Kadiri, Sudarsana Reddy

A Quantitative Comparison of Epoch Extraction Algorithms for Telephone Speech

Published in:
44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019; Brighton; United Kingdom; 12-17 May 2019 : Proceedings

DOI:
10.1109/ICASSP.2019.8683558

Published: 01/05/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Kadiri, S. R. (2019). A Quantitative Comparison of Epoch Extraction Algorithms for Telephone Speech. In *44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019; Brighton; United Kingdom; 12-17 May 2019 : Proceedings* (pp. 6500-6504). Article 8683558 (Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; Vol. 2019-May). IEEE. https://doi.org/10.1109/ICASSP.2019.8683558

A QUANTITATIVE COMPARISON OF EPOCH EXTRACTION ALGORITHMS FOR TELEPHONE SPEECH

Sudarsana Reddy Kadiri[1,2]

[1]International Institute of Information Technology-Hyderabad, India.
[2]Aalto University, Finland.

[1]sudarsanareddy.kadiri@research.iiit.ac.in; [2]sudarsana.kadiri@aalto.fi

ABSTRACT

Telephone speech is one of the degradations involved in building speech systems in practical environments. The potential use of the speech systems depends on the speech analysis algorithms that can handle different acoustic variations and degradations often found in the human speech communication. Detection of epochs/glottal closure instants (GCIs) is typically required in such analysis stages. In this paper, the effect of telephone channel speech on the accuracy of detection of epochs using state-of-art epoch extraction methods is investigated. Epoch is the instant of significant excitation to the vocal tract system in voiced speech. Most of the existing epoch extraction algorithms are shown to perform excellently well on the speech data collected under lab environment. The efficiency of these algorithms for the analysis of telephone quality speech is quantitatively studied and the strengths and weaknesses of the methods are discussed here. The methods are evaluated on six large databases containing speech and simultaneous EGG recordings as the ground truth. The state-of-art epoch extraction algorithms considered in this study for comparison are: ZFF, YAGA, DYPSA, SEDREAMS, SE-VQ and MMF. The performance of the algorithms is evaluated in terms of both reliability and accuracy measures.

Index Terms— Speech analysis, Excitation source, Epochs, Glottal Closure Instants, Telephone channel speech.

1. INTRODUCTION

The objective of this study is to examine the robustness of the state-of-art epoch extraction algorithms in one of the degraded environment namely telephone channel speech. Since this is the mostly used practical environment in daily life for speech communication, methods of detecting epochs/glottal closure instants (GCIs) is essential for various speech applications. Speech processed through telephone channel is one of the major degradation involved in developing speech systems. In this work, the effect of telephone channel on epoch extraction methods is investigated.

Epoch is the instant of significant excitation to the vocal tract system during the production of voiced speech and it takes place around the glottal closure. Identification of epoch locations plays a crucial role in many speech processing applications such as speech modification [1], excitation source modeling [2], inverse filtering [3, 4], joint optimization in concatenative speech synthesis [5], speech pathology [6, 7], etc. Apart from above applications, the high SNR property of the GCI was used in applications like glottal activity detection [8], pitch tracking [9, 10], formant frequencies [11], analysis and detection of phonation types [12, 13] and emotions [14, 15], speaker recognition [16], speech enhancement [8], multi-speaker separation, identification of number of speakers from multi-speakers data [17] etc. Due to wider range of applications, GCI detection has received a considerable amount of research attention. From the studies in [18], it was observed that most of the epoch detection methods were shown to provide good accuracy on the speech data collected in the lab environments. Also, some attempts were made to see the effectiveness of these methods for additive noise degraded conditions [19–23]. However, there are not many attempts in GCI detection for the degraded conditions like telephone quality speech.

Due to impulse-like excitation of glottal closure, the SNR of the speech signal is high in the region around the epochs. Hence, it is possible to enhance the speech by exploiting the characteristics of speech signals in the regions around the epochs [8]. Also, it is observed that features derived around epochs provide complementary information to the existing spectral features [16]. In the human perception also, the instants of significant excitation plays an important role. It is because of the epochs in speech, human beings able to perceive and process the distant speech, even though spectral components of the signal suffer an attenuation. In the whispered voice, we may not be able to get the message from a distance of 10 feet or more due to absence of epochs [24]. Human beings are able to perceive these microlevel events without much effort in extracting the information from speech even under the degradations such as noise, reverberation and channel variations such as telephone quality speech. Development of speech systems in practical environment gained a special interest in recent years in order to enable access to voice-based services. Speech processed through telephone channel is one of the major degradation involved in building the speech systems in practical environment. In this work, the effect of telephone quality speech on epoch extraction is assessed. The quality of speech in the case of telephone speech is effected by the bandwidth of the telephone channel. Telephone channel can be approximated by a bandpass filtering in the range of 300 Hz to 3400 Hz. The characteristics of the filter varies from 0-300 Hz, 300-3400 Hz and 3400-4000 Hz.

