Robust Extraction of Epoch and Pitch Contour from Emotional Speech

Md. Shah Fahad
Department of Computer Science
Central University of South Bihar
Patna, India 800014
Email: shahfahad@cub.ac.in

Jainath Yadav
Department of Computer Science
Central University of South Bihar
Patna, India 800014
Email: jainath@cub.ac.in

K. Sreenivasa Rao
School of Information Technology
Indian Institute of Technology
Kharagpur, India
Email: ksrao@iitkgp.ac.in

***Abstract*—The goal of our work is to enhance the performance of epoch detection in the case of emotional speech. Existing epoch estimation methods requires either modeling of the vocal-tract system or a priori information of the average pitch period. The performance of existing epoch estimation methods degrades significantly due to rapid variation of the pitch period in emotional speech. In the present work, we have utilized the advantage of zero time windowing method that provides instantaneous spectral features. These features show high spectral peaks near the epoch location. We compute sum of the three prominent spectral peaks at each sampling instant of the HNGD spectrum, which is called spectral energy profile. Epoch evidence is highlighted by convolving spectral energy profile with a Gaussian filter. Further, the positive peaks are selected to identify accurate epoch locations from epoch evidence. The experimental result shows that the average identification rate, root mean squared error and mean identification accuracy of epoch using proposed method is 95.44%, 0.43 ms and 0.30 ms, respectively in emotional speech which are significantly better than other existing methods. This method works well even for the aperiodic nature of the speech signal and is robust against signal degradation.**

***Index Terms*—Epoch extraction/detection, Emotional speech, Hilbert envelope of the Numerator Group Delay (HNGD), Zero frequency filter (ZFF), Zero time windowing (ZTW)**

I. Introduction

In recent years mobile phone market has grown exponentially, an estimated 62.9 percent of the population worldwide already owned a mobile phone and the number of mobile phone users in the world is expected to pass the five billion mark by 2019.As the usage of mobile phones increased and the cost to exchange Short Messaging Service(SMS) has decreased, SMS exchanged has grown rapidly. Short Messaging Service(SMS)is a text messaging service that allows mobile phone users to exchange short text messages usually of length less than 160 characters.

As Popularity of both mobile phones and SMS have increased, Mobile phones are becoming the main target for spammers using SMS. Spams are unsolicited messages sent electonically that may either be harmfull or commerical advertisements.

SMS are usually of different languages based, and the models available are not often applicable across distinct languages.Therefore a classification model that is not restricted for single language but for mulitple languages need to be applied.

Previous work on cross language classification uses mainly machine translation techniques and using different classifiers for differnet languages.

Machine translation method works by translating source language into target language and then applying various machine learning classification algorithms.But, several problems may occur after translation like actuall meaning in the source language may changed after translating to target language, accurate translation may not happen in all the cases this may drastically reduce the performance of the classification model. Even if translation is accurately done the main disadvantage is availabilty of open-source translatorsas most of the robust translation APIs are not free of cost. Translating large corpus may also makes it too expensive.

Using different classifiers for differnet languages the main drawback is that every message needs to be identified and then sent to their respective classifier.

we are proposing a model where we used only one classifier for different types of languages.

II. Emotional speech database

There are two databases that are used for evaluating the performance of epoch estimation namely database of German emotional speech [3] and Hindi emotional speech. Both emotional speech databases have simultaneously recording electro-glotto-graphy (EGG) signal. The pattern of the vibration of vocal cords is directly recorded by EGG, and it is used as reference for automatic epoch extraction methods. The German emotional speech database has 10 speakers (5 male, 5 female) and 10 sentences are recorded for seven emotions (Neutral, Happy, Angry, Sad, Boredom, Disgust and Fear). There are about 800 (10 speakers *10 sentences *7 emotions + some second versions) speech utterances in German emotional speech database.

The Hindi emotional speech database is collected by professional artist from Bhagalpur akashvani and Premchandra rangshala (Patna), India. The speech signal are recorded using α TechCadenza product electro-glotto-graph device and SHURE dynamic cardioid microphone with 16 kHz sampling frequency with the help of Cool Record Edit

Pro software. The Hindi emotional speech database has 10 speakers (5 male, 5 female) and 15 sentences are recorded for five emotions (Neutral, Happy, Angry, Sad and Fear). There are total 750 (10 speakers *15 sentences *5 emotions) speech utterance in Hindi emotional speech database.

III. MOTIVATION

Zero Time Windowing (ZTW) method has been proposed to extract instantaneous spectral features. This method is robust against all type of sound units even for semivowels, trills, fricatives, voice bar, burst and aspiration. ZTW method also works well for different type of signal degradation such as babble, vehicle and white noise. The main contribution of this method is that it uses impulse like window and numerator group delay function for each sampling instant and identify spectral features of each instant. It has been observed that spectral features is more strong and exhibits dynamic behavior at glottal closure instants other than their neighboring instants. This is the main reason which encouraged us to use the ZTW method for epoch detection. Some existing epoch detection methods require average pitch period of the utterance. The pitch period varies rapidly in an utterance of emotional speech. Therefore, these methods are not suitable for emotional speech. We have proposed a novel method using ZTW method that does not require information about average pitch period.

IV. PROPOSED METHOD

In this method, voiced regions are detected using combined evidences of zero crossing rate (ZCR), magnitude sum function (MSF) and pitch period. After, ZTW method is applied to get HNGD spectra of each voiced segments. The amplitude of the sum of the three prominent peaks is obtained from each spectrum of the HNGD. The resulting output reproduces the instantaneous energy profile of the windowed signal. The spectral energy profile, obtained from HNGD spectrum shows high energy at the epoch locations because of high SNR at this location, and it is normalized by mean smooth filter. Further, the normalized spectral energy profile is convolved with a Gaussian filter to highlight the peaks. The positive peaks are selected after removing the spurious peaks which are considered as epochs.

