Kadiri, Sudarsana; Alku, Paavo; Yegnanarayana, Bayya

Comparison of glottal closure instants detection algorithms for emotional speech

Published in:
2020 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020 - Proceedings

DOI:
10.1109/ICASSP40776.2020.9054737

Published: 01/05/2020

Document Version
Peer reviewed version

Please cite the original version:
Kadiri, S., Alku, P., & Yegnanarayana, B. (2020). Comparison of glottal closure instants detection algorithms for emotional speech. In *2020 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020 - Proceedings* (pp. 7379-7383). [9054737] (Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9054737

COMPARISON OF GLOTTAL CLOSURE INSTANTS DETECTION ALGORITHMS FOR EMOTIONAL SPEECH

Sudarsana Reddy Kadiri[1]*, Paavo Alku*[1] *and B. Yegnanarayana*[2]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland.
[2]Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India.
{sudarsana.kadiri, paavo.alku}@aalto.fi; yegna@iiit.ac.in

ABSTRACT

In production of voiced speech, epochs or glottal closure instants (GCIs) refer to the instants of significant excitation of the vocal tract. Extraction of GCIs is used as a pre-processing stage in many areas of speech technology, such as in prosody modification, speech synthesis and voice source analysis. In the past decades, several GCI detection algorithms have been developed and most of them provide excellent results for speech signals produced using modal (normal) type of phonation. There are, however, no studies comparing multiple state-of-the-art GCI detection methods in emotional speech. In this paper, we compare six GCI detection algorithms using emotional speech and known evaluation metrics. We use the Berlin EMO-DB acted emotional speech database which contains seven emotions and simultaneous electroglottography (EGG) recordings as ground truth. The results show that all six GCI detection algorithms give best performance in processing speech of neutral emotion and that the performance degrade particularly in emotions of high arousal (anger and joy). To improve the performance of GCI detection in emotional speech, the study underlines the importance of local average pitch period estimates.

Index Terms— Speech analysis, Excitation source, Epochs, Glottal Closure Instants, Emotions.

1. INTRODUCTION

Epochs or glottal closure instants (GCIs) refer to the instants of significant excitation of the vocal tract system in voiced speech [1]. GCI detection is needed as a pre-processing stage in many areas of speech technology, for example, in prosody modification [2], voice source analysis [3–5], concatenative speech synthesis [6], parametric speech synthesis [7] and detection of pathological speech [8, 9]. Many GCI detection algorithms have been developed in the past decades (for review, see [10, 11]). These algorithms have been typically evaluated using speech utterances produced in modal (normal) phonation collected in laboratory environments. Robustness of GCI detection has also been studied in degraded conditions using speech corrupted with additive noise and reverberations [10], as well as using telephone quality speech [12].

Due to the prominent excitation of the vocal tract system at GCIs, speech signal waveforms show high signal-to-noise ratios (SNRs) in temporal regions after GCIs. This phenomenon has been taken advantage of in many speech technology applications [1]. The processing of speech in high SNR regions, identified with GCI detection methods, has been studied, for example, in glottal activity detection [1], pitch extraction [13], estimation of formant frequencies [14], speaker recognition [15], speech enhancement [1], multi-speaker separation, and estimation of number of speakers from multi-speaker data [16]. Recently, the importance of features extracted around GCIs has been observed in several studies in emotional speech analysis and detection [17–19]. The importance of frequency of vibration of vocal folds (interval between GCIs) was studied in emotion detection and emotion conversion/synthesis in [18, 20]. In addition, some speech synthesis systems (e.g., [21–23]) use voice source modeling and GCI detection in generation of the synthesizer's excitation signal. These synthesis techniques can be adapted to different voice qualities or expressive voice thereby emphasizing the importance of GCI detection in synthesis of emotional speech.

Since analysis and synthesis of emotional speech benefits from accurate detection of GCI locations, the aim of this study is to compare the performance of several state-of-the-art GCI detection algorithms in emotional speech. GCI detection algorithms have not been previously compared using emotional speech or speech of different voice qualities, except in a few investigations [11, 24]. In [11], six GCI detection methods were compared using speech of six different voice qualities. In [24], two algorithms were studied in GCI detection of emotional speech from two databases. Accuracy of the existing algorithms may vary in processing of emotional speech because of extensive variations in the glottal source characteristics (e.g., in the strength of the excitation, in the amount of aspiration noise, in the pitch range) between different vocal emotions.