Epoch extraction: A Review In this section, a brief discussion of the key features of the epoch extraction methods is presented. The approaches for detecting epochs can be broadly classified into three. First approach is based on processing of the excitation signal (after source-filter decomposition) for epoch detection. The second approach involves directly processing of the speech signal based on the properties of the impulse-like excitation of epoch. The methods in the third approach uses both the excitation signal and speech signal. In this, excitation signal is used to accurately locate the epochs.

The methods in the first approach rely on the excitation signal derived from the speech waveform after removing the predictable portion and this is usually carried out by performing linear predic-

tion (LP) analysis. The large error value seen in the LP residual within a pitch period is supposed to indicate the epoch location. However, identification of unambiguous epoch locations from LP residual is difficult due to random polarity of residual around epochs. To overcome this Hilbert envelope of LP residual was proposed in [25]. Some methods also uses Gabor filtered or center of gravity of Hilbert envelope of LP residual to identify epoch locations [26]. Recently in [27], an integrated LP residual (ILPR) was used as a preprocessing signal and epoch candidates were selected by detecting transients in ILPR using nonlinear temporal measure called as plosion index. Some methods uses the group delay function of LP residual to locate epochs precisely [26,28]. However, it was observed that group delay based methods gives large number of false alarms. To reduce this effect, a dynamic programming based technique is used for selecting appropriate epoch candidates [29]. Instead of LP residual, some methods uses the glottal flow waveform to locate GCIs. YAGA is one such method [30], which uses the glottal flow waveform, wavelet transform, group delay and dynamic programming. Also, recently a method was proposed which uses the glottal flow waveform signal with the time domain criteria for detecting GCIs by forward-backward algorithm in [20].

The methods in the second approach uses the properties of the impulse-like excitation of epoch present in the speech signal. Zero frequency filtering (ZFF) is one such method which exploits the nature of the impulse-like excitation by filtering the speech signal around 0 Hz [24]. The lines of maximum amplitude (LoMA) is an another method in this category which uses the time-scale representation to locate the epochs [2]. The idea in this method is that, the discontinuities in speech (such as GCIs or GOIs) are reflected as amplitude maxima at each scale of wavelet transform. An optimal LoMA is computed within a pitch period using a dynamic programming to locate GCIs. In [19], a nonlinear formalism, namely microcanonical multiscale formalism was used to highlight the impulses present in the speech signal directly.

The methods in the third category uses the speech signal to identify the possible GCI locations in a certain interval, and then discontinuities in excitation signal was used to precisely locate the GCIs. SEDREAMS is one such method [18], which uses the mean based signal for finding the possible GCI locations in an interval and then an LP residual is inspected in that to detect accurate location. A modified version of SEDREAMS was proposed in [31] for handling GCI detection from various voice qualities. This method uses dynamic programming and post-processing techniques in addition to SEDREAMS method.

The present study investigates the strengths and weaknesses of the state-of-art epoch extraction methods in processing one of the practical environment data, namely telephone quality speech. More details of the epoch extraction methods, and the epoch-based analysis of speech processing can be found in [8, 18, 32].

The organization of the paper is as follows. Section 2 briefly describes the implementation details of the state-of-art epoch extraction algorithms used for comparison in this study. Section 3 describes the speech databases used and details of the experimental protocol, which includes the ground truth and evaluation metrics. The results of the experiments are presented in Section 4 along with their strengths and weaknesses. Finally, Section 5 gives a summary of the study.

2. METHODS FOR EPOCH EXTRACTION

The following six state-of-art epoch extraction algorithms are considered in this study. They are: zero frequency filtering (ZFF), Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS), SEDREAMS-voice quality (SE-VQ), dynamic programming phase slope algorithm (DYPSA), yet another GCI algorithm (YAGA) and microcanonical multi-scale formalism (MMF) [18, 19, 24, 29–31]. A brief implementation details of each of the chosen methods are given below.

