Excitation Epoch and Voicing Detection Using Hilbert Envelope with Single-Pass Processing

Hirak Dasgupta and Prem C. Pandey
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India
<hirakdgpt@ee.iitb.ac.in>, <pcpandey@ee.iitb.ac.in>

Abstract— A technique is presented for excitation epoch and voicing detection in speech signal using its Hilbert envelope and employing single-pass processing. The excitation epoch detection comprises dynamic range compression for reducing amplitude variability, Hilbert envelope calculation and dynamic peak detection for excitation saliency enhancement, and epoch marking by locating the maximum-sum subarray peaks. The voicing detection is based on thresholding the inter-epoch similarity calculated as the normalized covariance of the adjacent inter-epoch intervals of the Hilbert envelope. The total algorithmic delay is less than 60 ms. The epoch detection and the voicing detection for clean and telephone-quality speech showed a good match with those obtained from the EGG, and the detection performances compared favorably with the earlier techniques.

Keywords— excitation epochs; Hilbert envelope; voicing detection.

I. INTRODUCTION

Speech signal can be segmented into voiced and unvoiced segments depending on the mode of excitation of the vocal tract [1]–[4]. The excitation is characterized by significant glottal vibration during the voiced segments and its absence during the unvoiced segments. The voiced speech signal can be modelled as the output of a time-varying vocal tract filter excited by pulsatile airflow due to significant quasiperiodic vibration of the glottal folds [2]. The pulsatile excitation occurs around the instants of glottal closure, known as the excitation epochs (also as glottal closure instants and pitch marks) [1]. The duration between two successive excitation epochs in a voiced segment is termed the pitch period. The voicing and pitch detection is needed in most speech processing applications, namely speech coding [5]–[7], voice conversion [8], [9], speech synthesis [10], [11], voice disorder classification [12]–[14], speech recognition [15], [16], etc.

Excitation epoch detection involves epoch saliency enhancement and epoch marking [17]–[25]. The epoch saliency enhancement may be carried out by calculating the LP residual, decomposing the signal into multiple bands, phase spectrum analysis, etc. In [17], the speech signal is passed through two cascaded marginally-stable zero-frequency resonators (ZFRs) to reduce the vocal tract filter related responses. The positive zero-crossings of the near-sinusoidal signal obtained by repeated mean-trend subtraction are marked as the epochs. In [18], the epochs are initially marked between the local-minimum and positive zero-crossing on the speech signal after a moving-average filter with the Blackman window of length 1.75 times the average pitch period. These locations are refined by detecting the neighboring LP-residual peaks. Teager energy of the lowpass filtered speech signal is used for epoch detection in [19]. The technique in [20] uses integrated LP residual of the Hamming-windowed and pre-emphasized signal and the modified short-time crest factor, termed as dynamic plosion index.

The main difficulties in voicing detection are due to non-stationarity and quasiperiodicity of the voiced speech signal and interaction between the vocal tract and the glottal source [3], [31]. Several voicing detection techniques [7], [11], [15], [26]–[35] have been reported for use with pitch or epoch detection. They generally employ the processing steps of signal parameter calculation, voicing decisions, and post-processing to suppress errors. The parameters used include energy, zero-crossing rate, autocorrelation peaks, cepstral peaks, excitation strength, spectral entropy, etc. The voicing decisions can be based on thresholding or statistical approaches. In the thresholding-based decisions, the parameters are individually compared with empirically set thresholds and the comparison outputs are combined hierarchically. Errors may occur due to a significant overlap in the parameters of the voiced and unvoiced segments. In the statistical approaches, the distribution parameters of the features are estimated from the training data and subsequently used for voicing decisions.

The pitch and voicing detection technique by Hermes [26] is based on periodicity in the spectrum of voiced speech and its absence in unvoiced speech. The analysis uses 40-ms frames and 10-ms frameshift. The magnitude spectrum up to 1250 Hz is processed to enhance the peaks, the frequency axis is logarithmically compressed, the subharmonic-sum spectrum is obtained by log-step shifts, and the peak location is used to estimate the pitch. For each frame, the correlation coefficient is calculated between the signal segments of duration equal to the pitch period and selected symmetrically on either side of the frame's midpoint. The correlation coefficients are smoothed and thresholded for voicing decisions. The technique by Drugman and Alwan [27] uses summation of the LP residual harmonics for framewise pitch estimation and voicing detection. The summation residual harmonic (SRH) spectrum is calculated as $\mathrm{SRH}(k)$ from the LP residual spectrum $E(k)$ by adding $E(lk)$ and subtracting $E((l-1/2)k)$ to suppress peaks at even harmonics. Its peak location is used for fundamental frequency estimation, and the peak is thresholded for voicing detection. Optimal results were obtained for 100-ms frames, 10-ms frameshift, and summation over five harmonics. Processing with a second pass was used to suppress the detection errors.

