IEEE Signal Processing Letters

Cumulative Impulse Strength for Epoch Extraction

Journal:	*IEEE Signal Processing Letters*
Manuscript ID	SPL-17876-2015.R1
Manuscript Type:	Letter
Date Submitted by the Author:	n/a
Complete List of Authors:	Prathosh, A.P.; Xeror research Center India, P, Sujith; Ittiam systems, Ramakrishnan, A. G.; Indian Institute of Science, Electrical Engineering Kumar Ghosh, Prasanta ; Indian Institute of Science, Electrical Engineering
EDICS:	SPE-ANAL Speech coding, synthesis and analysis < SPE Speech processing

Cumulative Impulse Strength for Epoch Extraction

Prathosh A. P., *Member, IEEE* Sujith P, Ramakrishnan A. G. , *Senior Member, IEEE* and Prasanta Kumar Ghosh, *Senior Member, IEEE*

***Abstract*—Algorithms for extracting epochs or glottal closure instants (GCIs) from voiced speech typically fall into two categories: (i) ones which operate on linear prediction residual (LPR) and (ii) those which operate directly on the speech signal. While the former class of algorithms (such as YAGA and DPI) tend to be more accurate, the latter ones (such as ZFR and SEDREAMS) tend to be more noise-robust. In this paper, a temporal measure termed the cumulative impulse strength is proposed for locating the impulses in a quasi-periodic impulse-sequence embedded in noise. Subsequently, it is applied for detecting the GCIs from the inverted integrated LPR using a recursive algorithm. Experiments on two large corpora of speech with simultaneous electroglottographic recordings demonstrate that the proposed method is more robust to additive noise than the state-of-the-art algorithms, despite operating on the LPR.**

***Index Terms*—GCI detection, epoch extraction, cumulative impulse strength, impulse tracking.**

I. Introduction

Pitch-synchronous analysis of the voiced speech signal is a popular technique in which the glottal closure instants (GCIs or epochs) are used to define the analysis frames. Epochs are utilized in various applications including pitch tracking, voice source estimation [1], speech synthesis [2], [3], prosody modification [4], [5], [6], [7], voiced/unvoiced boundary detection [8] and speaker identification [9], [10]. Hence, automatic detection of the GCIs from the voiced speech signal is considered to be an important problem in speech research. Comprehensive reviews of the importance of the GCI detection problem and summary of the state-of-the-art algorithms may be found in [11], [12].

Many of the popular GCI detectors can be categorized into two classes. Detectors belonging to the first class adhere to the source-filter model of speech production and locate GCIs from an estimate of the glottal source signal such as linear prediction residual (LPR) and the voice source (VS) signal. Algorithms like Hilbert Envelope (HE) based epoch extractors [13], Dynamic Programming Phase Slope Algorithm (DYPSA) [14], Yet Another GCI Algorithm (YAGA) [15], Dynamic Plosion Index (DPI) [16] and sub-band decomposition method [17] fall into this category. The second class of algorithms such as Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) [18] and Zero-frequency resonator (ZFR) [19] operate directly on the speech signal without any model assumption or deconvolution. The former class of algorithms are more accurate than the latter ones [12]. This may be because the GCIs are associated with the source signal, which forms the basis for the analysis for these algorithms. However, they are believed to be more susceptible to noise compared to SEDREAMS and ZFR, mainly because of inaccurate estimation of the LPR in the presence of noise. Further, ZFR and SEDREAMS assume that the average pitch period (APP) is known a priori while the former class of algorithms do not require the information of APP. Motivated by these observations, in this paper, we explore whether an LPR based GCI detection scheme could be noise robust if the APP can be estimated *a-priori*. Specifically, we propose a generic measure named the cumulative impulse strength (CIS) to locate the impulses in a quasi-periodic impulse train corrupted by additive noise. Further, using CIS, we devise a recursive algorithm to extract GCIs from the integrated LPR (ILPR) [16] of the voiced speech and evaluate the proposed algorithm using two speech databases with simultaneous electroglottographic (EGG) recordings in both clean and noisy conditions.

Prathosh is with Xerox research center India, Sujith is with Ittiam systems India, and the other authors are with Indian Institute of Science, Bangalore - 560012, India. (e-mail: prathosh.AP@xerox.com, sujith.p4@gmail.com, ramkiag@ee.iisc.ernet.in, prasantg@ee.iisc.ernet.in.)

II. Impulse-location detection using CIS

A. Motivation

It is known that the GCIs coincide with the local negative peaks of the voice source signal [20]. Thus, a GCI extraction algorithm which uses the voice source signal typically involves two stages - (i) transformation of the speech signal into a domain where the voice source signal is best represented (such as ILPR), (ii) accurately picking the peaks corresponding to GCIs from the transformed signal. To reduce the error committed by the peak-picking algorithm, the temporal quasi-periodicity property of the voiced speech can be exploited. In a quasi-periodic impulse-train like sequence, the accuracy of detection of each impulse could be improved by using the knowledge of the location and the strength of the previous impulses. That is, the impulse-like behavior at a given instant of time may be determined not only by analyzing some local properties of the signal around that instant but also by taking into account the global behavior of the signal around all the previous impulse locations. Based on this intuition, we define a measure named the cumulative impulse strength to estimate the locations of the impulses in a quasi-periodic impulse train.