2.1. ZFF Algorithm

The epochs are detected in this method by directly exploiting impulsive property of the epoch present in the speech signal. The zero frequency filtering (ZFF) technique [24] is based on the observation that impulse-like excitation at GCIs have an effect across all frequencies. In this method, the differenced speech signal is passed through a cascade of two ideal zero-frequency resonators (ZFRs). The resulting signal grows/decays approximately as a polynomial function of time. The trend removed signal is called as zero-frequency filtered signal and the instants of negative-to-positive zero crossings (NPZCs) correspond to the epochs for positive polarity speech signal [33, 34].

2.2. SEDREAMS Algorithm

In this method, the epochs are detected using both the excitation signal and speech signal. The Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) [18] algorithm uses a mean based signal (which is obtained by calculating the mean of the sliding window whose length is 1.75 times average pitch period over the speech signal) and LP residual. In this, the first step determines the short intervals of GCIs presence using mean based signal and in the second step, the accurate GCIs are located by finding the largest discontinuity in the LP residual within those short intervals.

2.3. SE-VQ Algorithm

This method also uses both the excitation signal and speech signal in order to locate epochs. SE-VQ is a modified version of SEDREAMS algorithm, which is proposed to handle various voice qualities. Apart from the steps in SEDREAMS, the modifications involves applying a dynamic programming method (to select optimal path on GCI locations) and a finer post-processing to remove false positives, while at the same time not removing true positive GCIs.

2.4. DYPSA Algorithm

This method only uses the excitation signal to locate the epochs. The dynamic programming phase slope algorithm (DYPSA) [29] uses the LP residual of the speech signal. In this method, firstly zero crossings of the phase slope function calculated on the LP residual was used to obtain appropriate GCI candidates. And then, a phase slope projection technique was used to recover candidates for which the phase slope function does not include zero crossings. In order to identify true epochs by reducing the effect of false candidates, a dynamic programming was used by minimizing various cost functions. The cost function consists of five elements, they are: 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.

2.5. YAGA Algorithm

This method (yet another GCI algorithm (YAGA) [30]) uses the excitation signal (glottal flow waveform) to locate the epochs. The method combines various approaches used in other GCI detection methods, such as: wavelet analysis, group delay function and M-best dynamic programming. In order to highlight the discontinuities in the glottal flow waveform, the multi-scale product of the stationary wavelet transform is used and then the discontinuities are detected using negative going zero crossings of the group delay function. The falsely detected candidate GCIs are then removed using a similar M-best dynamic programming approach as is used in DYPSA algorithm [29]. YAGA also uses similar cost elements of DYPSA, with modification of the inter-pulse similarity cost and an additional element for distinguishing GCIs and GOIs.

2.6. MMF Algorithm

This method directly exploits the impulse-like discontinuity present in the speech signal using a nonlinear formalism. The method is based on the approach of microcanonical multi-scale formalism (MMF) [19]. It relies on the precise estimation of multi-scale parameter called as singularity exponent at each sampling instant in the time domain. The singularity exponent parameter quantifies the degree of signal singularity at each sample from a multi-scale point of view. The subset of samples with lowest singularity exponent values points towards the GCIs.

In all these methods, the main challenge lies in the detection of one GCI for one pitch period or glottal cycle, which lies closer to the actual GCIs.

3. EXPERIMENTAL PROTOCOL

The robustness of epoch extraction algorithms are examined for bandwidth degradation as in telephone quality speech. The performance of all the six algorithms on simulated telephone quality speech (as in [35, 36]) is validated, as there are no databases available which consists of actual telephone channel speech with simultaneous EGG recordings. The telephone channel is simulated as a bandpass filter with passband of 300-3400 Hz using an infinite impulse response filter implementation available in [37] and its frequency response is shown in Fig. 1. The clean speech data is filtered through the simulated telephone channel to obtain telephone quality speech.