In the technique by Gonzalez et al. [29], framewise fundamental frequency and voicing probability are estimated by applying a harmonic summing filter on the normalized power spectrum with 90-ms frames and 10-ms frameshift. The harmonic summing filter is realized as a convolution of the log-frequency power spectrum with an impulse response chosen to sum the harmonics and attenuate smoothly varying noise components. The frequency location of the highest peak in the harmonic-sum filtered spectrum is obtained, and its track is smoothed by applying a temporal continuity constraint to output the pitch track. Two features are calculated: log-mean power of the normalized spectrum and ratio of the sum

2022 National Conference on Communications (NCC) | 978-1-6654-5136-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/NCC55593.2022.9806818

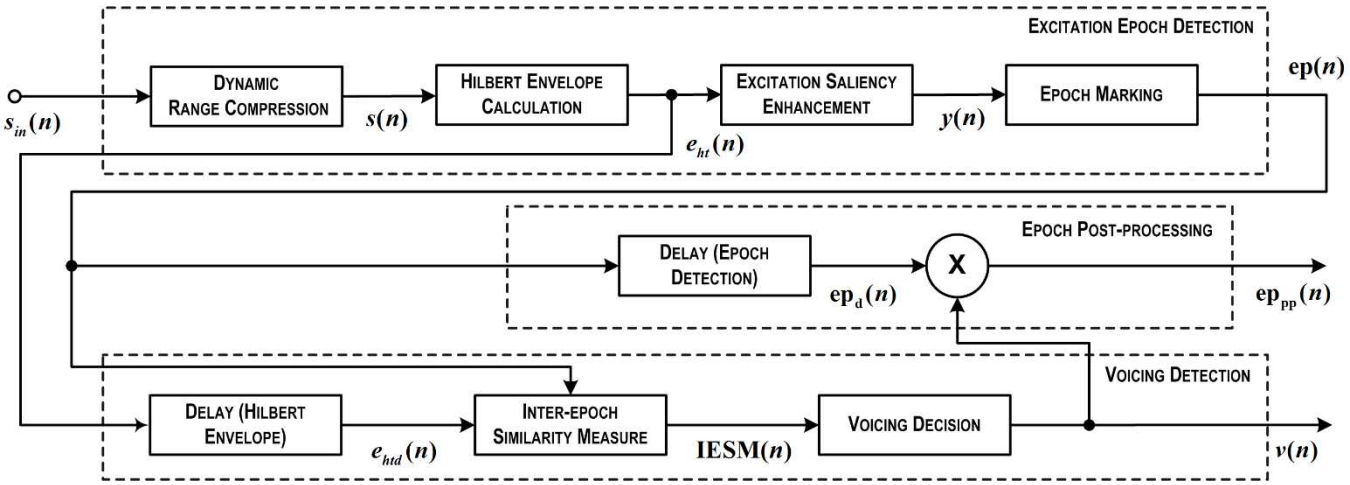

Figure 1: Block diagram of the proposed voicing and epoch detection technique.

of the highest three peaks in the harmonic-sum filtered spectrum to the mean power of the normalized spectrum. These features are applied as inputs to two GMM-based classifiers, one for voiced frames and the other for unvoiced frames. The voicing probability is calculated from the likelihood ratio of the two classifier outputs.

Qi and Hunt [30] used a multi-layer perceptron with cepstral coefficients, energy, and zero-crossing rate as the features to classify the signal frames as voiced, unvoiced, and silence. The voicing detection technique by Ahmadi and Spanias [31] thresholds the maximum cepstral peak within the quefrency range corresponding to the pitch range, energy, and zero-crossing rate, using the parameter medians over the speech signal as the thresholds. A frame is unvoiced if the cepstral peak is lower than the cepstral threshold, the energy is lower than the energy threshold, and the zero-crossing rate is higher than the zero-crossing rate threshold. The decision errors during weak voiced frames are suppressed using a time-varying cepstral threshold based on voicing decisions over three consecutive frames, and isolated errors are suppressed by a 5-point median filter.