The organization of the paper is as follows. Section 2 describes the GCI detection algorithms used for comparison in this study. Section 3 describes the emotional speech database and the evaluation metrics used. The results of the experiments are presented in Section 4 along with their discussion. Finally, Section 5 gives a summary of the study.

2. GCI DETECTION METHODS

The following state-of-art GCI detection algorithms are compared: the *zero frequency filtering* (ZFF) method [25], the *dynamic programming phase slope* (DYPSA) method [26], the *yet another GCI algorithm* (YAGA) method [27], the *speech event detection using the residual excitation and a mean based signal* (SEDREAMS) method [10], the *SEDREAMS algorithm modified to handle various voice qualities* (SE-VQ) method [11] and the *microcanonical multi-scale formalism* (MMF) method [28]. These six algorithms are briefly described in this section.

2.1. ZFF

The ZFF method [25] is based on the observation that impulsive nature of the glottal excitation is reflected across all frequencies in-

cluding the zero frequency (0 Hz). Hence, the GCI locations can be detected by confining the analysis around 0 Hz. The following steps are involved in the ZFF method.

1. The speech signal $s[n]$ is differentiated to remove any low-frequency bias present in the signal.

$$x[n] = s[n] - s[n-1]. \quad (1)$$

2. The signal $x[n]$ is passed through a cascade of two ideal zero-frequency resonators, given by:

$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n], \quad (2)$$

where $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$. The resulting signal $y_0[n]$ grows/decays approximately as a polynomial function of time.

3. The trend in $y_0[n]$ is removed by subtracting the local mean computed over the average pitch period (computed using the autocorrelation function in 30-ms segments of $x[n]$), at each sample. The resulting signal is given by

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m]. \quad (3)$$

The signal defined in Eq. (3) is the zero frequency filtered signal. Here $2N+1$ corresponds to the number of samples used for trend removal. The instants of negative-to-positive zero crossings (NPZCs) correspond to GCIs by considering the positive polarity of the signal [25,29]. If the speech signal is reversed in polarity, the signal has to be negated before conducting ZFF analysis [30].

2.2. DYPSA

The DYPSA algorithm [26] estimates GCIs by using linear prediction (LP) residual signals. The algorithm consists of three components. The first component generates candidate GCIs using zero crossings of the phase slope function, which is calculated from the LP residual. The energy-weighted group delay is used as a measure to derive the phase slope function. The second component uses a phase slope projection technique to recover candidates for which the phase slope function does not include zero crossings. The above mentioned two components detect almost all the true GCIs, but they also generate a large number of false candidates. Therefore, the third component of DYPSA uses dynamic programming to identify the true GCIs from the hypothesized candidates by minimizing a cost function. The cost function consists of five elements: inter-pulse similarity, pitch deviation, costs derived from the projected phase slope, normalized energy values and deviations from an ideal phase slope function. Each of the five elements are weighted with constant values given in [26]. The MATLAB implementation of DYPSA available in VOICEBOX [31] is used in the current study.

2.3. YAGA

The YAGA algorithm [27] combines several methods used in the GCI detection including wavelet analysis, group delay analysis and M-best dynamic programming. In YAGA, the GCI candidates are detected by first estimating the voice source signal using iterative adaptive inverse filtering (IAIF) [3]. In order to highlight the discontinuities in the voice source signal, the multi-scale product of the stationary wavelet transform is utilized by using information across the wavelet scales. The discontinuities are detected using the group delay function, and the GCI candidates are measured as negative-going zero crossings. The falsely detected GCIs are removed using M-best dynamic programming in a similar way as in DYPSA [26]. The YAGA method can also be used in the detection of glottal opening instants (GOIs). YAGA uses cost elements that are similar to those in DYPSA, with modifications in the inter-pulse similarity cost and in the use of an additional element to distinguish GCIs and GOIs. In the current study, the MATLAB implementation of the YAGA algorithm obtained from the authors of [26] is used.