A. Voiced Activity Detection (VAD)

Epochs are present in the voiced regions due to vibration of the vocal cords. Hence, we first discriminate the speech into voiced and unvoiced regions based on its characteristics. VAD is identified based on the principle that if the extracted features from the speech frame exceeds the threshold value then the frame is marked as voiced region (VAD=1) otherwise it is marked as unvoiced region (VAD=0). In general, a VAD algorithm results a binary decision by considering the short frame length of 20-30 ms of input speech signal. In the present work, voiced regions are detected [12] using the combined evidences from ZCR, MSF and pitch period.

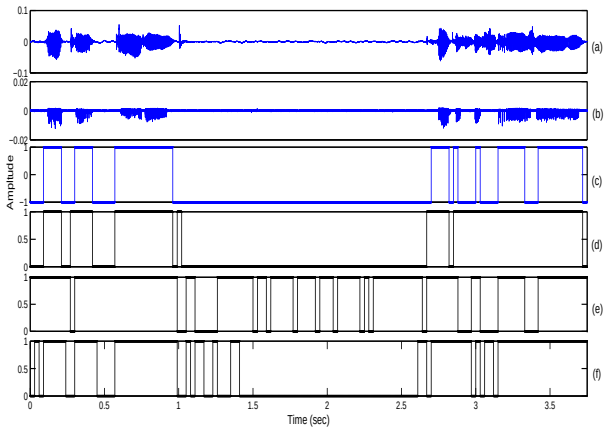

Fig. 1. Detection of voiced and unvoiced regions. (a) Speech signal. (b) Differenced EGG signal. (c) Detected voiced and unvoiced regions using evidences from ZCR, MSF and pitch period. (d) Detected voiced and unvoiced regions using evidence from MSF only. (e) Detected voiced and unvoiced regions using evidence from ZCR only. (f) Detected voiced and unvoiced regions using evidence from pitch period only.

- **Zero Crossing Rate (ZCR):** ZCR for speech signal is defined as the number of times the amplitude of the signal changes its sign. ZCR can be calculated as follows:

$$Z_{cr} = \sum_{m=1}^{n} |sgn[x(m)] - sgn[x(m-1)]| \quad (1)$$

  The number of ZCR follows a definite range in the voiced speech. ZCR in unvoiced regions is higher than voiced regions.
- **Magnitude Sum Function (MSF):** The MSF gives the short time energy of the speech signal. It is calculated as follows:

$$MSF = \sum_{m=1}^{n} s(n) \quad (2)$$

  The energy of the voiced frame is more than the energy of the unvoiced frames due to lower frequency. In unvoiced speech, most of the energy is present at higher frequencies. Since low frequencies imply low zero-crossing rates and high frequencies imply high zero crossing rates, there is a strong correlation between energy distribution and zero-crossing rate with frequency.
- **Pitch Period:** Pitch period can be estimated using the autocorrelation function. Pitch period in voiced speech follows regular pattern, but in unvoiced speech it is random. Using only one evidence, it may fail in some cases. Therefore, we have combined the three evidences to get the reliable voiced segment. The threshold values are decided empirically by experimental result. Fig. 1 shows the discriminated voiced-unvoiced regions.

Fig. 1(a) shows the speech signal and its corresponding differenced EGG signal in Fig. 1(b). The EGG signal gives

the basis for well voiced-unvoiced discrimination, because the differenced EGG signal is around zero during unvoiced speech segment where the glottis is always open. The decision of voiced and unvoiced speech has been taken using pitch period, ZCR and MSF, respectively in the Fig. 1(d), (e) and (f). Fig. 1(c) shows detection of voiced regions using combined evidences of ZCR, MSF and pitch period. Detection of voiced regions using combined evidences of ZCR, MSF and pitch period is approximately same as using differenced EGG signal.

B. Zero Time Windowing (ZTW)

Zero time windowing method [2] [16] is used to capture the spectral feature with good temporal resolution. The speech signal is multiplied by a window function that gives more weight to the samples around the starting sampling instants known as zero time. The effect of the decaying window function $h_1[n]$ in the time domain is approximately same as the zero frequency filtering in the frequency domain. The speech signal is differenced using sampling frequency 16 kHz to reduce any low frequency biased in the speech signal. Speech segment of 3 ms is considered (i.e., M = 48 samples) at each sampling point and it is appended enough number of zeros to make its size equal to DFT length (N). The windowed signal is computed by multiplying the N samples segment with window functions $h_1^2[n]$. The window function h_1 is defined as:

$$h_1[n] = \begin{cases} 0 & n = 0 \\ h_1[n] = \frac{1}{4sin^2(\frac{\pi n}{N})} & n = 1, 2, .., N-1 \end{cases} \quad (3)$$

where $s[n]$ is the differentiated speech of length N samples. $h_1^2[n]$ is the twice of the window function $h_1[n]$ that does not smear spectral feature like any other arbitrary window. The formant feature of spectrum is highlighted by numerator of the group delay function of $x[n]$ due to its additive and high resolution property. The numerator group delay function $g[k]$ of a signal $x[n]$ is computed by

$$g[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k], k = 0, 1, 2..., N-1 \quad (4)$$

where $X_I[k]$ and $X_R[k]$ are the imaginary and real parts of the N-point DFT $X[k]$ of $x[n]$, respectively. $Y_I[k]$ and $Y_R[k]$ are the imaginary and real parts of the N-point DFT $Y[k]$ of $y[n] = nx[n]$, respectively. To highlight the peaks in the spectrum, double differencing of NGD function is performed. The low amplitude formant do not seem well because of some spectral valleys. To overcome this problem , we compute the Hilbert envelope of the differenced NGD is known as HNGD spectrum.