B. Cumulative impulse strength

Let $r[n]$ be an amplitude-perturbed, quasi-periodic impulse train of length N represented as follows:

$$r[n] = \sum_{k=1}^{N} A_k \delta[n - n_k], \tag{1}$$

$$n_k = n_{k-1} + N_0 + \Delta_k, \ 2 \leq k \leq N. \tag{2}$$

where n_k is the location of the k-th impulse with amplitude A_k, $\delta[n-n_k]$ denotes the Kronecker delta function, N_0 is the average period of $r[n]$ and Δ_k is the deviation of $n_k - n_{k-1}$ from N_0. The measure CIS is defined recursively at each location n, by combining the effect of the signal r and the CIS C around the previous impulse location. That is, if $\rho = \max_k |\Delta_k|$, the CIS $C[n]$ at the n-th sample is defined as follows:

$$C[n] = \max_{n-N_0-\rho \le m \le n-N_0+\rho} \big(C[m] + r[m]\big) \qquad (3)$$

In order to locate the impluses from $C[n]$, we define one more sequence $V[n]$ as follows.

$$V[n] = \operatorname*{argmax}_{n-N_0-\rho \le m \le n-N_0+\rho} \big(C[m] + r[m]\big). \qquad (4)$$

That is, at each sample n, $V[n]$ stores the location that maximizes $C[n]$ within the search interval defined in Eq. 3. Once the location of the last impulse is known, a back tracking procedure is employed to locate all the impulses from $V[n]$ as follows: if n_k corresponds to the k^{th} impulse location, the $(k-1)^{th}$ impulse location is given by $V[n_k]$. The location of the final impulse is defined to be that which maximizes $r[m]$, $N-1-N_0+\rho \le m \le N-1$. This is because the location of the maxima of the $r[m]$ within the last periodic interval corresponds to the final impulse.

C. Illustration of CIS on synthetic data

In this section we report an experiment where the objective is to estimate the locations of the impulses using the CIS, from an impulse train (N_0=150) of 10 impulses spanning over 1500 samples, having perturbations in amplitudes (up to 30% of a fixed amplitude) and period (up to 10 % of N_0) and corrupted with additive white Gaussian noise at -10 dB signal to noise ratio (SNR). To account for the random nature of the noise, we consider the mean and standard deviation (SD) of the deviation (σ) of the estimate from the actual location over 1000 noisy realizations of the impulse train. Fig. 1 depicts the five different experiments conducted: (a) exactly periodic impulse train without amplitude perturbation and noise, (b) and (c) are exactly periodic noisy impulse trains without and with amplitude perturbation, respectively. Fig. 1 (d) and (e) are quasi-periodic noisy impulse trains without and with amplitude perturbation, respectively. The impulse locations are estimated without any error for the cases (a), (b) and (c). For the cases (d) and (e), the mean and standard deviation of the σ for all impulse locations are approximately zero and less than five samples, respectively. This result suggests that the perturbation in the amplitudes of the impulses has no effect on the estimation of impulse locations using the CIS whereas the estimation error depends on the extent of fluctuation of the period. Further, in most of the cases there are well-defined peaks in the CIS, at the locations of impulses even at -10 dB SNR.

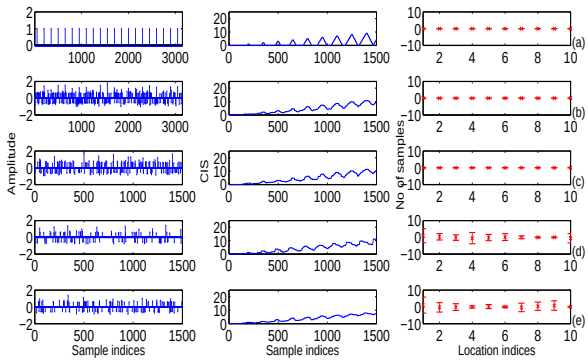

Figure 1. Illustration of the cummulative impulse strength (CIS) (for the cases described in the text of section II. C) of a quasi-periodic impulse train (left panels are the impulse trains, middle panels are the CIS and last panels show the error in the estimated locations)

D. GCI detection using CIS on ILPR

It has been shown that the use of the ILPR is more robust for GCI detection compared to LPR [15], [16]. Since the GCIs manifest as local negative peaks in the ILPR [20], ILPR samples other than the local minima, do not contain information regarding the GCIs. Thus we first consider the inverted ILPR and then convert the inverted ILPR (call it $c[n]$) to a peak-strength sequence $ps[n]$, which is non-zero only at the local maxima of $c[n]$. In $r[n]$, if l_{max} represents the location of a maximum between two successive local minima l_{min^-} and l_{min^+}, the $ps[n]$ at l_{max} is defind as

$$ps[l_{max}] = c[l_{max}]/\sqrt{(|c[l_{min^-}]|) \times (|c[l_{min^+}]|)} \qquad (5)$$

The CIS is computed using the $ps[n]$ of the ILPR to locate the GCIs. Note that, given a speech signal, the computation of the CIS can be initiated at any point in time, in the speech signal. The back tracking algorithm ensures that the peaks picked are the GCIs at the voiced segments and arbitrary locations at the unvoiced segments, that occur post the initialization point. However, in practice, computation of CIS is started at the beginning of the utterance so that the GCIs within the entire utterance are detected. Figure 2 illustrates the workflow of the algorithm on three pitch periods of the inverted ILPR. The search interval (required for back-tracking) for an arbitrary instant n_0 which appears between the final and penultimate GCI locations (n_k and n_{k-1}) is indicated between $n_0 - T_0 - \Delta$ and $n_0 - T_0 + \Delta$. It is seen that once the final GCI is detected, the CIS measure along with the back-tracking function ensures that the previous GCIs are correctly located. Figure 3 illustrates the estimation of GCIs using the proposed method on a segment of the voiced speech corrupted with white Gaussian noise at different SNR levels down to -10 dB. It is seen that the $ps[n]$ serves two purposes: (a) emphasizing the local peaks and (b) reducing the number of locations considered for analysis. The locations of the GCIs are correctly (i.e., there are no misses and false insertions) estimated for all the cases. However, the deviation of the estimated locations from the true locations increases with decreasing SNR.