The technique in [33] is based on robustness of the voiced epochs and the excitation strength. The epochs are detected after ZFR-filtering the signal, and the excitation strengths are calculated from the mean-subtracted output as the slope at the excitation epochs. The epochs are also detected from two noisy speech signals with each signal obtained by adding a white noise at 10-dB SNR. An epoch is declared voiced if the inter-epoch interval is less than 15 ms, mismatch of the epochs detected from the two noisy signals is less than 1 ms, and the excitation strength is above 1% of the peak excitation strength. The technique in [11] uses the excitation strength of the epochs detected from the ZFR-filtered signal with 25-ms frames and 10-ms frameshift. The excitation strength calculated as the slope of the mean-subtracted ZFR-filtered signal at the epoch location varies with the window size for the mean subtraction, and this variation may cause voicing detection errors in weak segments. Therefore, epoch detection and excitation strength calculation are carried out by varying the window over 2–15 ms in 1-ms steps. The maximum-sum of the excitation strengths at the epochs detected in the frame for these windows is used for voicing detection.

The technique by Ananthapadmanabha et al. [34] detects the excitation epochs by processing the LP residual and uses signal similarity in two successive inter-epoch intervals for voicing detection. The maximum of the normalized cross-correlation function for two successive inter-epoch intervals is calculated as the similarity measure. An inter-epoch interval is declared voiced if the smoothed measure is above 0.6.

The voicing detection techniques based on periodicity in the voiced speech spectrum [26], [27], [29] require a relatively long frame length. Hence, the detections may be erroneous in segments with pitch changes, vocal tract transitions, and nonmodal voicing characterized by amplitude and frequency perturbed glottal vibrations. Similar errors may occur due to the long window length used in the techniques based on cepstral peaks, energy, ZCR, etc. [30], [31]. The voicing detectors based on epoch detection [11], [33], [34] do not employ a long window length. However, detection errors may occur due to erroneous epoch detection, particularly in highpass-filtered speech. Some of the techniques are not suited for single-pass processing [11], [27], [31], [33]. For addressing these problems, an excitation epoch and voicing detection technique employing single-pass processing and based on the Hilbert envelope of the speech signal and an inter-epoch similarity measure is presented and evaluated. The proposed technique is described in the second section. The evaluation method and results are presented in the third and fourth sections, followed by the conclusion in the last section.

II. Proposed Technique

The proposed technique comprises excitation epoch detection, voicing detection, and epoch post-processing, as shown in Figure 1. It employs single-pass processing. The excitation epoch detection is a variant of the Hilbert envelope-based epoch detection in [36], modified for single-pass processing with a low algorithmic delay. It detects the excitation epochs during voiced segments and may detect spurious epochs during unvoiced and silence segments. The voicing detection is based on the observation that the signals in the successive inter-epoch intervals have high similarity during the voiced segments and low similarity during the unvoiced segments, and this difference is enhanced in the squared Hilbert envelope. A framewise inter-epoch similarity measure is calculated from the squared Hilbert envelope, smoothed, and thresholded to obtain the voicing decisions, and these decisions are used in the epoch post-processing to suppress the spurious epochs during the unvoiced and silence segments.

The excitation epoch detection in [36] has been reported to be robust against highpass filtering, and its performance for

speech signals with voice disorders is shown to be better than or similar to state-of-the-art techniques. Therefore, the proposed technique may suit applications involving highpass-filtered speech and voice disorders. The processing blocks are described in the following subsections.

A. Excitation Epoch Detection

The excitation epoch detection is shown as the upper block in Figure 1. It comprises dynamic range compression, Hilbert envelope calculation, excitation saliency enhancement, and epoch marking.

The dynamic range compression employs feed-forward compression based on A-law [37] applied on the input speech signal's magnitude envelope, to reduce the misdetection possibility due to the signal amplitude variability. The squared Hilbert envelope $e_{ht}(n)$ is calculated on the dynamic-range compressed output. It enhances the instants of significant excitation, even in the fundamental's absence. The squared Hilbert envelope $e_{ht}(n)$ is calculated using a Hilbert transformer realized as a causal linear-phase FIR filter with its impulse response obtained from the ideal noncausal impulse response by applying a 15-ms Hamming window. The excitation saliency enhancement uses a dynamic peak detection and a two-step median-mean smoothing to further reduce the residual ripples in $e_{ht}(n)$ without reducing the epoch saliency. A 5-point differentiation is applied on the smoothed peak envelope to emphasize the instants with a high rate of change to obtain the saliency-enhanced peak envelope $y(n)$.

The epoch marking comprises segmenting $y(n)$ in overlapping frames with the frame length equal to the largest pitch period, multiplying the frames with an exponentially decaying window, obtaining contiguous maximum-sum subarrays, and locating the subarray peaks. It is a modification of the epoch marking in [36] to reduce the algorithmic delay.