C. Sequence of Steps for Epoch Extraction From Emotional Speech

The steps for proposed method to extract epoch location can be summarized as follows:

1) Determine the voiced segment using the combined information obtained from ZCR, MSF and average pitch period.
2) Differenced the voiced speech signal to remove any low frequency biased in the speech signal at sampling frequency 16 kHz. i.e.
$$y[n] = s[n] - s[n-1] \quad (5)$$
3) Take 3 ms segments of the differenced speech signal(i.e, M = 48 samples) at each sampling point and append it with $N - M$ (2048-48) zeros to obtain sufficient resolution in the frequency domain.
4) Multiply the time domain signal with window function two times h_1^2 to achieve the smoothed spectrum by integration in the frequency domain. The window function h_1 is defined as:
$$h_1[n] = \begin{cases} 0 & n = 0 \\ h_1[n] = \frac{1}{4sin^2(\frac{\pi n}{N})} & n = 1, 2, .., N-1 \end{cases} \quad (6)$$
5) The ripple effect due to truncation is reduced by multiplying the window h_2. The resultant signal $x[n]$ is called windowed signal. Window function $h_2[n]$ is defined as:
$$h_2[n] = 4cos^2(\frac{\pi n}{2M}), n = 0, 1, 2..., M-1 \quad (7)$$
6) To highlight the spectral features, compute the numerator group delay $g[k]$ of windowed signal. It is further required double differencing operation to remove the trend. The resultant signal is known as DNGD signal. The numerator group delay $g[k]$ is computed as follows:
$$g[k] = X_R[k]Y_R[k] + X_I[k]Y_R[k], k = 0, 1, 2..., N-1 \quad (8)$$
7) Compute the Hilbert envelope of the DNGD spectrum to show the spectral peaks prominently. The Hilbert envelop $h_e[k]$ of DNGD signal $g[k]$ is computed as
$$h_e[k] = \sqrt{g^2[k] + g_h^2[k]} \quad (9)$$
where $g_h[k]$ is the Hilbert transform of the sequence $g[k]$, and obtained as follows:
$$g_h[k] = IDFTE_h(w) \quad (10)$$
where
$$E_h(\omega) = \begin{cases} -jE(\omega), & 0 < \omega < \pi \\ jE(\omega, & -\pi < \omega < 0 \end{cases} \quad (11)$$
and $E(\omega)$ is the DTFT of the sequence $g(k)$.
The HNGD spectrum of a voiced speech segment is plotted in Fig. 2. The voiced speech segment is shown in Fig. 2(a) and its corresponding HNGD spectrum is shown in Fig. 2(b).
8) Determine the sum of the three prominent peaks of the HNGD spectrum at each sampling instant. The resultant amplitude shows high SNR around glottal

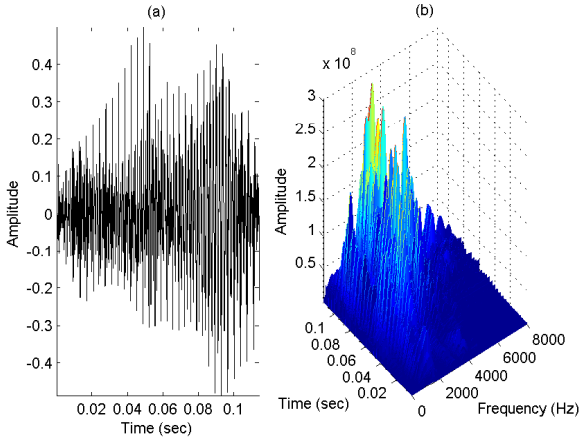

Fig. 2. HNGD plot of voiced speech segment. (a) Speech signal. (b) HNGD spectrum.

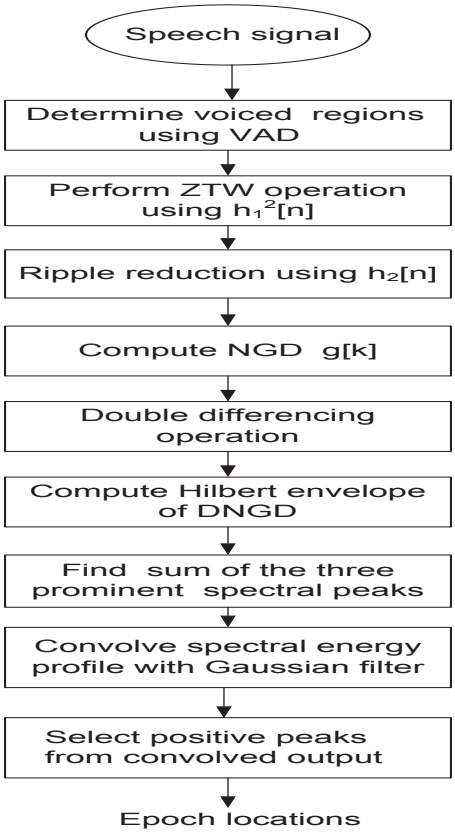

Fig. 3. Schematic block diagram of the proposed method.

closure. Further, smooth the amplitude contour using 5-point mean smoothing filter to eliminate any outliers.

9) The sum of the three prominent peaks obtained from each HNGD spectra is called spectral energy profile. The spectral energy profile is convolved with a Gaussian filter of size 2 m sec for high arousal emotions like Happy and Angry and 6 m sec for low arousal emotions like Fear and Sad to emphasize spectral peaks. The convolved output is called epoch evidence plot. A Gaussian filter of length L is given by

$$G[n] = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}}, n = 1, 2, ..., L \qquad (12)$$

The standard deviation σ in this work is $\frac{1}{4^{th}}$ of Gaussian filter length.