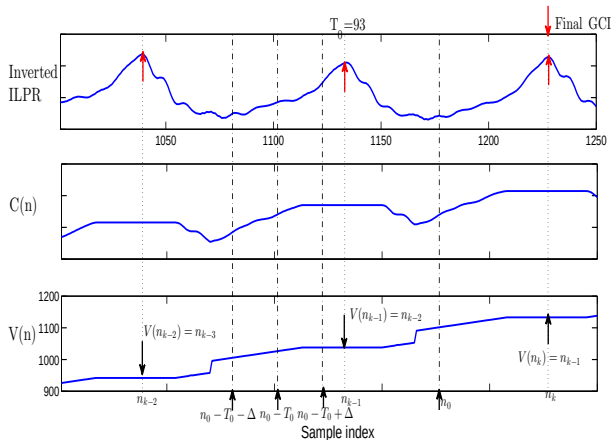

Figure 2. Illustration of the CIS algorithm on three pitch periods of the inverted ILPR. The search interval for computation of CIS for the point n_0 is indicated. Further, the location of the final GCI and the preceeding GCIs as determined from the back tracking using $V(n)$ are also marked.

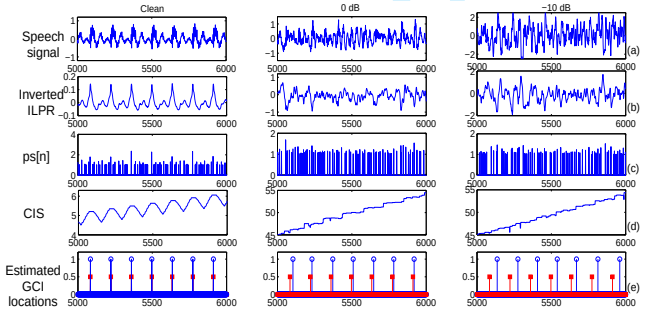

Figure 3. Illustration of the GCI estimation at different noise levels (a) speech signal at different SNRs, (b) inverted ILPR signal, (c) peak strength signal (d) CIS and (e) the estimated (square beads) and actual (circular beads) locations of the GCIs.

III. Experiments and Results

A. Databases and performance measures

The proposed technique is evaluated on two corpora, comprising simultaneous recordings of the speech and the EGG signals - (i) the data provided with the book by D. G. Childers [21], henceforth referred to as the Childers' data. This is recorded from 52 speakers (both male and female) in a single-wall sound room. The Childers' data consists of utterances of 12 sustained vowels, 16 sustained fricatives, an utterance counting one to ten, one counting one to ten with a progressively increasing loudness, singing the musical scale using 'la' and three sentences. In this study, all the speech materials of the Childers' data except the fricative stimuli are used. (ii) a subset of the CMU ARCTIC databases which contain 1132 phonetically balanced sentences. Each of these is a single speaker database corresponding to BDL-US male, JMK-Canadian male and SLT-US female. We use a negative threshold (1/6 of the maximum value [22]) on the dEGG signal to distinguish the voiced from the unvoiced speech. The negative peaks of dEGG provide the ground truth GCIs for validation, which is done only on the voiced speech. We use the standard performance measures of identification rate (IDR), miss rate (MR), false alarm rate (FAR) and the standard deviation of error (SDE) or identification accuracy (IDA) and the accuracy to 0.25 ms (A_{25}) which are illustrated in Fig. 6 of [12]. Experiments are carried out on clean speech and speech degraded with additive white Gaussian and babble noise at SNR 20 to -20 dB in steps of 5 dB. The noise samples are taken from the NOISEX-92 database [23]. We compare the results with four state-of-the-art algorithms: DPI, SEDREAMS, ZFR and DYPSA. The average pitch period required for ZFR, SEDREAMS and CIS are derived from the pitch estimation algorithm [24] (both for clean and noisy speech) and the maximum pitch deviation parameter ρ, is empirically set at 0.3 times the average pitch period[1]. ILPR is estimated by inverse filtering the speech signal (over each disjoint voiced segment), with prediction 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.

Table I
Results of different GCI estimation algorithms on clean speech. The two entries correspond to the results on Childers' data and CMU ARCTIC databases, respectively.

Method	IDR %	SDE in ms	A_{25} %
CIS	95.85, 98.82	0.35, 0.26	82.83, 78.85
DPI	96.12, 99.25	0.38, 0.28	86.09, 80.10
SED	95.47, 99.10	0.35, 0.26	87.81, 77.81
ZFR	94.15, 94.12	0.42, 0.21	52.61, 78.01
DYP	97.01, 98.02	0.43, 0.45	83.28, 71.72

B. Results and discussion

1) Clean speech: Table I summarizes the performance of the five GCI detection algorithms on clean speech. The first entries in Table 1, show that, on Childer's data, the IDR of the CIS method (95.85%) is marginally better than that of the ZFR (94.15%) and SEDREAMS (95.47%), which are based on direct processing of speech signal. However, DYPSA and DPI algorithms have higher IDR because they do not use any APP information and hence GCIs from these algorithms are not affected by the erroneous APP estimates. On the CMU ARCTIC data (second entries in Table 1), all the measures IDR, SDE and A_{25} of the CIS algorithm are comparable to those of the other algorithms. However, as corroborated by the observations made in the previous studies [16], [12], the DPI algorithm and the SEDREAMS are the best in terms of the GCI estimation accuracy on clean speech.

2) Noisy speech: Figures 4 and 5 depict the results of the algorithms on the speech corrupted with additive white Gaussian and babble noise, respectively. In the case of the white Gaussian noise, the IDR, of the CIS method is better than all the algorithms at SNRs between 0 and -15 dB. The accuracy measures namely, SDE and A_{25} are also consistently the lowest and the highest for the CIS method, respectively.

[1]It is experimentally observed that the choice of the value of ρ is not very critical for a wide range of values. Specifically, the IDR varies (on a subset of the database) is about 1 % when ρ varies from 0.1 and 0.45. IDR is maximum for $\rho = 0.3$ and hence this value is used in all further experiments.