3.1. Speech Material and Ground Truth

The six state-of-art methods are evaluated on six large databases containing speech and simultaneous EGG recordings as the ground truth. Among these, first three databases are from CMU ARCTIC database [38]. These databases were collected for the purpose of developing speech synthesis systems. Each of these three databases consists of around 1132 phonetically balanced English sentences, spoken by a single speaker, they are: BDL (US male), JMK (US male) and SLT (US female). In the fourth database, a set of nonsense words containing all phone-phone transitions in English were recorded and is referred as RAB database (spoken by the UK male speaker). The fifth database is the KED TIMIT database spoken by a US male speaker. All these five databases are available on the Festvox webpage [38, 39]. The sixth database is the APLAWD database [30], which contains ten repetitions of five phonetically balanced English sentences spoken by 5 male and 5 female speakers

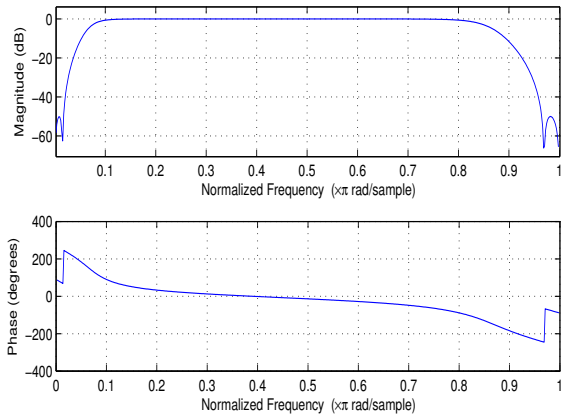

Fig. 1. Frequency response of the simulated telephone channel (for sampling frequency of 8 kHz).

Table 1. *Summary of the databases used for validation.*

Database	Speaker(s)	Duration (approx.)
BDL	1 male	54 min.
JMK	1 male	55 min.
SLT	1 female	54 min.
KED	1 male	20 min.
RAB	1 male	29 min.
APLAWD	5-male, 5-female	20 min.
Total	9-male, 6-female	232 min.

and is available in [40]. In all these databases, the EGG and speech signals sampled at 8 kHz are considered for evaluation. Reference epoch locations were extracted by finding the negative peaks in the dEGG signal. The EGG and speech signals were aligned to compensate the larynx-to-microphone delay. The epoch extraction methods are validated only on the voiced segments as epochs are meaningful only for voiced segments. The total data collectively consists of 9 male and 6 female speakers. The description about the databases is summarized in Table 1.

3.2. Evaluation Measures

The measures defined in [29] are used to evaluate the performance of the epoch extraction algorithms. The first three measures: identification rate (IDR), miss rate (MR) and false alarm rate (FAR) are called as reliability measures, and the remaining two: identification accuracy (IDA) and identification rate within ± 0.25 ms (IDR within ± 0.25 ms) are called as accuracy measures.

4. RESULTS AND DISCUSSION

To examine the effect of telephone quality speech on epoch extraction algorithms, the evaluation metrics were calculated for each database in both male and female speech separately. The obtained results are presented in Table 1 (for each database in both genders) and Table 2 (averaged across all databases). From Table 1, it can be seen that the performance of the epoch extraction algorithms is drastically lower for all the databases (which has larger FAR) in terms of both IDR and IDR within $\pm 0.25ms$. It can also be observed from Tables 1 and 2 that, the DYPSA and MMF algorithms are consistently giving higher performance for almost all

Table 2. *Performance comparison of the six methods of GCI detection for the six databases. IDR - Identification rate, MR - Miss rate, FAR - False alarm rate, IDA - Identification accuracy (σ) in ms, IDR within ± 0.25 ms (%).*