10) The spurious peaks are eliminated by using following sub steps:
   (a) The spurious peaks are eliminated on the basis that the difference between two peaks are not less than 2 ms. Because, it is the minimum range of the pitch period. If this condition follows then remove the peak location which has less amplitude.
   (b) It is also bounded that there should be a negative region between two successive peaks. This criteria also eliminates some spurious peak locations.

11) The positive peaks in epoch evidence plot represent epoch locations.

Steps of the proposed method are shown in the schematic block diagram of the Fig. 3. Epoch detection using proposed method is shown in Fig. 4. The Angry emotional speech segment is shown in Fig. 4(a) and its differenced EGG signal in Fig. 4(b). The spectral energy profile obtained from HNGD spectrum of the speech signal using ZTW analysis is plotted in Fig. 4(c). The epoch evidence plot after convolving spectral energy profile with a Gaussian window of 2 m sec is shown in Fig. 4(d). Epoch locations are shown in Fig. 4(e).

Epoch detection using proposed and ZFF methods are shown in Fig. 5. The voiced Sad speech segment is shown in Fig. 5(a) and its corresponding EGG signal in Fig. 5(b). Zero frequency filtered signal and its corresponding positive zero crossings which are considered as epochs are shown in Fig. 5(c) and Fig. 5(d), respectively. Epoch evidence obtained from proposed method and the positive peak locations which are considered as epochs are shown in Fig. 5(e) and Fig. 5(f), respectively. From Fig. 5, it is observed that spurious epochs present in ZFF method are eliminated in the proposed method.

V. Performance of the Proposed Epoch Extraction Method for Emotional Speech

The performance of the proposed method is tested on German emotional speech database and Hindi emotional speech databases for five emotions (Neutral, Happy, Angry, Sad and Fear) with simultaneous EGG signal. There are 500 and 750 utterances are taken for evaluation of German

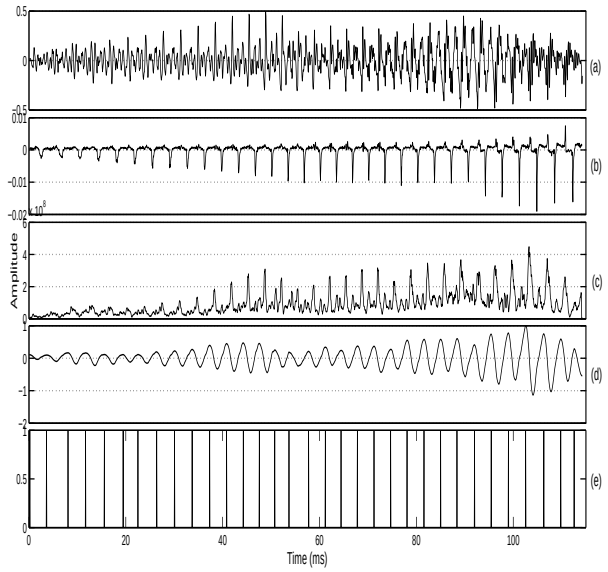

Fig. 4. Epoch extraction using proposed method. (a) Angry speech segment. (b) Differenced EGG signal. (c) spectral energy profile obtained from HNGD spectrum. (d) Epoch evidence plot. (e) Epoch locations.

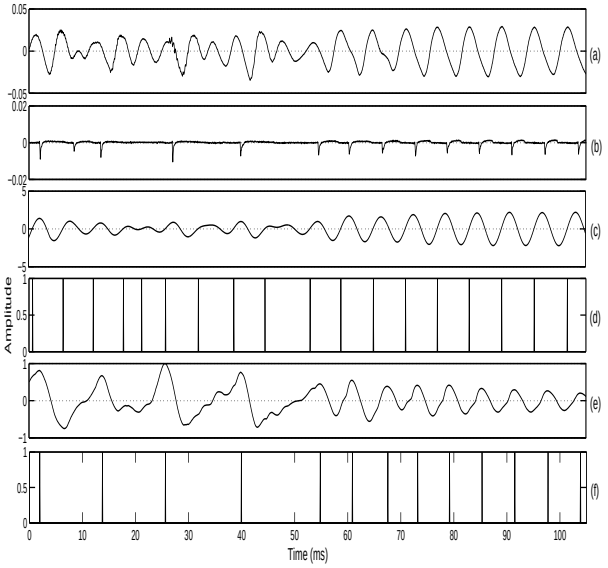

Fig. 5. Epoch extraction using ZFF and proposed methods. (a) Sad speech segment. (b) Differenced EGG signal. (c) ZFF signal. (d) Epoch locations using ZFF method. (e) Epoch evidence plot. (f) Epoch locations using proposed method.

emotional speech database and Hindi emotional speech databases, respectively. Fig. 6 shows the characterization of epoch estimation parameters showing each of the possible decisions from the epoch detection algorithms. The following measures are in [13] used to evaluate the performance of epoch extraction method.

- **Glottal Cycle:** The range of samples $\frac{1}{2}(g_{r-1}+g_r) \leq n \leq \frac{1}{2}(g_r+g_{r+1})$ given an epoch reference at sample g_r with preceding and succeeding epoch references at samples g_{r-1} and g_{r+1}, respectively.
- **Identification Rate (IDR):** The percentage of glottal cycles for which exactly one epoch is detected.
- **Miss Rate (MR):** The percentage of glottal cycles for which epoch is not detected.
- **False Alarm Rate (FAR):** The percentage of glottal cycles for which more than one epoch is detected.
- **Identification Error ζ :** The timing error between the reference epoch location and the detected epoch location in glottal cycles for which exactly one epoch is detected. In this paper, identification error is calculated in terms of root mean squared error (RMSE).
- **Identification Accuracy (IDA) σ :** The standard deviation of the identification error ζ . Small values of σ indicate high accuracy of identification ζ.