2.4. SEDREAMS

SEDREAMS [10] uses a mean-based signal and LP residual for the detection of GCIs. The algorithm consists of two steps. In the first step, a mean-based signal, inspired by [25], is computed from the speech signal using a simple moving average operation and a window function (e.g., Blackman window). By searching for the minima and maxima of the mean-based signal, short intervals where GCIs are expected to occur are defined. In the second step, refined GCIs are detected in the intervals determined in the first step by searching for the instants where the LP residual reaches its maximum value. The MATLAB implementation of SEDREAMS available in COVAREP [32] is used in the current study.

2.5. SE-VQ

The SE-VQ algorithm [11] is a modification of SEDREAMS. SE-VQ applies a dynamic programming method to select the optimal path of GCIs. This is based on the strength of LP residual peaks and on a transition cost, which takes care of transition from one GCI to another. A finer post-processing is included to remove false positives, while at the same time preserving true positive GCIs. For this, the F_0 filtered LP residual is used. The MATLAB implementation of the SE-VQ algorithm available in COVAREP [32] is used in the current study.

2.6. MMF

MMF [28] estimates a multi-scale parameter (singularity exponent) at each instant in the signal domain. This parameter quantifies the degree of signal singularity at each sample from a multi-scale point of view and its value can be used to classify signal samples accordingly. The subset of samples with lowest singularity exponent values points to a GCI. This property is used in the MMF algorithm with an appropriate estimation of singularity exponents. In the singularity exponent values, GCIs attain smaller values compared to its surrounding signal samples in one fundamental period. However, in a wider neighborhood, they show higher values compared to non-GCI samples. Hence, the application of a global threshold for the whole segment cannot refine GCIs from non-GCIs. To overcome this drawback, sudden fall in singularity exponent values right before each GCI and local averages of singularity exponents before and after each GCI are used. A MATLAB implementation of the MMF algorithm available in [28] is used in the current study.

3. EXPERIMENTS

3.1. Database and ground truth

The Berlin emotion speech database (EMO-DB) [33] was used to compare the selected six GCI detection methods. The database consists of about 800 emotional utterances spoken by ten professional

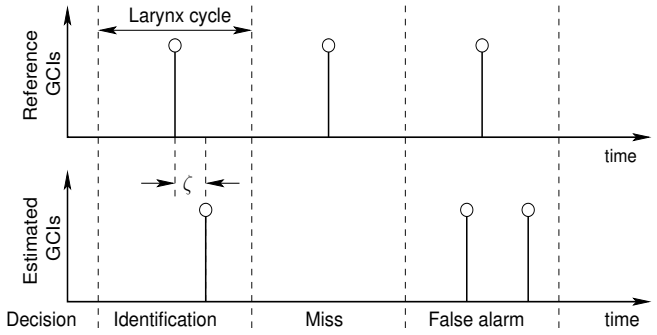

Fig. 1. Characterization of GCI/epoch estimates showing 3 larynx cycles with examples of each possible outcome from GCI detection [25, 26].

native German actors (5 males, 5 females). The data consists of 7 different emotions (neutral, anger, fear, joy, sadness, disgust and boredom) recorded in one or more sessions in an anechoic chamber. The duration of each utterance is approximately 3 sec. In addition to speech, the database also contains the simultaneously recorded electroglottography (EGG) signals. The speech and EGG signals were time-aligned to compensate for the larynx-to-microphone delay. The data was originally sampled at 48 kHz, but it was downsampled to 16 kHz in the present study. Reference GCIs (i.e., the ground truth) were determined from the voiced segments of the EGG signals by searching for the peaks of the differentiated EGG (dEGG) signal. In this study, the ground truth GCIs are obtained from the dEGG signal, whose values are above 10% of the maximum dEGG value of the entire signal. The performance of the algorithms was evaluated only in the voiced segments (detected from EGG) between the reference GCIs and the estimated GCIs.