The N_w-sample frame $z_n(m)$ for the sample index n is obtained from the saliency enhanced peak envelope $y(n)$ as

$$z_n(m) = y(n - N_w + 1 + m), \qquad 0 \le m < N_w - 1 \tag{1}$$

The frame $z_n(m)$ is multiplied with an exponentially decaying window to get

$$z_{nw}(m) = z_n(m)\exp(-m/N_M), \quad 0 \le m < N_w - 1 \tag{2}$$

where N_M is empirically selected as two-thirds of the mean of the previous five pitch periods for 95% decay in two pitch periods. The average pitch period is initially set as 10 ms. The N_M value of less than half or more than three-fourths of the mean pitch period deteriorates the epoch detection. The contiguous maximum-sum subarray of $z_{nw}(m)$ is obtained using Kadane's one-pass optimal search algorithm (described in [38], [39]). With i_s and i_e as the subarray boundary indices, the subarray peak location is calculated as

$$m_e = \underset{m}{\operatorname{argmax}}\, z_n(m), \qquad i_s \le m \le i_e \tag{3}$$

and used as the candidate epoch location. The sample index n is marked as an epoch if the number of samples after the last epoch location is more than a refractory period N_r and the subarray peak location m_e is less than $N_r/2$. The refractory period N_r is the smallest pitch period, and the frame length N_w is the largest pitch period.

B. Voicing Detection and Epoch Post-processing

The voicing detection, lower block in Figure 1, comprises frame segmentation of the delayed squared Hilbert envelope, inter-epoch similarity measure calculation, and voicing decision. The squared Hilbert envelope $e_{ht}(n)$ is delayed by a delay equal to that introduced in the excitation saliency enhancement and epoch marking. The delayed squared Hilbert envelope $e_{htd}(n)$ is segmented into overlapping frames, and the normalized covariance of the signal segments in the first two inter-epoch intervals is calculated as the frame's inter-epoch similarity measure. The measure is set as zero if there are less than three epochs. This calculation requires at least three excitation epochs in the frame, and hence the frame has to be $3N_w$ samples or longer.

For the delayed squared Hilbert envelope frame at the sample index *n,* first three epoch locations within the frame are obtained as n_{e1}, n_{e2}, and n_{e3}. The shorter of the first two inter-epoch intervals is obtained as

$$N_s = \min(n_{e3} - n_{e2},\, n_{e2} - n_{e1}) \tag{4}$$

This interval is used to mark two segments $e_{hn1}(m)$ and $e_{hn2}(m)$ from the frame $e_{htd}(n)$ as the following:

$$e_{hn1}(m) = e_{htd}(n_{e1} + m), \qquad 0 \le m \le N_s - 1 \tag{5}$$

$$e_{hn2}(m) = e_{htd}(n_{e2} + m), \qquad 0 \le m \le N_s - 1 \tag{6}$$

These two segments are used to calculate the mean-subtracted segments $\hat{e}_{hn1}(m)$ and $\hat{e}_{hn2}(m)$ as the following:

$$\hat{e}_{hn1}(m) = e_{hn1}(m) - (1/N_s)\sum_{i=0}^{N_s-1} e_{hn1}(i) \tag{7}$$

$$\hat{e}_{hn2}(m) = e_{hn2}(m) - (1/N_s)\sum_{i=0}^{N_s-1} e_{hn2}(i) \tag{8}$$

From these mean-subtracted segments, the normalized covariance is calculated as

$$\rho(n) = \frac{\sum_{m=0}^{N_s-1}[\hat{e}_{hn1}(m)\,\hat{e}_{hn2}(m)]}{\max\{\sqrt{\sum_{m=0}^{N_s-1}[\hat{e}_{hn1}^2(m)]\;\sum_{m=0}^{N_s-1}[\hat{e}_{hn2}^2(m)]},\;\varepsilon\}} \tag{9}$$

It is used as the inter-epoch similarity measure for the frame ending at the sample index *n.* Here ϵ is a floor parameter to avoid numerical instability during low-level segments without significantly affecting the ρ value during the voiced segments. This floor parameter is empirically selected as 10^{-4} for an input signal range [−1, 1]. The measure $\rho(n)$ is smoothed using 10-ms median and 10-ms moving average filters to get smoothed inter-epoch similarity measure IESM, and it is thresholded for voicing decision. The resulting voicing decisions have some isolated errors, and a 30-ms median filter is used to suppress them and provide the voicing detector output $v(n)$.

The epoch post-processing is shown between the excitation epoch detection and voicing detection blocks in Figure 1. The excitation epoch detection output $\mathrm{ep}(n)$ is delayed by the delay introduced in the voicing detection. The delayed epoch detection output $\mathrm{ep_d}(n)$ is multiplied by the voicing detection output $v(n)$ to suppress the spurious excitation epochs during unvoiced and silence segments and to output the processed excitation epoch detection $\mathrm{ep_{pp}}(n)$.