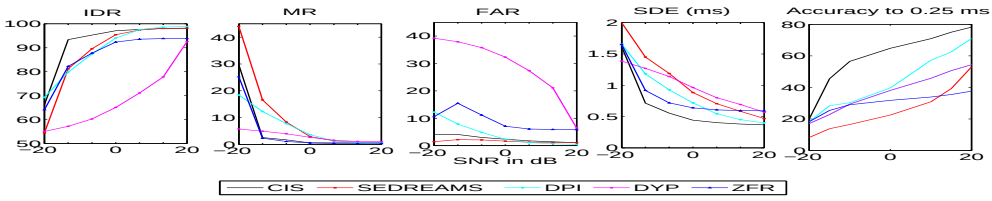

Figure 4. Performance of the five different algorithms averaged over both the databases at different SNRs (-20 to 20 dB) with additive white Gaussian noise.

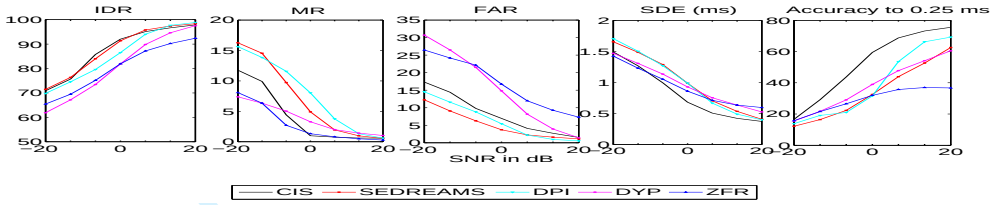

Figure 5. Performance of the five different algorithms averaged over both the databases at different SNRs (-20 to 20 dB) with additive babble noise.

The superior performance of the CIS method may be attributed to the fact that the sequence of CIS uses locations of all the previous impulses to estimate the location of the current impulse in a recursive manner. In the case of the babble noise, the IDR and A_{25} for all the algorithms are worse than those in the case of the white Gaussian noise. This may be due to the speech-like characteristics of the babble noise. The performance of the CIS method is comparable to that of SEDREAMS and ZFR, in terms of IDR. However, CIS performs better than all the other algorithms considered in terms of accuracy measure A_{25}. In summary, for the experiments in clean and noisy conditions, it is observed that the performance of the CIS method is comparable (superior in some cases) to that of all the algorithms examined despite being based on the ILPR. CIS method is found to be superior than the other algorithms which are based on the LPR (DPI and DYPSA) in the presence of noise. It is known that DYPSA algorithm degrades the most with noise. The DPI algorithm, despite using ILPR is comparable to SEDREAMS and ZFR. Based on these experiments, it may be concluded that if the average pitch information is available *a-priori,* then an algorithm based on the linear prediction residual can reach a performance comparable to those based on the speech signal alone in the presence of noise.

3) Dependency on average pitch period: In the earlier sections, it was mentioned that the proposed algorithm, along with ZFR and SEDREMS require the average pitch information a-priori. To quantify the dependency of these algorithms on the accuracy of average pitch value, the IDR obtained with different noisy average pitch estimates on ARCTIC databases is shown in Fig. 6. The base estimate for the average pitch period is obtained using the dEGG signal to ensure that the errors in its computation do no affect the experiments. Subsequently pitch period is varied such that the error between the actual and the estimated pitch periods are in the range of -0.6 to 1.4 (with respect to the actual pitch period) in steps of 0.1. The performance of all the three algorithms degrade with error in average pitch estimate. However, the degradation trends corresponding to different algorithms are slightly different. If the estimated pitch period is less than the actual pitch, the degradation in ZFR is more severe compared to the other two, which are comparable with each other. However ZFR is more robust than the other two if the estimated pitch is more than the actual pitch, with a decrease in IDR from 98 % to just above 85 % when the error in the estimated pitch varies from 0 to 140 % of the actual pitch. SEDREAMS and CIS have their IDR more than 80 % when the estimated pitch is within $\pm$ 35 % of the actual average pitch whereas IDR for ZFR degrades to 50 % if the error in the estimated average picth is -0.2.

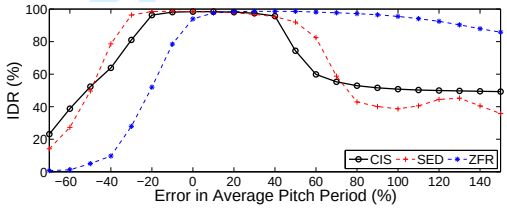

Figure 6. Illustration of dependency of three GCI detection algorithms on average pitch period. The variation in IDR with varying error in average pitch period is shown for the CMU ARCTIC data.

IV. Conclusions

We propose a non-linear measure called the cumulative impulse strength to locate the impulses in a noisy quasi-periodic impulse train. We apply the CIS measure on the ILPR to detect the GCIs of the voiced speech, using an estimate of average pitch period. Experiments with different noisy conditions on data with simultaneous speeech and EGG data reveal that the CIS method is comparable to the best state-of-the-art algorithms indicating its robustness to noise despite operating on the linear prediction residual.

REFERENCES

[1] D. Wong, J. Markel, and A. Gray Jr, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[2] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[3] V. R. Lakkavalli, P. Arulmozhi, and A. G. Ramakrishnan, "Continuity metric for unit selection based text-to-speech synthesis," in *Signal Processing and Communications (SPCOM), 2010 International Conference on*. IEEE, 2010, pp. 1–5.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.

[5] M. R. Shanker, R. Muralishankar, and A. G. Ramakrishnan, "Bauer method of MVDR spectral factorization for pitch modification in the source domain," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, 2007, pp. 263–266.

[6] R. Muralishankar, M. Ravi Shanker, and A. G. Ramakrishnan, "Perceptual-MVDR based analysis-synthesis of pitch synchronous frames for pitch modification," in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 81–84.

[7] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Communication*, vol. 42, no. 2, pp. 143–154, 2004.

[8] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 460–471, 2014.

[9] 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, no. 5, pp. 569–586, 1999.

[10] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *J. Acoust. Soc. Amer. EL*, vol. 137, p. EL469, 2015.

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

[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, no. 3, pp. 994–1006, Mar. 2012.

[13] 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.

[14] 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.

[15] 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.

[16] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, no-12, pp. 2471–2480, Dec. 2013.