3.2. Evaluation measures

The performance of the GCI detection algorithms was evaluated using measures utilized in [26]. These measures, demonstrated in Fig. 1, are defined as follows:

(a) *Identification rate* (IDR): The percentage of larynx cycles for which exactly one GCI is detected.

(b) *Miss rate* (MR): The percentage of larynx cycles for which no GCI is detected.

(c) *False alarm rate* (FAR): The percentage of larynx cycles for which more than one GCI is detected.

(d) *Identification accuracy (IDA):* The standard deviation of the timing between the reference GCI and the detected GCI in larynx cycles, for which exactly one GCI was detected.

(e) *IDR within* $\pm 0.25\ msec$: The percentage of GCIs detected within $\pm 0.25\ msec$ to the reference GCIs.

Measures (a)–(c) are collectively called the reliability measures, and measures (d) and (e) are called the accuracy measures.

4. RESULTS AND DISCUSSION

The evaluation metrics calculated separately for each emotion are presented in Table 1. The metrics averaged over all other six emotions except neutral (i.e., anger, fear, joy, sadness, disgust, boredom)

Table 1. Performance of the six GCI detection methods for emotions of the EMO-DB database. IDR - Identification rate, MR - Miss rate, FAR - False alarm rate, IDA - Identification accuracy in ms, and IDR(± 0.25 ms) - IDR within ± 0.25 ms (%).

Emotion	Method	IDR%	MR %	FAR %	IDA (ms)	IDR(± 0.25 ms)
Neutral	ZFF	**96.41**	**00.21**	**03.38**	00.31	55.36
	YAGA	94.84	00.32	04.84	**00.21**	**75.38**
	DYPSA	92.16	02.59	05.24	00.33	67.16
	SEDREAMS	95.01	00.59	04.40	00.29	68.13
	SE-VQ	95.30	00.30	04.40	00.26	67.92
	MMF	88.61	06.48	04.91	00.33	68.47
Anger	ZFF	88.25	**00.41**	11.33	00.31	58.45
	YAGA	**89.72**	01.44	08.84	00.35	57.47
	DYPSA	85.97	07.23	06.80	**00.26**	**61.90**
	SEDREAMS	88.76	00.84	10.40	00.38	58.14
	SE-VQ	88.55	00.49	10.96	00.37	59.46
	MMF	59.32	37.25	**03.43**	00.28	41.15
Fear	ZFF	**92.23**	00.18	07.59	**00.31**	54.80
	YAGA	90.74	00.81	08.45	00.35	59.41
	DYPSA	86.84	05.66	07.50	00.34	**59.93**
	SEDREAMS	91.57	00.36	08.07	00.37	56.17
	SE-VQ	92.19	**00.12**	07.69	00.39	54.25
	MMF	66.00	30.26	**03.74**	00.37	41.64
Joy	ZFF	**88.61**	**00.47**	10.92	00.32	53.49
	YAGA	87.96	01.46	10.58	00.32	58.18
	DYPSA	85.21	06.29	08.49	**00.27**	**64.86**
	SEDREAMS	88.20	00.86	10.94	00.35	55.58
	SE-VQ	88.54	00.76	10.70	00.37	53.88
	MMF	58.95	37.17	**03.87**	00.29	40.04
Sadness	ZFF	**95.13**	00.99	**03.88**	00.36	51.63
	YAGA	93.11	**00.51**	06.38	**00.27**	**72.12**
	DYPSA	90.79	02.37	06.84	00.52	53.07
	SEDREAMS	95.06	01.02	03.92	00.39	62.07
	SE-VQ	93.99	00.55	05.46	00.32	64.86
	MMF	87.81	04.06	08.13	00.56	57.16
Disgust	ZFF	**93.98**	**00.20**	05.82	00.35	51.71
	YAGA	92.77	00.50	06.72	**00.34**	**62.32**
	DYPSA	89.64	03.51	06.85	00.40	55.75
	SEDREAMS	93.72	00.43	05.85	00.36	59.05
	SE-VQ	93.45	00.46	06.10	00.43	50.32
	MMF	81.87	13.45	**04.68**	00.49	48.42
Boredom	ZFF	**96.63**	**00.59**	**02.79**	00.35	51.37
	YAGA	95.46	00.64	03.91	**00.20**	**76.37**
	DYPSA	93.28	02.45	04.27	00.35	64.68
	SEDREAMS	96.03	00.73	03.24	00.31	64.72
	SE-VQ	94.80	01.47	03.73	00.28	66.06
	MMF	88.06	06.85	05.09	00.37	64.74