Database	Method	IDR%	MR %	FAR %	IDA (ms)	IDR(± 0.25 ms)
BDL(male)	ZFF	50.73	**0.12**	49.15	0.70	19.98
	YAGA	49.78	0.97	49.25	1.11	22.42
	DYPSA	91.59	2.13	06.28	0.48	**82.71**
	SEDREAMS	72.49	**0.12**	27.40	**0.47**	58.55
	SE-VQ	87.43	2.77	09.80	0.86	16.75
	MMF	**94.04**	3.33	**02.63**	0.49	71.02
JMK(male)	ZFF	94.20	1.55	04.25	1.06	26.72
	YAGA	63.72	2.14	34.14	1.25	09.87
	DYPSA	88.72	1.97	09.31	0.70	64.25
	SEDREAMS	**98.60**	**0.33**	**01.06**	**0.60**	**65.08**
	SE-VQ	87.60	4.11	08.30	1.20	04.30
	MMF	94.17	2.65	03.18	0.68	53.22
SLT(female)	ZFF	**98.91**	**0.05**	**01.04**	0.35	**65.82**
	YAGA	71.01	8.06	20.93	1.10	21.21
	DYPSA	93.43	2.89	03.69	0.33	64.19
	SEDREAMS	95.44	2.77	01.79	**0.28**	63.57
	SE-VQ	84.34	5.98	09.68	0.87	09.90
	MMF	91.87	7.09	**01.04**	0.37	61.35
KED(male)	ZFF	21.79	**0.07**	78.14	0.51	21.89
	YAGA	55.66	0.91	43.42	1.07	33.44
	DYPSA	96.03	1.14	02.83	**0.32**	**87.01**
	SEDREAMS	35.85	1.53	62.62	0.97	38.59
	SE-VQ	69.51	0.71	29.78	0.96	06.06
	MMF	**96.65**	1.53	**01.82**	0.33	79.90
RAB(male)	ZFF	53.98	0.22	45.81	1.37	26.64
	YAGA	18.18	2.70	79.12	1.21	13.62
	DYPSA	**76.74**	1.53	**21.73**	0.50	74.51
	SEDREAMS	61.64	**0.20**	38.16	**0.48**	69.24
	SE-VQ	65.39	3.47	31.14	1.03	08.14
	MMF	73.61	1.75	24.64	0.70	**75.33**
APLAWD(male)	ZFF	72.51	**0.03**	27.46	0.72	22.99
	YAGA	57.95	0.66	41.39	1.09	18.30
	DYPSA	89.52	1.39	09.08	**0.48**	**78.83**
	SEDREAMS	89.60	0.06	10.34	**0.48**	64.86
	SE-VQ	84.33	1.95	13.72	0.98	16.60
	MMF	**91.49**	2.72	**5.78**	0.54	66.07
APLAWD(female)	ZFF	77.45	**0.05**	22.54	**0.31**	64.89
	YAGA	66.39	1.01	32.60	0.97	31.85
	DYPSA	77.82	1.41	20.77	0.38	**79.82**
	SEDREAMS	75.90	0.06	24.04	0.37	57.99
	SE-VQ	72.42	1.12	26.46	1.08	05.36
	MMF	**80.77**	7.21	**12.03**	0.61	58.86

the databases compared to other four methods. In both of these algorithms, DYPSA is better than MMF in most of the cases. Also, it was observed that in comparison with various algorithms, MMF offers the good reliability and DYPSA offers the good accuracy, with consistency.

The performance of the ZFF method in terms of IDR and IDR within ± 0.25 ms is very low for the male speakers data (except for JMK), whereas for the female speakers data, the performance is relatively high. The performance of the YAGA method is low in terms of IDR and IDR within ± 0.25 ms for all the databases including male and female speakers. The SEDREAMS method gives good IDR and IDR within ± 0.25 ms for some databases, such as JMK, SLT and APLAWD (especially male speaker). However, the performance of the method is poorer for other databases. It is interesting to note that, even though SEVQ method is a modified version of SEDREAMS, its performance is significantly lower especially in terms of IDR within ± 0.25 ms compared to SEDREAMS. Whereas IDR is improved in some databases such as BDL, KED and RAB (slightly) compared to SEDREAMS. The IDR of SEVQ algorithm is reduced in some cases such as JMK, SLT and APLAWD. It can be seen that, there is a consistency in the performance of the DYPSA and MMF methods in terms of IDR and IDR within ± 0.25 ms (except for the RAB speaker, which has lower IDR) for both the male and female speakers data. It is interesting to note that, the performance of these two methods (DYPSA and MMF) is relatively lower in clean speech data compared to other four methods (see [18,19]). But this is not case for telephone quality speech. Infact, these two methods are performing better than all the other four methods.

From the results in Tables 1 and 2, it is to be noted that, the methods that uses the smoothed signal (such as zero frequency filtered signal in ZFF method, mean based signal in SEDREAMS and SEVQ) for detecting GCIs gives lower performance in some databases and its effect is varying for databases. This is mainly due to the spurious zero crossings of the smoothed signal which leads to large number of false alarms. It is interesting that, the SEVQ method performance is lower compared to SEDREAMS method. The decremental performance in this method may be due to the cost functions/thresholds involved in the method and also due to post-processing technique. It can be seen that, YAGA method has larger number of false alarms, which leads lowered performance. One reason for this might be due to the inaccurate estimates of glottal source waveform. The other reason might be the effect of the cost functions/thresholds involved in dynamic programming which are optimized for clean speech data. The performance in the case of DYPSA and MMF methods consistently higher in most of the cases. The reason for the good performance of DYPSA method might be due to its depends on the LP residual signal to locate the epochs and the higher performance of MMF method may be due to its exploitation of impulse-like discontinuity directly from the time domain signal.