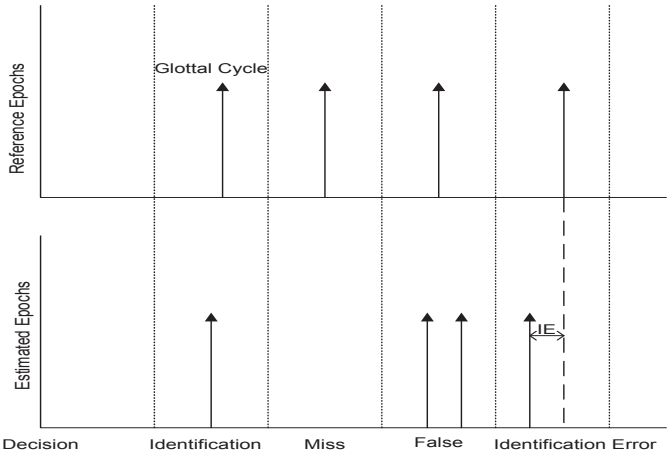

Fig. 6. Characterization of epoch estimation parameters showing 5 glottal cycles with examples of each possible outcome from epoch extraction.

In this work, performance of the proposed method for epoch extraction is compared with existing methods DYPSA, ILPR and ZFF. The performance of DYPSA, ILPR, ZFF, and the proposed methods is shown in Table I on German emotional database and in Table II on Hindi emotional database.

Column 1 of both the table indicates different emotions Neutral, Angry, Fear, Happy and Sad. Column 2 indicates the different epoch extraction methods for analysis. Columns 3 to 7 indicate different parameters IDR, MR, FR, RMSE and IDA, respectively for evaluating the performance. The performance evaluation using IDR, MR, and FAR is represented in terms of percentage. Performance evaluation using RMSE and IDA is represented in terms of millisecond. Less value of MR, FAR, RMSE and IDA means better performance. From Table I and II, it is observed that both German and Hindi emotional database have similar trend in performance. The graphical representation of all the five epoch estimation parameter is also plotted. For

TABLE I
PERFORMANCE EVALUATION OF EPOCH EXTRACTION BY PROPOSED AND OTHER METHODS USING NEUTRAL, ANGRY, HAPPY, FEAR AND SAD EMOTIONAL SPEECH UTTERANCE FROM GERMAN EMOTIONAL SPEECH DATABASE

Emotion	Method	IDR-(%)	MR-(%)	FAR-(%)	RMSE-(ms)	IDA-(ms)
Neutral	DYPSA	96.37	1.80	1.83	0.77	0.45
	ILPR	98.97	0.01	1.02	0.69	0.40
	ZFF	99.06	0.18	0.76	0.58	0.32
	Proposed	98.88	1.07	0.05	0.39	0.27
Angry	DYPSA	83.36	4.55	12.09	0.79	0.52
	ILPR	86.01	10.59	3.40	0.71	0.47
	ZFF	88.46	0.15	11.39	0.63	0.36
	Proposed	96.04	3.87	0.09	0.46	0.30
Fear	DYPSA	84.37	9.47	6.16	0.85	0.54
	ILPR	86.99	8.48	4.53	0.77	0.42
	ZFF	88.05	3.25	8.70	0.69	0.34
	Proposed	93.09	6.78	0.13	0.55	0.33
Happy	DYPSA	86.77	8.97	4.26	0.80	0.47
	ILPR	84.63	12.01	3.36	0.72	0.52
	ZFF	87.33	3.21	9.46	0.68	0.32
	proposed	96.29	3.57	0.14	0.44	0.28
Sad	DYPSA	86.22	10.48	3.30	0.82	0.48
	ILPR	87.03	10.58	2.39	0.73	0.49
	ZFF	88.05	3.57	8.38	0.67	0.33
	Proposed	93.23	6.69	0.08	0.43	0.29

TABLE II
PERFORMANCE EVALUATION OF EPOCH EXTRACTION BY PROPOSED AND OTHER METHODS USING NEUTRAL, ANGRY, HAPPY, FEAR AND SAD EMOTIONAL SPEECH UTTERANCE FROM HINDI EMOTIONAL SPEECH DATABASE

Emotion	Method	IDR-(%)	MR-(%)	FAR-(%)	RMSE-(ms)	IDA-(ms)
Neutral	DYPSA	96.42	1.77	1.81	0.80	0.43
	ILPR	98.37	0.02	1.61	0.69	0.39
	ZFF	99.16	0.16	0.68	0.56	0.31
	Proposed	98.01	1.93	0.06	0.36	0.28
Angry	DYPSA	84.59	4.28	11.13	0.82	0.53
	ILPR	86.43	10.19	3.38	0.76	0.46
	ZFF	88.03	0.17	11.80	0.64	0.36
	Proposed	96.13	3.74	0.13	0.42	0.31
Fear	DYPSA	83.89	10.29	5.82	0.84	0.53
	ILPR	87.00	8.53	4.47	0.76	0.42
	ZFF	87.79	3.24	8.97	0.64	0.35
	Proposed	93.27	6.64	0.09	0.53	0.34
Happy	DYPSA	85.72	9.47	4.81	0.82	0.46
	ILPR	85.21	12.63	2.16	0.73	0.50
	ZFF	86.93	3.18	9.89	0.61	0.34
	proposed	96.14	3.77	0.09	0.41	0.29
Sad	DYPSA	85.61	10.77	3.62	0.78	0.46
	ILPR	87.15	10.48	2.37	0.66	0.51
	ZFF	87.85	3.27	8.88	0.59	0.33
	Proposed	93.33	6.53	0.14	0.40	0.30

graphical representation average of evaluation parameters of both German and Hindi emotional speech is taken. Fig.7 shows identification rate (IDR) of five emotions for different methods.