[17] V. R. L., G. K.V., H. S, A. G. Ramakrishnan, and T. Ananthapadmanabha, "Subband analysis of linear prediction residual for the estimation of glottal closure instants," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 945–949.

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

[19] 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.

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

[21] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. Wiley, Newyork, 2000.

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

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

[24] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 333–337.

Cumulative Impulse Strength for Epoch Extraction

Prathosh A. P., *Member, IEEE* Sujith P, Ramakrishnan A. G. , *Senior Member, IEEE* and Prasanta Kumar Ghosh, *Senior Member, IEEE*

Abstract

Algorithms for extracting epochs or glottal closure instants (GCIs) from voiced speech typically fall into two categories: (i) ones which operate on linear prediction residual (LPR) and (ii) those which operate directly on the speech signal. While the former class of algorithms (such as YAGA and DPI) tend to be more accurate, the latter ones (such as ZFR and SEDREAMS) tend to be more noise-robust. In this paper, a temporal measure termed the cumulative impulse strength is proposed for locating the impulses in a quasi-periodic impulse-sequence embedded in noise. Subsequently, it is applied for detecting the GCIs from the inverted integrated LPR using a recursive algorithm. Experiments on two large corpora of speech with simultaneous electroglottographic recordings demonstrate that the proposed method is more robust to additive noise than the state-of-the-art algorithms, despite operating on the LPR.

Index Terms

GCI detection, epoch extraction, cumulative impulse strength, impulse tracking.

I. Introduction

Pitch-synchronous analysis of the voiced speech signal is a popular technique in which the glottal closure instants (GCIs or epochs) are used to define the analysis frames. Epochs are utilized in various applications including pitch tracking, voice source estimation [1], speech synthesis [2], [3], prosody modification [4], [5], [6], [7], voiced/unvoiced boundary detection [8] and speaker identification [9], [10]. Hence, automatic detection of the GCIs from the voiced speech signal is considered to be an important problem in speech research. Comprehensive reviews of the importance of the GCI detection problem and summary of the state-of-the-art algorithms may be found in [11], [12].

Many of the popular GCI detectors can be categorized into two classes. Detectors belonging to the first class adhere to the source-filter model of speech production and locate GCIs from an estimate of the glottal source signal such as linear prediction residual (LPR) and the voice source (VS) signal. Algorithms like Hilbert Envelope (HE) based epoch extractors [13], Dynamic Programming Phase Slope Algorithm (DYPSA) [14], Yet Another GCI

Prathosh is with Xerox research center India, Sujith is with Ittiam systems India, and the other authors are with Indian Institute of Science, Bangalore - 560012, India. (e-mail: prathosh.AP@xerox.com, sujith.p4@gmail.com, ramkiag@ee.iisc.ernet.in, prasantg@ee.iisc.ernet.in.)

Algorithm (YAGA) [15], Dynamic Plosion Index (DPI) [16] and sub-band decomposition method [17] fall into this category. The second class of algorithms such as Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) [18] and Zero-frequency resonator (ZFR) [19] operate directly on the speech signal without any model assumption or deconvolution. The former class of algorithms are more accurate than the latter ones [12]. This may be because the GCIs are associated with the source signal, which forms the basis for the analysis for these algorithms. However, they are believed to be more susceptible to noise compared to SEDREAMS and ZFR, mainly because of inaccurate estimation of the LPR in the presence of noise. Further, ZFR and SEDREAMS assume that the average pitch period (APP) is known a priori while the former class of algorithms do not require the information of APP. Motivated by these observations, in this paper, we explore whether an LPR based GCI detection scheme could be noise robust if the APP can be estimated *a-priori*. Specifically, we propose a generic measure named the cumulative impulse strength (CIS) to locate the impulses in a quasi-periodic impulse train corrupted by additive noise. Further, using CIS, we devise a recursive algorithm to extract GCIs from the integrated LPR (ILPR) [16] of the voiced speech and evaluate the proposed algorithm using two speech databases with simultaneous electroglottographic (EGG) recordings in both clean and noisy conditions.

II. Impulse-location detection using CIS

A. Motivation

It is known that the GCIs coincide with the local negative peaks of the voice source signal [20]. Thus, a GCI extraction algorithm which uses the voice source signal typically involves two stages - (i) transformation of the speech signal into a domain where the voice source signal is best represented (such as ILPR), (ii) accurately picking the peaks corresponding to GCIs from the transformed signal. To reduce the error committed by the peak-picking algorithm, the temporal quasi-periodicity property of the voiced speech can be exploited. In a quasi-periodic impulse-train like sequence, the accuracy of detection of each impulse could be improved by using the knowledge of the location and the strength of the previous impulses. That is, the impulse-like behavior at a given instant of time may be determined not only by analyzing some local properties of the signal around that instant but also by taking into account the global behavior of the signal around all the previous impulse locations. Based on this intuition, we define a measure named the cumulative impulse strength to estimate the locations of the impulses in a quasi-periodic impulse train.

B. Cumulative impulse strength

Let $r[n]$ be an amplitude-perturbed, quasi-periodic impulse train of length N represented as follows:

$$r[n] = \sum_{k=1}^{N} A_k \delta[n - n_k], \tag{1}$$

$$n_k = n_{k-1} + N_0 + \Delta_k, \ 2 \leq k \leq N. \tag{2}$$

where n_k is the location of the k-th impulse with amplitude A_k, $\delta[n - n_k]$ denotes the Kronecker delta function, N_0 is the average period of $r[n]$ and Δ_k is the deviation of $n_k - n_{k-1}$ from N_0. The measure CIS is defined recursively at each location n, by combining the effect of the signal r and the CIS C around the previous impulse location. That is, if $\rho = \max_k |\Delta_k|$, the CIS $C[n]$ at the n-th sample is defined as follows:

$$C[n] = \max_{n-N_0-\rho \leq m \leq n-N_0+\rho} \big(C[m] + r[m]\big) \tag{3}$$

In order to locate the impluses from $C[n]$, we define one more sequence $V[n]$ as follows.

$$V[n] = \operatorname*{argmax}_{n-N_0-\rho \leq m \leq n-N_0+\rho} \big(C[m] + r[m]\big). \tag{4}$$

That is, at each sample n, $V[n]$ stores the location that maximizes $C[n]$ within the search interval defined in Eq. 3. Once the location of the last impulse is known, a back tracking procedure is employed to locate all the impulses from $V[n]$ as follows: if n_k corresponds to the k^{th} impulse location, the $(k-1)^{th}$ impulse location is given by $V[n_k]$. The location of the final impulse is defined to be that which maximizes $r[m]$, $N-1-N_0+\rho \leq m \leq N-1$. This is because the location of the maxima of the $r[m]$ within the last periodic interval corresponds to the final impulse.

