the case of the speaker KED. This is because, the relative spectral level of fundamental is lower for this speaker than that of the others even in clean conditions. This explains the lowest score of ZFR method for this speaker (87% IDR) under clean conditions whereas it is consistently more than 98% for all others. However, DPI method and DYPSA suffer very little degradation in the performance on telephone quality speech. This may be due to the absence of lowpass filtering. DPI method is not only reliable but also accurate.

The scenario is completely different for the female speaker SLT. There is no degradation in the performance in any of the algorithms since the telephone channel does not degrade the fundamental. There is a slight lowering of accuracy in the case of SEDREAMS and ZFR.

V. CONCLUSION

In this paper, we have proposed an algorithm, named the DPI algorithm, for epoch extraction. Half wave rectified and negated integrated linear prediction residual is used as the pre-processed signal which appears to be relatively less ambiguous to identify epochs compared to other signal representations. The effect of phase of formants on ILPR has been dealt with appropriately. A new temporal measure, Plosion Index proposed to detect 'transients' in speech signals has been used. An extension of PI, called the Dynamic Plosion Index (DPI) is applied on the pre-processed signal to detect the epochal candidates. The method has been validated using six large databases comprising 15 speakers against EGG recordings. It is tested for its robustness in the presence of additive white and babble noise. Also, robustness is studied on simulated telephone quality speech. The performance of DPI algorithm is compared with several state-of-the-art algorithms. It has been found the performance of DPI algorithm is comparable to the best in the literature, for all the cases studied. DPI algorithm is effective even for low-level voiced segments. It does not require *a priori* pitch information which suggests that it may be applied to speech with large range of pitch as in the case of emotional speech or music.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Thomas Drugman for lending APLAWD database and other important data required for comparison. They would also like to acknowledge the ananymous reviewers and the associated editor for their constructive comments which helped to improve the presentation. The first author would like to thank his friend Mr. Abhiram for the help rendered during the preparation of some of the figures.

APPENDIX

ALGORITHM FOR ASCERTAINING THE CHOICE OF APPROPRIATE SIGNAL

- A given utterance is divided into non-overlapping frames of 20 ms for LP analysis.
- ILPR and its Hilbert transform (HTILPR) are estimated for each frame.
- The ratio of the absolute value of the maximum negative peak in ILPR to that in HTILPR is calculated for each cycle.
- The median of such ratios is calculated.
- If the median is greater than one then ILPR is taken to be the appropriate signal else HTILPR is used.

The same algorithm is used with noisy speech as well.

REFERENCES

[1] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Springer,Newyork, 1972.

[2] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Amer.*, vol. 31, pp. 667–677, 1959.

[3] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 6, pp. 562–570, Dec.1975.

[4] ——, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no.7, pp. 309–319, Aug.1979.

[5] Y. M. Cheng and D.O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no.12, pp. 1805–1815, Dec.1989.

[6] Y. K. C. Ma and L. F. Willeams, "A frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 258–265, Apr.1994.

[7] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no.5, pp. 325–333, Sep.1995.

[8] V. N. Tuan and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Proc. Eurospeech*, Budapest, 1999, pp. 2805–2808.

[9] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 2002, pp. 349–352.

[10] 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 Lang. Process.*, vol. 15, no.1, pp. 34–43, Jan.2007.

[11] 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.

[12] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

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

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

[15] 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, no. 3, pp. 994–1006, Mar. 2012.

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

[17] K. S. S. Srinivas and K. Prahallad, "An 'FIR implementation of zero frequency filtering of speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20 no.9, pp. 2613–2617, Nov. 2012.

[18] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50 no.2, pp. 637–655, 1971.

[19] J. Markel, "Digital inverse filtering - a new tool for formant trajectory estimation," *IEEE Trans. on Audio and Electro.*, vol. Au-20, pp. 129–137, Jun. 1972.

[20] T. V. Ananthapadmanabha, "Acosutic factors determining perceived voice quality," in *Vocal fold Physiology - Voice quality control*, O.Fujimura and M. Hirano, Eds. San Diego,Cal.: Singualr publishing group, 1995, ch. 7, pp. 113–126.

[21] T. Drugman and T. Dutoit, "Oscillating statistical moments for speech polarity detection," in *Proc. of Non-Linear Speech Processing Workshop (NOLISP11)*, Las Palmas, Gran Canaria, Spain, 2011, pp. 48–54.

[22] Noiseus. [Online]. Available: http://www.utdallas.edu/~loizou/speech/noizeus/

[23] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of closure-burst transitions of stops and affricates in continuous speech using plosion index," *J. Acoust. Soc. Amer.*, revised manuscript under review.

[24] S. Boyd, "Multitone signal with low crest factor," *IEEE Transactions On Circuits and Systems*, vol. 10, pp. 1018–1022, 1986.

[25] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, Nov.1991.

[26] D. G. Childers and A. K. Krishnamurthy, "A critical review of electroglottography," *CRC Crit. Rev. Bioeng.*, vol. 12, pp. 131–164, 1985.

[27] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 505–519, Jan.1995.

[28] The festvox website. [Online]. Available: http://festvox.org

[29] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," *Univ. College London, London, Tech. Rep.*, 1987.

[30] voqual. [Online]. Available: http://archives.limsi.fr/VOQUAL/voicematerial.html

[31] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *J. Acoust. Soc. Amer.*, vol. 103, pp. 2649–2658, 1998.

[32] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication*, vol. 55, pp. 295–314, 2013.

[33] Noisex-92. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Sectionl/Data/noisex.html