All the existing methods give better result for neutral speech but the performance degrades significantly in the emotional speech. Identification accuracy (IDA) of epoch extraction is more in the proposed method compared to existing methods. The identification rate (IDR) of epoch extraction is significantly increases in the proposed method. The average identification rate of epoch using proposed method is 95.44% for emotional speech.

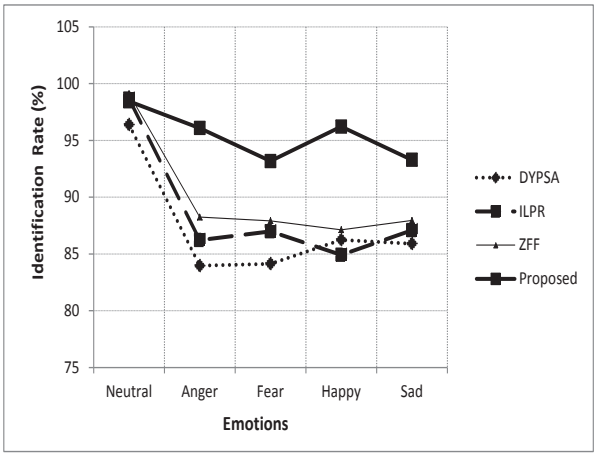

Fig. 7. Identification Rate (IDR) of the four epoch extraction methods across the five emotions

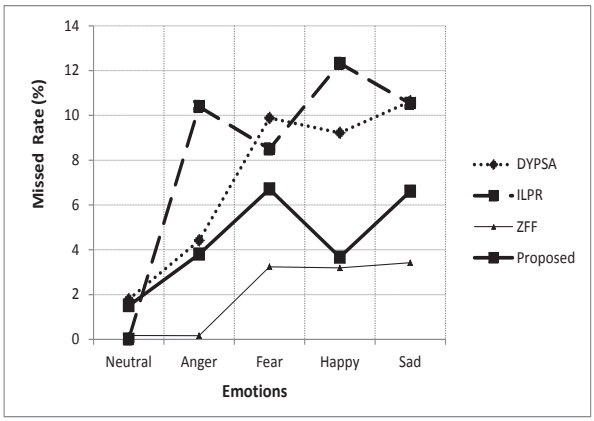

Fig. 8. Missed Rate (MR) of the four epoch extraction methods across the five emotions

IDR of positive emotions like Angry and Happy is higher compared to negative emotions like Fear and Sad in the proposed method. The proposed method is dependent on the spectral energy contour. The energy content of Angry and Happy emotions is more compared to Fear and Sad emotions. Sometimes the energy of voiced consonant is very low in the Sad and Fear emotions. Therefore, epoch locations may be missed in the case of Sad and Fear emotions. The graphical representation of missed rate for different methods is shown in Fig. 8. The average missed rate using proposed method is 4.48%.

The false alarm rate (FAR) decreases significantly in the proposed method due to strong assumption used to eliminate false epoch locations. False alarm rate across all emotions for different methods is shown in Fig.9. The average false

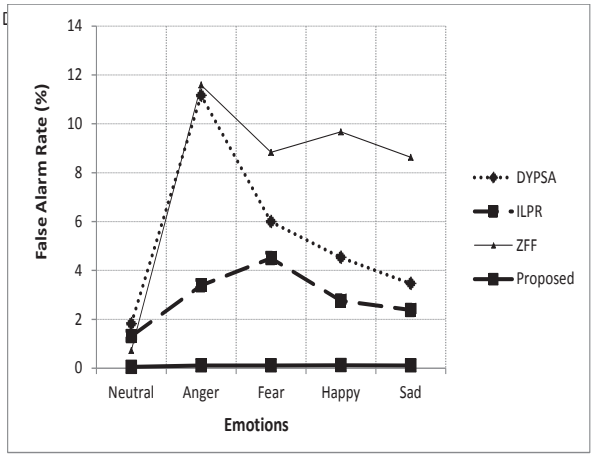

Fig. 9. False Alarm Rate (FAR) of the four epoch extraction methods across the five emotions

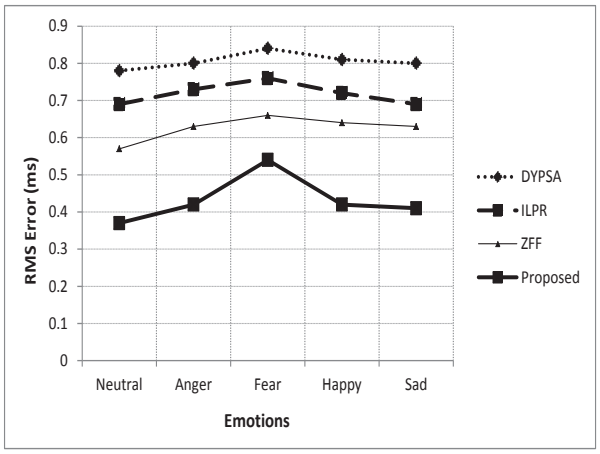

Fig. 11. Root mean squared (RMS) error of the four epoch extraction methods across the five emotions

alarm rate of the proposed method is only 0.10%. Fig. 10 shows the combined identification accuracy (IDA) of