C. Technique Implementation

The proposed technique was implemented using MATLAB (MathWorks, Inc., Natick, MA, USA). The

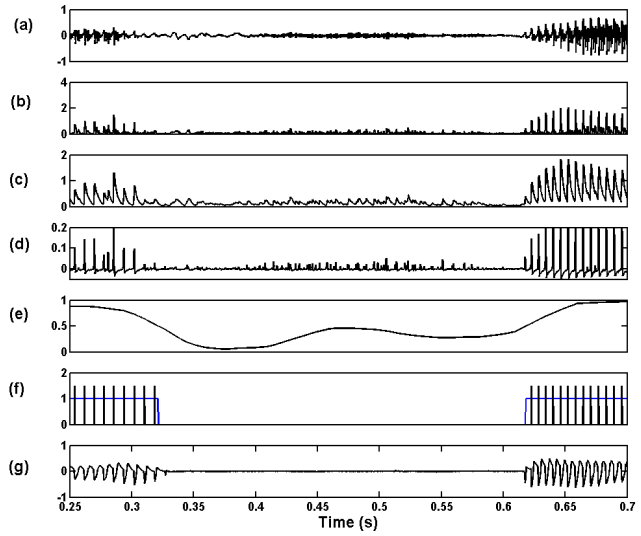

Figure 2. A processing example (/asa/ from a male speaker): (a) speech signal, (b) Hilbert envelope, (c) nonlinear smoother output, (d) differentiator output, (e) IESM, (f) detected excitation epochs (black) and voicing (blue), and (g) EGG.

processing parameters were selected for the pitch range of 100−500 Hz, with the refractory period N_r for the epoch detection as 2 ms (smallest pitch period) and the frame length N_w as 10 ms (largest pitch period). The processing was carried out with 30-ms frames and one-sample frameshift. It has an algorithmic delay of 54.8 ms (dynamic range compression: 12.5 ms, Hilbert envelope calculation: 7.5 ms, saliency enhancing: 0.8 ms, epoch marking: 9 ms, voicing detection: 25 ms with 10 ms for smoothing the inter-epoch similarity measure and 15 ms due to the 30-ms median filter).

An example of the processing with a sampling frequency of 16 kHz is shown in Figure 2. It shows the detected excitation epochs and the voicing detector output in good agreement with the glottal excitation instants and voicing in the EGG signal, respectively.

III. Evaluation Method

The technique was evaluated using the excitation epochs and voicing detected from the EGG as the reference for signals with a sampling frequency of 16 kHz. The performance measures for excitation epoch detection are identification rate (IDR), miss rate (MR), and false-alarm rate (FR) as reliability measures; identification bias (IDB) and identification accuracy (IDA) as accuracy measures; and accuracy-weighted identification rate (AWIDR) as a combined reliability-precision measure as in [24] and [36]. The performance measures for voicing detection are unvoiced-to-voiced error (UV-V), voiced-to-unvoiced error (V-UV), and the voicing decision error (VDE = UV-V + V-UV) as in [27] and [31].

Some studies have used manually marked epochs and voicing tracks as the reference for evaluation, and some have used automated epoch detection and voicing decisions from the EGG signal with manual corrections to avoid manual marking and the associated variability. The databases with parallel speech and EGG signals in these evaluations include APLAWD [40], CSTR [41], Keele [42], and CMU-ARCTIC [43]. The number of speakers and the speech material in these databases are summarized in Table 1. CSTR and Keele have reference pitch tracks with pitch breaks during the unvoiced segments. The CSTR track was obtained by detecting the excitation epochs from the EGG signal by amplitude thresholding and subsequently using the pitch periods for voicing decisions [41]. The Keele track was obtained from the EGG's autocorrelation and manual correction by observing the speech and EGG signals [42]. Such tracks are not available in APLAWD and CMU-ARCTIC. In the evaluation using CMU-ARCTIC in [33], the excitation strength was calculated from the mean-subtracted ZFR-filtered EGG as the slope of the sinusoid-like signal at the excitation epochs, and it was thresholded to detect the voiced segments. The evaluations in [11] and [27] used CSTR and Keele with reference pitch tracks. In [27], APLAWD was used with the reference voicing track obtained from the autocorrelation of the EGG signal.

Table 1: Databases with speech and EGG signals.

Database	Speakers and speech material
APLAWD [40]	5 males, 5 females. Ten repetitions of five sentences by each speaker.
CSTR [41]	1 male, 1 female. 50 sentences by each speaker
Keele [42]	5 males, 5 females. A paragraph by each speaker
CMU-ARCTIC [43]	4 males, 1 female. 1132 sentences by two male speakers and one female speaker, a set of nonsense words by one male speaker, 452 TIMIT sentences by one male speaker.