C. Illustration of CIS on synthetic data

In this section we report an experiment where the objective is to estimate the locations of the impulses using the CIS, from an impulse train (N_0=150) of 10 impulses spanning over 1500 samples, having perturbations in amplitudes (up to 30% of a fixed amplitude) and period (up to 10 % of N_0) and corrupted with additive white Gaussian noise at -10 dB signal to noise ratio (SNR). To account for the random nature of the noise, we consider the mean and standard deviation (SD) of the deviation (σ) of the estimate from the actual location over 1000 noisy

realizations of the impulse train. Fig. 1 depicts the five different experiments conducted: (a) exactly periodic impulse train without amplitude perturbation and noise, (b) and (c) are exactly periodic noisy impulse trains without and with amplitude perturbation, respectively. Fig. 1 (d) and (e) are quasi-periodic noisy impulse trains without and with amplitude perturbation, respectively. The impulse locations are estimated without any error for the cases (a), (b) and (c). For the cases (d) and (e), the mean and standard deviation of the σ for all impulse locations are approximately zero and less than five samples, respectively. This result suggests that the perturbation in the amplitudes of the impulses has no effect on the estimation of impulse locations using the CIS whereas the estimation error depends on the extent of fluctuation of the period. Further, in most of the cases there are well-defined peaks in the CIS, at the locations of impulses even at -10 dB SNR.

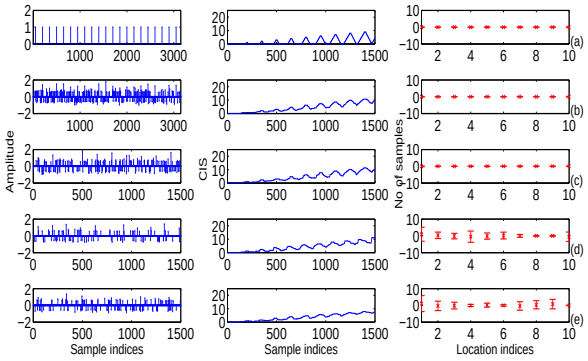

Figure 1. Illustration of the cummulative impulse strength (CIS) (for the cases described in the text of section II. C) of a quasi-periodic impulse train (left panels are the impulse trains, middle panels are the CIS and last panels show the error in the estimated locations)

D. GCI detection using CIS on ILPR

It has been shown that the use of the ILPR is more robust for GCI detection compared to LPR [15], [16]. Since the GCIs manifest as local negative peaks in the ILPR [20], ILPR samples other than the local minima, do not contain information regarding the GCIs. Thus we first consider the inverted ILPR and then convert the inverted ILPR (call it $c[n]$) to a peak-strength sequence $ps[n]$, which is non-zero only at the local maxima of $c[n]$. In $r[n]$, if l_{max} represents the location of a maximum between two successive local minima l_{min^-} and l_{min^+}, the $ps[n]$ at l_{max} is defind as

$$ps[l_{max}] = c[l_{max}]/\sqrt{(|c[l_{min^-}]|) \times (|c[l_{min^+}]|)} \tag{5}$$

The CIS is computed using the $ps[n]$ of the ILPR to locate the GCIs. Note that, given a speech signal, the computation of the CIS can be initiated at any point in time, in the speech signal. The back tracking algorithm ensures that the peaks picked are the GCIs at the voiced segments and arbitrary locations at the unvoiced segments, that occur post the initialization point. However, in practice, computation of CIS is started at the beginning of the

utterance so that the GCIs within the entire utterance are detected. Figure 2 illustrates the workflow of the algorithm on three pitch periods of the inverted ILPR. The search interval (required for back-tracking) for an arbitrary instant n_0 which appears between the final and penultimate GCI locations (n_k and n_{k-1}) is indicated between $n_0 - T_0 - \Delta$ and $n_0 - T_0 + \Delta$. It is seen that once the final GCI is detected, the CIS measure along with the back-tracking function ensures that the previous GCIs are correctly located. Figure 3 illustrates the estimation of GCIs using the

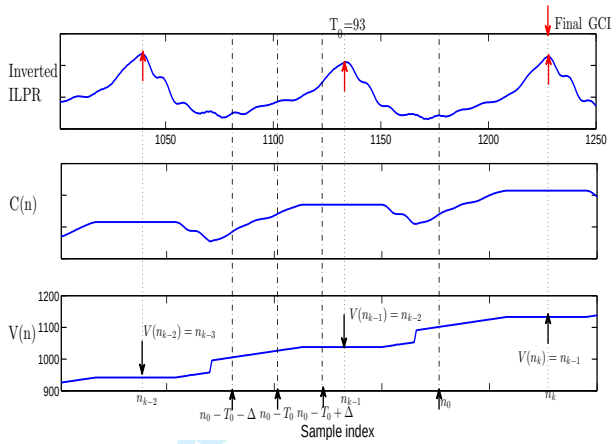

Figure 2. Illustration of the CIS algorithm on three pitch periods of the inverted ILPR. The search interval for computation of CIS for the point n_0 is indicated. Further, the location of the final GCI and the preceeding GCIs as determined from the back tracking using $V(n)$ are also marked.

proposed method on a segment of the voiced speech corrupted with white Gaussian noise at different SNR levels down to -10 dB. It is seen that the $ps[n]$ serves two purposes: (a) emphasizing the local peaks and (b) reducing the number of locations considered for analysis. The locations of the GCIs are correctly (i.e., there are no misses and false insertions) estimated for all the cases. However, the deviation of the estimated locations from the true locations increases with decreasing SNR.