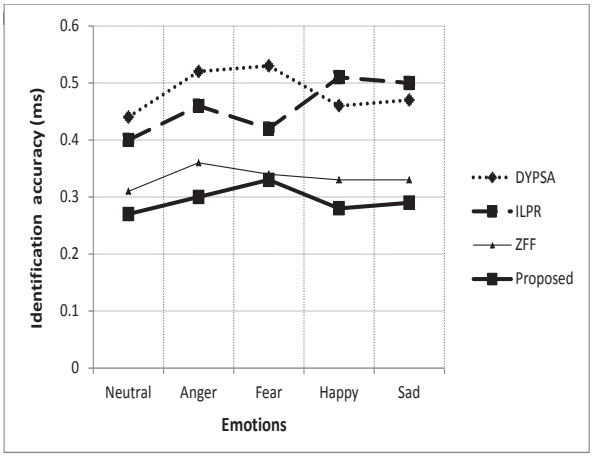

Fig. 10. Identification accuracy (IDA) of the four epoch extraction methods across the five emotions

both the German and Hindi databases for five emotions. IDA is calculated using the three existing standard epoch extraction methods and our proposed method. The figure clearly shows that the identification accuracy is better for the proposed method as compared to existing methods. IDA varies with different emotions but it remains more accurate using proposed method for each of the different emotions other than existing methods. The combined root mean squared (RMS) error of both the German and Hindi databases for five emotions is shown in Fig. 11. The RMS error also varies with different emotions but the RMS error of proposed epoch extraction method is less for each of the different emotions than existing methods. The identification accuracy (IDA) and RMS error are more in DYPSA and ILPR methods because these methods perform Hilbert envelope on the LP-residual signal. The Hilbert envelope operation shifts the LP-residual signal from actual location. ZFF method is independent of LP-residual signal and Hilbert envelope operation. Therefore, both IDA and RMS error are less in this method. These parameters are also affected in the ZFF method with the local mean. If the calculated local mean is more or less than actual mean that occurs in emotional speech then the epoch locations are not accurate. The above discussed problem does not occur in the proposed method. Therefore, the proposed method detects accurate epoch locations. The experimental result shows that there is a significant improvement in the performance of the proposed method in emotional speech from other methods like DYPSA, ILPR and ZFF in terms of IDR, MR, FAR and IDA.

Further, proposed method also performs well in low voiced speech regions while detecting epoch locations. Even though this method is dependent on spectral energy contour, performance of the proposed method is not affected as much as other methods based on energy contour. This is due to the zero time windowing of the speech signal. Epoch detection using proposed method is shown in Fig. 12. The Angry emotional speech segment having low voiced region is shown in Fig. 12(a) and its differenced EGG signal in Fig. 12(b). The spectral energy profile obtained from HNGD spectrum of the speech signal using ZTW analysis is plotted in Fig. 12(c). The epoch evidence plot after convolving spectral energy profile with a Gaussian window of 2 m sec is shown in Fig. 12(d). Epoch locations are shown in Fig. 12(e). In Fig. 12, the amplitude of some peaks are too low that it is not visible in the epoch evidence plot. But, the epoch locations are detected accurately.

A. Extraction of pitch contour using proposed method

The instantaneous pitch period is the duration between two successive epoch locations. Pitch can be obtained as the reciprocal of the pitch period [9] [14]. The pitch contours derived from estimated epoch locations using ZFF and

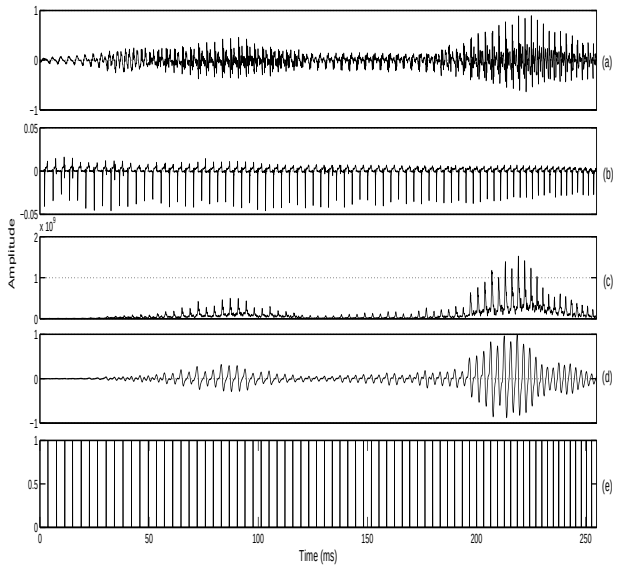

Fig. 12. Epoch extraction using proposed method. (a) Angry speech segment. (b) Differenced EGG signal. (c) Amplitude contour obtained from HNGD spectrum. (d) Epoch evidence plot. (e) Epoch locations.

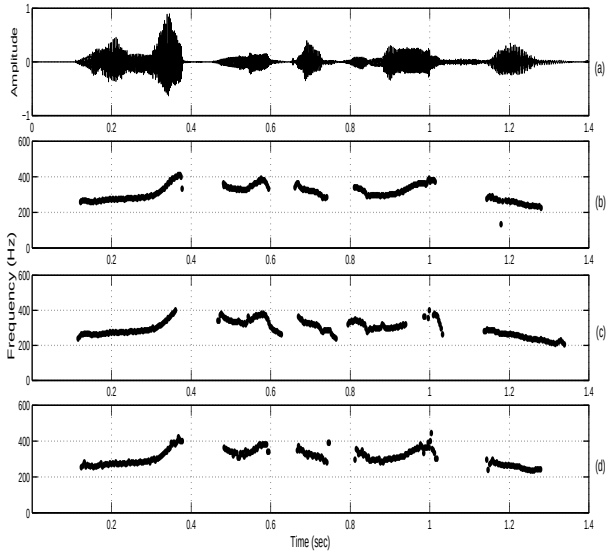

Fig. 13. Pitch contour using ZFF and proposed method. (a) Angry speech signal. (b) Pitch contour of the differenced EGG signal. (c) Pitch contour using ZFF method. (d) Pitch contour using proposed method.

proposed methods for Angry emotional speech signals are shown in Fig. 13. The pitch contour derived using ZFF and proposed methods are compared by referenced EGG pitch contour. Fig. 13(a) shows Angry speech utterance and pitch contour derived from its EGG signal is shown in Fig. 13(b). The pitch contour derived using ZFF and proposed methods are shown in Fig. 13(c) and 13(d), respectively. It is clear from figure that the pitch contour derived from proposed method is almost similar to the referenced EGG pitch contour. The pitch contour derived from ZFF method does not show similar trend. Hence, the proposed method for epoch estimation is useful for both better recognition and synthesis of emotion.