Table 3. *Performance comparison averaged over all databases for the six algorithms of epoch extraction.*

Method	IDR%	MR %	FAR %	IDA (ms)	IDR(± 0.25 ms)
ZFF	67.08	**0.30**	32.63	0.72	35.56
YAGA	54.67	2.35	42.98	1.11	21.53
DYPSA	87.69	1.78	10.53	**0.46**	**75.90**
SEDREAMS	75.65	0.72	23.63	0.52	59.70
SE-VQ	78.72	2.87	18.41	0.99	9.59
MMF	**88.94**	3.75	**7.30**	0.53	66.53

5. SUMMARY AND CONCLUSION

A quantitative comparison of six state-of-art epoch extraction algorithms for automatic detection of epochs from the telephone quality speech was investigated. The algorithms considered in this study are: ZFF, YAGA, DYPSA, SEDREAMS, SE-VQ and MMF. The performance of these algorithms was assessed on simulated telephone quality speech of six large databases which contains 9 male and 6 female speakers data. From the experimental results, it was observed that the performance for telephone quality speech is degraded heavily for all the methods. It was also observed that, ZFF algorithm seems to provide good performance for female speech, but the method performance is reduced drastically in male speech especially in IDR within ± 0.25 ms even though for some speakers IDR is high (such as JMK). Similar the case in SEDREAMS method also. YAGA method performance is poorer in both male and female speech. Among the six methods, it appears that DYPSA and MMF methods seems to work well in most of the cases even though the performance of the methods are lower compared to clean speech (can be seen from results in [18, 19]). In all these methods, the effect is mainly due to the bandwidth of the telephone channel involved, which makes the impulse-like discontinuity as less evident. The experimental results clearly display that there is a clear lack and need for more reliable and accurate algorithms of epoch extraction for practical degraded data like telephone quality speech.

6. ACKNOWLEDGEMENTS

This study was partly funded by the Academy of Finland (project 312490).

7. REFERENCES

[1] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Signal Process. Letters*, vol. 14, no. 3, pp. 972–980, May 2006.

[2] 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.

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

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

[7] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 9:1–9:9, Jan. 2009.

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

[9] 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.

[10] S. R. Kadiri and B. Yegnanarayana, "Estimation of fundamental frequency from singing voice using harmonics of impulse-like excitation source," in *INTERSPEECH*, 2018, pp. 2319–2323.

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

[12] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs)," in *INTERSPEECH*, 2018, pp. 232–236.

[13] ——, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," in *INTERSPEECH*, 2018, pp. 441–445.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] A. Koutrouvelis, G. Kafentzis, N. Gaubitch, and R. Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Transactions on Audio, Speech, and Lang. Process.*, vol. 24, no. 2, pp. 316–328, Feb. 2016.

[21] G. Seshadri and B. Yegnanarayana, "Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 1853–1864, 2011.

[22] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52 – 63, 2017.

[23] G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Detection of glottal closure instants in degraded speech using single frequency filtering analysis," in *Interspeech*, 2018, pp. 2300–2304.

[24] 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.

[25] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Speech Audio Processing*, vol. 27, pp. 309–319, 1979.

[26] 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. Letters*, vol. 14, no. 10, pp. 762–765, 2007.

[27] A. Prathosh, T. Ananthapadmanabha, and A. 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.

[28] 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.

[29] 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.

[30] 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 & Lang. Process.*, vol. 20, no. 1, pp. 82–91, 2012.

[31] 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.

[32] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

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

[34] ——, "Speech polarity detection using strength of impulse-like excitation extracted from speech epochs," in *ICASSP*, 2017, pp. 5610–5614.

[35] C. Vikram and S. Mahadeva Prasanna, "Epoch extraction from telephone quality speech using single pole filter," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 3, pp. 624–636, 2017.

[36] K. Vijayan and K. S. R. Murty, "Epoch extraction by phase modelling of speech signals," *Circuits, Systems, and Signal Processing*, vol. 35, no. 7, pp. 2584–2609, 2016.

[37] M. Brookes. Voicebox: A speech processing toolbox for MATLAB.2006. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[38] J. Kominek and A. Black, "The CMU Arctic speech databases," in *5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[39] T. F. Website, Source: http://festvox.org/cmu_arctic/index.html.

[40] Source: http://www.commsp.ee.ic.ac.uk/~sap/resources/aplawdw/.