Table 2. Performance of the six GCI detection methods averaged over all emotions (except neutral) of the EMO-DB database. IDR - Identification rate, MR - Miss rate, FAR - False alarm rate, IDA - Identification accuracy in ms, and IDR(± 0.25 ms) - IDR within ± 0.25 ms (%).

Method	IDR%	MR %	FAR %	IDA (ms)	IDR(± 0.25 ms)
ZFF	**92.47**	**00.47**	07.05	00.33	53.56
YAGA	91.62	00.89	07.48	**00.31**	**64.30**
DYPSA	88.62	04.59	06.79	00.36	60.00
SEDREAMS	92.22	00.71	07.07	00.36	59.28
SE-VQ	91.92	00.64	07.44	00.36	58.14
MMF	73.67	21.50	**04.82**	00.39	48.86

are presented in Table 2. By comparing the evaluation metrics computed for neutral emotion (Table 1, the first row) to the corresponding metrics in Table 2, it can be seen that the performance of GCI detection from speech in neutral was better than in speech pooled over the six other emotions. Most importantly, the high GCI detection performance in neutral was observed for each GCI detection algorithm and in all evaluation metrics except for the MMF method using the FAR metrics. In addition, it can be observed that the performance of all algorithms is approximately equal in terms of IDR in neutral with slightly lower performance in DYPSA and MMF. Among the emotions involved, the high-pitched emotions anger and joy show a large reduction (6–8%) in IDR for all GCI estimation algorithms. IDR in disgust and fear is slightly lower compared to neutral, whereas IDR in sadness and boredom is equal to that in neutral.

In analysing the metrics averaged over the six non-neutral emotions (Table 2), it can be seen that IDR of the ZFF algorithm is highest, but the scores of SEDREAMS, SE-VQ and YAGA are close to ZFF. In terms of IDR$\pm 0.25\ msec$, it can be observed that YAGA showed the highest accuracy. The DYPSA algorithm gives almost comparable accuracy (some times higher) performance, just below the accuracy of YAGA, and SEDREAMS. The accuracy of the ZFF algorithm seems to be not changing much (its accuracy around 55%) among the various emotion categories, and it is closer or better to SEDREAMS and SE-VQ in some of the emotion categories (anger, happy and fear). The identification rates of DYPSA and MMF are lower compared to other four algorithms. Between these two methods, DYPSA, however, showed better performance than MMF in all emotions.

The results indicate clearly that the performance of the GCI extraction deteriorates when comparing neutral to the other six emotions. The performance degradation is more severe in anger and joy that include high-pitched utterances compared to sadness and boredom that show lower pitch values and lower pitch variations. One of the main reasons for the degradation of the GCI detection performance is that most GCI algorithms either directly or indirectly depend on the estimation of the average pitch period and this estimation fails for high-pitched utterances that are prevalent particularly in emotions of high arousal such as anger and joy. Moreover, the use of excitation signals (such as the LP residual or the glottal source waveform) in some of the GCI detection algorithms results in degraded performance due to the smoothing of the abruptness of the excitation waveform at glottal closure in high-pitched utterances. In addition, the rapid variations in pitch in emotional speech results in vocal fold vibrations with lower suction, which also leads to reduced accuracy in detection of the glottal closure instants.