The APLAWD database has ten speakers (5 male, 5 female) but only five sentences and a significant noise in some EGG signals. Keele has ten speakers (5 male, 5 female) but only a small paragraph with four sentences and significant artifacts in the EGG signals due to electrode movements. CSTR has 50 sentences and disturbance-free EGG, but only two speakers (1 male, 1 female). CMU-ARCTIC has speech material with significant diversity and disturbance-free EGG, but only five speakers with only one female speaker. Considering the EGG quality, speech material diversity, and the number of speakers with gender balance, the evaluation was carried out using the CSTR and CMU-ARCTIC databases. The sampling frequency is 20 kHz in CSTR. It is 32 kHz in CMU-ARCTIC for three speakers and 16 kHz for the others. The database signals with 20 and 32 kHz sampling frequencies were resampled to have all signals with a sampling frequency of 16 kHz for use in the evaluation. The silences were deleted to avoid affecting the performance measures, using a voice activity detector reported in [44] with the implementation in the VOICEBOX toolbox [45].

The pitch track in CSTR did not show any significant anomaly with respect to the speech and EGG waveforms. It was used to obtain the reference voicing track. In CMU-ARCTIC, the reference epochs were marked as the negative peaks of the differentiated EGG (DEGG) by amplitude-duration thresholding as described in [36]. The voicing decisions were obtained by thresholding the EGG amplitude envelope and autocorrelation peak. In this method, the EGG is bandpass filtered using a 20-ms FIR filter to suppress the baseline drift and noise. Its squared Hilbert envelope is calculated using a 15-ms FIR Hilbert transformer and smoothed using a 20-ms moving-median filter followed by a 2-ms moving-average filter to obtain the amplitude envelope. The autocorrelation of the bandpass-filtered EGG is calculated using 25-ms frames and one-sample frameshift. A sample is marked voiced if the amplitude is higher than one-fifth of its maximum value and the autocorrelation peak within the 2–10 ms delay (for 100–500 Hz pitch) is higher than one-

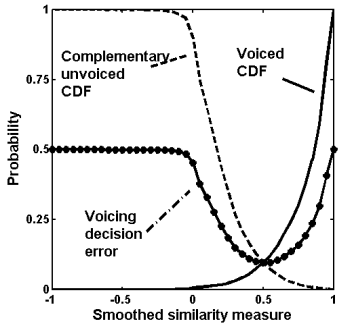

Figure 3. CDF for the voiced frames, complementary CDF for the unvoiced frames, and voicing decision error.

Table 2: Performance measures of excitation epoch detection (908,371 epochs across the speakers).

Speech	Technique	Performance measures						
		IDR (%)	MR (%)	FR (%)	IDB (ms)	IDA (ms)	A-0.25 (%)	AWIDR (%)
Clean speech	SEDREAMS	**97.42**	**1.86**	0.72	0.61	0.63	68.85	**83.28**
	ZFR	95.12	4.32	**0.57**	−0.82	**0.60**	50.22	81.96
	DYPSA	93.70	2.60	3.70	0.59	0.68	72.41	78.84
	MMF	91.39	3.15	5.46	0.50	0.80	**73.85**	74.79
	HEMSS	95.08	2.84	2.08	0.54	0.62	69.67	81.32
	HECEVD	93.91	2.90	3.19	**0.38**	0.62	72.93	80.49
Tele-phone-quality speech	SEDREAMS	92.32	**1.50**	6.18	0.38	0.81	44.78	74.80
	ZFR	74.13	2.65	23.22	−0.77	0.68	47.98	63.22
	DYPSA	92.83	3.28	3.88	0.37	0.67	**73.58**	78.41
	MMF	90.43	3.85	5.72	0.46	0.78	72.06	74.30
	HEMSS	92.77	3.78	3.45	0.57	0.65	66.07	79.15
	HECEVD	**92.85**	4.33	**2.81**	0.43	**0.61**	70.97	**79.85**

Table 3: Voicing detection performance measures (%).

Technique	Normal speech			Telephone-quality speech		
	UV-V	V-UV	VDE	UV-V	V-UV	VDE
PEFAC	**3.95**	9.26	13.21	5.35	9.86	15.21
SRH	5.90	**3.79**	9.69	6.30	**3.99**	10.28
HECEVD	4.22	5.04	**9.26**	**4.42**	5.44	**9.86**

third of its zero-delay value. On the resulting voicing decisions, a 20-ms median filter is applied to suppress isolated errors and output the reference voicing track. The use of this method on the CSTR database and comparison with the available reference pitch track showed a voicing decision error of 6.96% (1.33% UV-V error and 5.62% V-UV error). A visual examination of the tracks for the two databases did not show a significant anomaly.