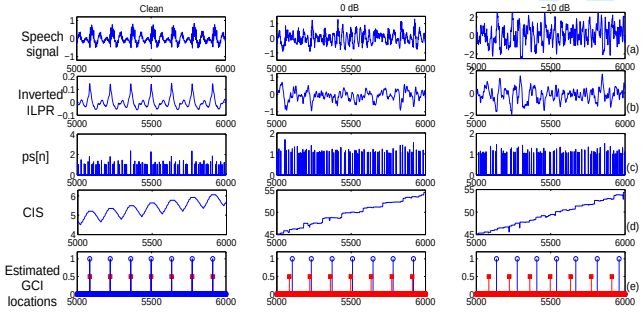

Figure 3. Illustration of the GCI estimation at different noise levels (a) speech signal at different SNRs, (b) inverted ILPR signal, (c) peak strength signal (d) CIS and (e) the estimated (square beads) and actual (circular beads) locations of the GCIs.

III. Experiments and results

A. Databases and performance measures

The proposed technique is evaluated on two corpora, comprising simultaneous recordings of the speech and the EGG signals - (i) the data provided with the book by D. G. Childers [21], henceforth referred to as the Childers' data. This is recorded from 52 speakers (both male and female) in a single-wall sound room. The Childers' data consists of utterances of 12 sustained vowels, 16 sustained fricatives, an utterance counting one to ten, one counting one to ten with a progressively increasing loudness, singing the musical scale using 'la' and three sentences. In this study, all the speech materials of the Childers' data except the fricative stimuli are used. (ii) a subset of the CMU ARCTIC databases which contain 1132 phonetically balanced sentences. Each of these is a single speaker database corresponding to BDL-US male, JMK-Canadian male and SLT-US female. We use a negative threshold (1/6 of the maximum value [22]) on the dEGG signal to distinguish the voiced from the unvoiced speech. The negative peaks of dEGG provide the ground truth GCIs for validation, which is done only on the voiced speech. We use the standard performance measures of identification rate (IDR), miss rate (MR), false alarm rate (FAR) and the standard deviation of error (SDE) or identification accuracy (IDA) and the accuracy to 0.25 ms (A_{25}) which are illustrated in Fig. 6 of [12]. Experiments are carried out on clean speech and speech degraded with additive white Gaussian and babble noise at SNR 20 to -20 dB in steps of 5 dB. The noise samples are taken from the NOISEX-92 database [23]. We compare the results with four state-of-the-art algorithms: DPI, SEDREAMS, ZFR and DYPSA. The average pitch period required for ZFR, SEDREAMS and CIS are derived from the pitch estimation algorithm [24] (both for clean and noisy speech) and the maximum pitch deviation parameter ρ, is empirically set at 0.3 times the average pitch period[1]. ILPR is estimated by inverse filtering the speech signal (over each disjoint voiced segment), with prediction 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.

Table I
Results of different GCI estimation algorithms on clean speech. The two entries correspond to the results on Childers' data and CMU ARCTIC databases, respectively.

Method	IDR %	SDE in ms	A_{25} %
CIS	95.85, 98.82	0.35, 0.26	82.83, 78.85
DPI	96.12, 99.25	0.38, 0.28	86.09, 80.10
SED	95.47, 99.10	0.35, 0.26	87.81, 77.81
ZFR	94.15, 94.12	0.42, 0.21	52.61, 78.01
DYP	97.01, 98.02	0.43, 0.45	83.28, 71.72

[1]It is experimentally observed that the choice of the value of ρ is not very critical for a wide range of values. Specifically, the IDR varies (on a subset of the database) is about 1 % when ρ varies from 0.1 and 0.45. IDR is maximum for $\rho = 0.3$ and hence this value is used in all further experiments.

B. Results and discussion

1) Clean speech: Table I summarizes the performance of the five GCI detection algorithms on clean speech. The first entries in Table 1, show that, on Childer's data, the IDR of the CIS method (95.85%) is marginally better than that of the ZFR (94.15%) and SEDREAMS (95.47%), which are based on direct processing of speech signal. However, DYPSA and DPI algorithms have higher IDR because they do not use any APP information and hence GCIs from these algorithms are not affected by the erroneous APP estimates. On the CMU ARCTIC data (second entries in Table 1), all the measures IDR, SDE and A_{25} of the CIS algorithm are comparable to those of the other algorithms. However, as corroborated by the observations made in the previous studies [16], [12], the DPI

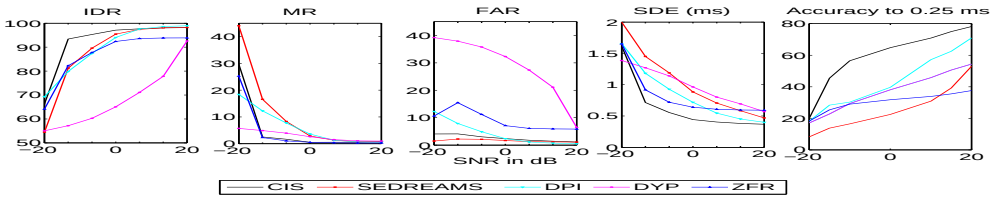

Figure 4. Performance of the five different algorithms averaged over both the databases at different SNRs (-20 to 20 dB) with additive white Gaussian noise.

algorithm and the SEDREAMS are the best in terms of the GCI estimation accuracy on clean speech.