5. SUMMARY AND CONCLUSIONS

Six state-of-art GCI detection algorithms (ZFF, YAGA, DYPSA, SEDREAMS, SE-VQ and MMF) were compared for automatic detection of GCIs from emotional speech. The performance of the algorithms was assessed using known evaluation metrics on speech utterances of the EMO-DB database representing seven emotions. From the experimental results, it was observed that the GCI detection performance in neutral emotion is nearly equal for all the algorithms (except DYPSA and MMF). Results on the remaining six other emotions indicate that the GCI detection performance degrades heavily especially in anger and joy. This is due to deterioration in the estimation of the average pitch period in these emotions compared to neutral. In addition, emotions such as anger and joy show reduced abruptness in the excitation signal waveforms, computed by the GCI detection algorithms, near GCIs. This reduced abruptness increased the number of erroneously identified GCIs.

Based on the experiments conducted, it can be argued that improved GCI extraction algorithms are needed to enhance the performance of GCI detection in emotional speech. To reach this goal, the estimation of the average pitch period should be improved by, for example, adapting to the local variation of pitch in the signal [24, 34, 35]. In addition, improved GCI extraction algorithms can also be developed by combining known methods such as ZFF (as it provides good IDR) and YAGA (as it provides good IDR within $\pm 0.25\%$ ms).

6. ACKNOWLEDGEMENTS

This study was partly funded by the Academy of Finland (project no. 312490). The third author would like to thank the Indian National Science Academy (INSA) for their support.

7. REFERENCES

[1] B. Yegnanarayana and S. V. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, no. 5, pp. 651–697, 2011.

[2] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 3, pp. 972–980, May 2006.

[3] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

[4] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *J. Acoust. Soc. Am.*, vol. 120, pp. 3289–3305, 2009.

[5] C. D Alessandro and N. Sturmel, "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude," *Sadhana*, vol. 36, no. 5, pp. 601–622, 2011.

[6] 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, Jan 2001.

[7] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, 2018.

[8] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *J. Acoust. Soc. Am.*, vol. 35, pp. 344–353, 1963.

[9] D. Silva, L. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, 2009.

[10] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 20, no. 3, pp. 994–1006, 2012.

[11] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication*, vol. 55, no. 2, pp. 295–314, 2013.

[12] S. R. Kadiri, "A quantitative comparison of epoch extraction algorithms for telephone speech," in *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6500–6504.

[13] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, May 2009.

[14] M. A. Joseph, S. Guruprasad, and Y. B, "Extracting formants from short segments using group delay functions," in *Proc. INTERSPEECH*, Sep. 2006, pp. 1009–1012.

[15] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[16] R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 481–484, July 2007.

[17] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *Proc. INTERSPEECH*, August 2013, pp. 1916–1920.

[18] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. INTERSPEECH*, 2015, pp. 1324–1328.

[19] R. Sun, E. Moore, and J. Torres, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," *ICASSP*, pp. 4509–4512, 2009.

[20] H. K. Vydana, S. R. Kadiri, and A. K. Vuppala, "Vowel-based non-uniform prosody modification for emotion conversion," *Circuits, Systems, and Signal Processing*, vol. 35, no. 5, pp. 1643–1663, 2016.

[21] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio Speech and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[22] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," *Proc. INTERSPEECH*, pp. 1829–1832, 2008.

[23] T. Drugman, J. Kane, and C. Gobl, "Modeling the creaky excitation for parametric speech synthesis," *Proc. INTERSPEECH*, pp. 1424–1427, 2012.

[24] D. Govind and S. R. M. Prasanna, "Epoch extraction from emotional speech," in *Proc. SPCOM*, July 2012, pp. 1–5.

[25] 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.

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

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

[28] V. Khanagha, K. Daoudi, and H. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 1941–1950, Dec 2014.

[29] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," *IEEE Sig. Pro. Letters*, vol. 20, no. 4, pp. 387–390, 2013.

[30] S. R. Kadiri and B. Yegnanarayana, "Speech polarity detection using strength of impulse-like excitation extracted from speech epochs," in *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5610–5614.

[31] M. Brookes. Voicebox: A speech processing toolbox for MATLAB. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[32] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," in *ICASSP*, 2014, pp. 960–964.

[33] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[34] S. R. Kadiri and B. Yegnanarayana, "Analysis of singing voice for epoch extraction using zero frequency filtering method," in *Proc. Proc. ICASSP*, April 2015, pp. 4260–4264.

[35] ——, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52 – 63, 2017.