The IESM histograms for the voiced and unvoiced frames in CSTR with its reference voicing track showed only a small overlap. For equal occurrences of the voiced and unvoiced frames, the voicing decision error for a threshold is mean of the cumulative distribution function (CDF) for the voiced frames and the complementary CDF for the unvoiced frames. Based on the plots in Figure 3, the threshold for minimum error is set as 0.5 and also used for CMU-ARCTIC.

For evaluation with telephone-quality speech, the speech signal in CMU-ARCTIC was bandpass-filtered according to ITU-T P.862 [46]. The excitation epoch detection performance was compared with four earlier techniques: ZFR [17], SEDREAMS [18], and MMF [21], DYPSA [24], and HEMSS [36]. The voicing decision performance was compared with two earlier techniques: SRH [27] with its implementation in [47] and PEFAC [29] with its implementation in [45].

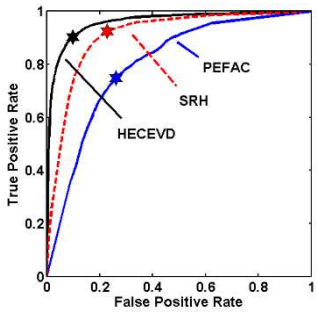

Figure 4. ROC curves of voicing detection techniques, with the thresholds for results in Table 3 marked by stars.

IV. Results

For result tabulation, the proposed technique is referred to as HECEVD (*H*ilbert *E*nvelope and *C*ovariance-Based *E*poch and *V*oicing *D*etection). The performance measures for the two databases used in the evaluation showed similar results. Table 2 shows the epoch detection performances for clean and telephone-quality speech for the signals in the CMU-ARCTIC database. In case of the clean speech, the IDR is 93.91% for HECEVD, while it is highest as 97.42% for SEDREAMS and lowest as 91.39% for MMF. The IDA is 0.62 ms for HECEVD, while it is lowest as 0.60 ms for ZFR. The AWIDR is 80.49% for HECEVD, similar to the other techniques. In case of the telephone-quality speech, HECEVD has the best performance measures. The voicing detection performance measures for the CMU-ARCTIC database are shown in Table 3. The VDE for HECEVD is 9.26% in case of the clean speech and 9.86% in case of the telephone-quality speech, and lower than the other two techniques.

In a threshold-based binary classifier, the errors may vary with the threshold setting and the class distribution. For examining it, the receiver operating characteristics (ROC) curve [48] was obtained by plotting the true-positive rate against the false-positive rate at various threshold settings, with the area under the ROC curve (AUC) as a performance index. The ROC curves for the three techniques for the CSTR database with the reference pitch track are shown in Figure 4. The AUC for HECEVD is 0.96, higher than 0.87 for SRH and 0.80 for PEFAC, indicating a lower sensitivity of its threshold to class distributions.

V. Conclusion

A technique has been presented for excitation epoch and voicing detection. It uses Hilbert envelope of the speech signal for epoch detection and a normalized covariance of the adjacent inter-epoch segments for voicing detection. It employs single-pass processing with an algorithmic delay of less than 60 ms. The epoch detection and the voicing detection for clean and telephone-quality speech showed a good match with those obtained from EGG, and the performances compared favorably with the earlier techniques. Its performance needs to be evaluated for speech signals with voice disorders. The algorithmic delay and computational requirements of the proposed technique need to be compared with state-of-the-art techniques employing single-pass processing for assessing its suitability in real-time applications.

Acknowledgements

The work is supported by the project "Visual Speech Training System Phase-2", MEITY, Government of India.

REFERENCES

[1] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, Berlin, Germany: Springer, 1965, pp. 119–164.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ, USA: Prentice-Hall, 1978, pp. 54–125.

[3] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Berlin, Germany: Springer-Verlag, 1983.

[4] D. O'Shaughnessy, *Speech Communication: Human and Machine*, 2nd ed. Piscataway, NJ, USA: IEEE Press, 2000.

[5] B. S. Atal, V. Cuperman, and A. Gersho, *Advances in Speech Coding*, Boston, MA, USA: Springer Science and Business Media, 1991.

[6] A. S. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, 1994.

[7] S. Ghaemmaghami, M. Deriche, and B. Boashash, "A new approach to pitch and voicing detection through spectrum periodicity measurement," in *Proc. TENCON 1997*, Brisbane, Australia, pp. 743–746.

[8] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Commun.*, vol. 8, pp. 147–158, 1989.

[9] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, 2017.

[10] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 1071–1079, 2010.