2) Noisy speech: Figures 4 and 5 depict the results of the algorithms on the speech corrupted with additive white Gaussian and babble noise, respectively. In the case of the white Gaussian noise, the IDR, of the CIS method is better than all the algorithms at SNRs between 0 and -15 dB. The accuracy measures namely, SDE and A_{25} are also consistently the lowest and the highest for the CIS method, respectively. The superior performance of the CIS method may be attributed to the fact that the sequence of CIS uses locations of all the previous impulses to estimate the location of the current impulse in a recursive manner. In the case of the babble noise, the IDR and A_{25} for all the algorithms are worse than those in the case of the white Gaussian noise. This may be due to the speech-like characteristics of the babble noise. The performance of the CIS method is comparable to that of SEDREAMS and ZFR, in terms of IDR. However, CIS performs better than all the other algorithms considered in terms of accuracy measure A_{25}. In summary, for the experiments in clean and noisy conditions, it is observed that the performance of the CIS method is comparable (superior in some cases) to that of all the algorithms examined despite being based on the ILPR. CIS method is found to be superior than the other algorithms which are based on the LPR (DPI and DYPSA) in the presence of noise. It is known that DYPSA algorithm degrades the most with noise. The DPI algorithm, despite using ILPR is comparable to SEDREAMS and ZFR. Based on these experiments, it may be concluded that if the average pitch information is available *a-priori,* then an algorithm based on the linear prediction residual can reach a performance comparable to those based on the speech signal alone in the presence

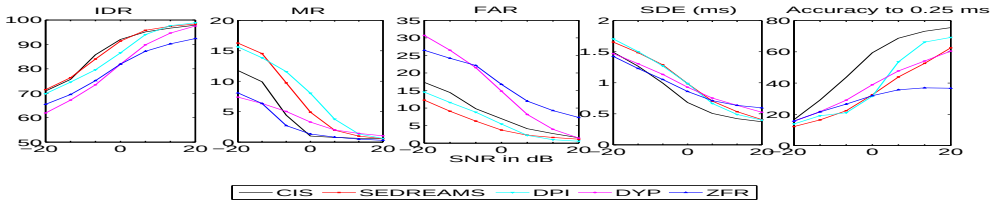

Figure 5. Performance of the five different algorithms averaged over both the databases at different SNRs (-20 to 20 dB) with additive babble noise.

of noise.

3) Dependency on average pitch period: In the earlier sections, it was mentioned that the proposed algorithm, along with ZFR and SEDREMS require the average pitch information a-priori. To quantify the dependency of these algorithms on the accuracy of average pitch value, the IDR obtained with different noisy average pitch estimates on ARCTIC databases is shown in Fig. 6. The base estimate for the average pitch period is obtained using the dEGG signal to ensure that the errors in its computation do no affect the experiments. Subsequently pitch period is varied such that the error between the actual and the estimated pitch periods are in the range of -0.6 to 1.4 (with respect to the actual pitch period) in steps of 0.1. The performance of all the three algorithms degrade with error in average pitch estimate. However, the degradation trends corresponding to different algorithms are slightly different. If the estimated pitch period is less than the actual pitch, the degradation in ZFR is more severe compared to the other two, which are comparable with each other. However ZFR is more robust than the other two if the estimated pitch is more than the actual pitch, with a decrease in IDR from 98 % to just above 85 % when the error in the estimated pitch varies from 0 to 140 % of the actual pitch. SEDREAMS and CIS have their IDR more than 80 % when the estimated pitch is within $\pm$ 35 % of the actual average pitch whereas IDR for ZFR degrades to 50 % if the error in the estimated average picth is -0.2.

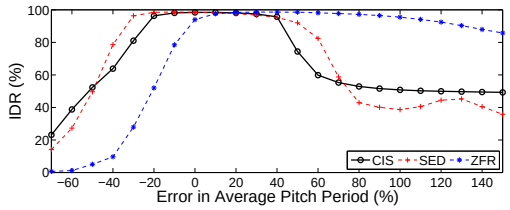

Figure 6. Illustration of dependency of three GCI detection algorithms on average pitch period. The variation in IDR with varying error in average pitch period is shown for the CMU ARCTIC data.

IV. Conclusions

We propose a non-linear measure called the cumulative impulse strength to locate the impulses in a noisy quasi-periodic impulse train. We apply the CIS measure on the ILPR to detect the GCIs of the voiced speech, using an estimate of average pitch period. Experiments with different noisy conditions on data with simultaneous speeech and

EGG data reveal that the CIS method is comparable to the best state-of-the-art algorithms indicating its robustness to noise despite operating on the linear prediction residual.

REFERENCES

[1] D. Wong, J. Markel, and A. Gray Jr, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[2] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[3] V. R. Lakkavalli, P. Arulmozhi, and A. G. Ramakrishnan, "Continuity metric for unit selection based text-to-speech synthesis," in *Signal Processing and Communications (SPCOM), 2010 International Conference on*. IEEE, 2010, pp. 1–5.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.

[5] M. R. Shanker, R. Muralishankar, and A. G. Ramakrishnan, "Bauer method of MVDR spectral factorization for pitch modification in the source domain," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, 2007, pp. 263–266.

[6] R. Muralishankar, M. Ravi Shanker, and A. G. Ramakrishnan, "Perceptual-MVDR based analysis-synthesis of pitch synchronous frames for pitch modification," in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 81–84.

[7] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Communication*, vol. 42, no. 2, pp. 143–154, 2004.

[8] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 460–471, 2014.

[9] 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, no. 5, pp. 569–586, 1999.

[10] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *J. Acoust. Soc. Amer. EL*, vol. 137, p. EL469, 2015.

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

[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, no. 3, pp. 994–1006, Mar. 2012.

[13] 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.

[14] 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.

[15] 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.

[16] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, no-12, pp. 2471–2480, Dec. 2013.

[17] V. R. L., G. K.V., H. S, A. G. Ramakrishnan, and T. Ananthapadmanabha, "Subband analysis of linear prediction residual for the estimation of glottal closure instants," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 945–949.

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

[19] 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.

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

[21] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. Wiley, Newyork, 2000.

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

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

[24] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 333–337.