B. Extraction of SOE Contour using Proposed method

The strength of excitation (SOE) is determined as the difference between two successive epoch values. The SOE contours derived from estimated epoch locations using ZFF and proposed methods for Angry emotional speech signals are shown in Fig. 14. The SOE contour derived using ZFF and proposed methods are compared by speech signal. It is clear from figure that the SOE contour derived from proposed method is almost similar to the speech signal. The pitch contour derived from ZFF method does not show similar trend. Hence, the proposed method for SOE estimation is useful for both better recognition and synthesis of emotion.

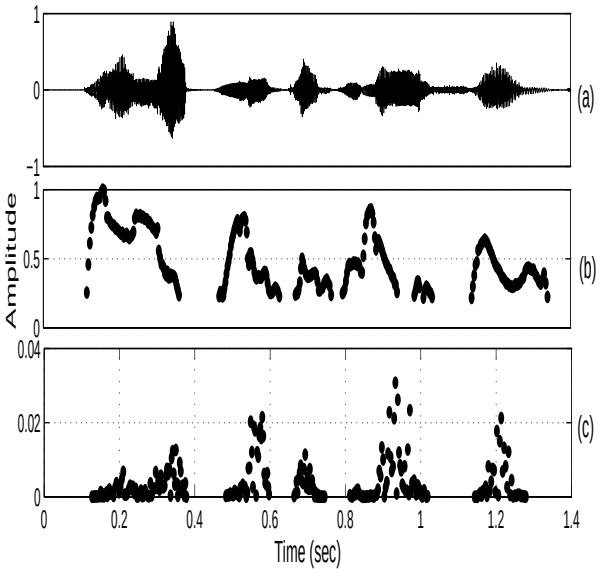

Fig. 14. SOE contour using ZFF and proposed method. (a) Angry speech signal. (b) SOE contour using ZFF method. (c) SOE contour using proposed method.

VI. CONCLUSION

Our present work identifies the problem of estimating the epoch in the emotional speech. We have proposed a novel method using ZTW analysis that overcomes the problem because this method is not based on the assumption of quasi-periodicity of the signal. Epoch evidence is obtained by convolving amplitude of the sum of three prominent peaks from the HNGD spectrum with a Gaussian filter. The performance evaluation shows robustness of this method for accurate and reliable epoch estimation for emotional speech. This method neither require modeling of the system, nor

depends on a priori pitch period. This method works well even in the case of significant degradation. Our future work is to explore proposed method for emotion conversion and emotion recognition systems.

REFERENCES

[1] Tirupattur V Ananthapadmanabha and B Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(4):309–319, 1979.

[2] Yegnanarayana Bayya and Dhananjaya N Gowda. Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. *Speech Communication*, 55(6):782–795, 2013.

[3] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[4] KT Deepak and SRM Prasanna. Epoch extraction using zero band filtering from speech signal. *Circuits, Systems, and Signal Processing*, pages 1–25, 2014.

[5] Thomas Drugman and Thierry Dutoit. Glottal closure and opening instant detection from speech signals. In *Interspeech*, pages 2891–2894, 2009.

[6] Christer Gobl, Ailbhe Ni, et al. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1):189–212, 2003.

[7] D Govind, SR Mahadeva Prasanna, and K Ramesh. Improved method for epoch extraction in high pass filtered speech. In *India Conference (INDICON), 2013 Annual IEEE*, pages 1–5. IEEE, 2013.

[8] D Govind and SRM Prasanna. Epoch extraction from emotional speech. In *Signal Processing and Communications (SPCOM), 2012 International Conference on*, pages 1–5. IEEE, 2012.

[9] Shashidhar G Koolagudi, Raghu Reddy, and K Sreenivasa Rao. Emotion recognition from speech signal using epoch parameters. In *Signal Processing and Communications (SPCOM), 2010 International Conference on*, pages 1–5. IEEE, 2010.

[10] Vinay Kumar Mittal and B Yegnanarayana. Significance of aperiodicity in the pitch perception of expressive voices. In *INTERSPEECH*, pages 504–508, 2014.

[11] Vinay Kumar Mittal, B Yegnanarayana, and Peri Bhaskararao. Study of the effects of vocal tract constriction on glottal vibration. *The Journal of the Acoustical Society of America*, 136(4):1932–1941, 2014.

[12] Syed Mohamed, Syed Abd Rahman Al-Hadad, M Iqbal Saripan, Shyamala C Doraisamy, Abd Ramli, Mohammad Ali Nematollahi, et al. A method for speech watermarking in speaker verification, 2013.

[13] K Murty and B Yegnanarayana. Epoch extraction from speech signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(8):1602–1613, 2008.

[14] NP Narendra and K Sreenivasa Rao. Robust voicing detection and f_{0} estimation for hmm-based speech synthesis. *Circuits, Systems, and Signal Processing*, 34(8):2597–2619, 2015.

[15] Patrick Naylor, Anastasis Kounoudes, Jon Gudnason, Mike Brookes, et al. Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):34–43, 2007.

[16] RaviShankar Prasad and B Yegnanarayana. Acoustic segmentation of speech using zero time liftering (ztl). In *INTERSPEECH*, pages 2292–2296, 2013.

[17] AP Prathosh, TV Ananthapadmanabha, and AG Ramakrishnan. Epoch extraction based on integrated linear prediction residual using plosion index. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(12):2471–2480, 2013.

[18] Roel Smits and B Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *Speech and Audio Processing, IEEE Transactions on*, 3(5):325–333, 1995.

[19] Mark RP Thomas, Jon Gudnason, and Patrick A Naylor. Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):82–91, 2012.