[11] N. P. Narendra and K. S. Rao, "Robust voicing detection and F0 estimation for HMM-based speech synthesis," *Circuits, Syst., Signal Process.*, vol. 34, no. 8, pp. 2597–2619, 2015.

[12] J. I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, "Automatic detection of voice impairments from text-dependent running speech," *Biomed. Signal Process. Control*, vol. 4, pp. 176–182, 2009.

[13] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: review of concepts and an insight to the state of the art," *Biomed. Signal Process. Control,* vol. 51, pp. 181–199, 2019.

[14] C. Middag, Y. Saeys, and J. P. Martens, "Towards an ASR-free objective analysis of pathological speech," in *Proc. Interspeech 2010*, Makuhari, Japan, pp. 294–297.

[15] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 201–212, 1976.

[16] A. Zolnay, R. Schlüter, and H. Ney, "Extraction methods of voicing feature for robust speech recognition," in *Proc. Eurospeech 2003*, Geneva, Switzerland, pp. 497–500.

[17] 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, 2008.

[18] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech 2009*, Brighton, UK, pp. 2891–2894.

[19] H. A. Patil and S. Viswanath, "Effectiveness of Teager energy operator for epoch detection from speech signals," *Int. J. Speech Tech.,* vol. 14, no. 4, pp. 321–337, 2011.

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

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

[22] K. Wu, D. Zhang, and G. Lu,"GMAT: glottal closure instants detection based on the multiresolution absolute Teager-Kaiser energy operator," *Digital Signal Process.,*vol. 69, pp. 286–299, 2017.

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

[24] 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, 2007.

[25] T. Drugman, M. Thomas, J. Gudnason, P. A. 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, 2012.

[26] D. J. Hermes, "Measurement of pitch by subharmonic summation, " *J. Acoust. Soc. Amer.*, vol. 83, pp. 257–264, 1988.

[27] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech 2011*, Florence, Italy, pp. 1973–1976.

[28] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. 1990*, Albuquerque, NM, USA, pp. 249–252.

[29] S. Gonzalez and M. Brooke, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE Trans. Audio, Speech,Lang. Process.,*vol. 22 , no. 2, pp. 518–528, 2014.

[30] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech, Audio, Process.*, vol. 1, no. 2, pp. 250–255, 1993.

[31] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech, Audio, Process.*, vol. 7, no. 3, pp. 333–338, 1999.

[32] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proc. Interspeech 2005*, Lisbon, Portugal, pp. 369–372.

[33] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Process. Lett*., vol. 17, no. 3, pp. 273–276, 2010.

[34] 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. Amer.*, vol. 135, no. 1, pp. 460–471, 2014.

[35] G. J. Lal, E. A. Gopalakrishnan, and D. Govind, "Glottal activity detection from the speech signal using multifractal analysis," *Circuits, Syst., Signal Process.*, vol. 39, no. 4, pp. 2118–2150, 2020.

[36] H. Dasgupta, P. C. Pandey, and K. S. Nataraj, "Epoch detection using Hilbert envelope for glottal excitation enhancement and maximum-sum subarray for epoch marking," *IEEE J. Selected Topics Signal Process.*, vol. 14, no. 2, pp. 461–471, 2019.

[37] U. Zölzer, *Digital Audio Signal Processing*, Chippenham, UK: Wiley,2008, pp. 225–239.

[38] J. Bentley, "Programming pearls: Algorithm design techniques,"*Commun. ACM*, vol. 27, no. 9, pp. 865-871, 1984.

[39] T. Takaoka, "Efficient algorithms for the maximum subarray problem by distance matrix multiplication," *Electron. Notes Theoretical Comput. Sci.*, vol. 61, pp. 191–200, 2002.

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

[41] P. Bagshaw, S. Hiller, and M. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," in *Proc. Eurospeech 1993* , Berlin, Germany, pp. 1003–1006.

[42] F. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech 1995*, Madrid, Spain, pp. 837–840.

[43] J. Kominek and A. W. Black, *CMU-ARCTIC Speech Database*. Accessed: Oct. 20, 2019. [Online] Available: festvox.org

[44] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection,"*IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[45] M. Brookes, *Voicebox toolbox*, Accessed: Mar. 27, 2020. [Online]. Available: ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[46] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Rec. ITU-T P.862, International Telecommunications Union, Geneva, Switzerland, 2001. [online]. Available: itu.int/rec/T-REC-P.862-200102-I/en

[47] T. Drugman, *Gloat Toolbox*, Accessed: Oct. 05, 2018. [Online]. Available: tcts.fpms.ac.be/~drugman/Toolbox/

[48] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Lett.,* vol. 27, pp. 861–874, 2006.