Circuits Syst Signal Process (2018) 37:3245–3274
https://doi.org/10.1007/s00034-018-0804-x

Epoch Estimation from Emotional Speech Signals Using Variational Mode Decomposition

G. Jyothish Lal[1] · **E. A. Gopalakrishnan**[1] · **D. Govind**[1]

Received: 4 August 2017 / Revised: 13 March 2018 / Accepted: 15 March 2018 /
Published online: 20 March 2018

Abstract This paper presents a novel approach for the estimation of epochs from the emotional speech signal. Epochs are the locations of significant excitation in the vocal tract during the production of voiced sound by the vibration of vocal folds. The estimation of epoch locations is essential for deriving instantaneous pitch contours for accurate emotion analysis. Many well-known algorithms for epoch extraction are found to show degraded performance due to the varying nature of excitation characteristics in the emotional speech signal. The proposed approach exploits the effectiveness of a new adaptive time series decomposition technique called variational mode decomposition (VMD) for the estimation of epochs. The VMD algorithm is applied on the emotional speech signal for decomposition of the signal into various sub-signals. Analysis of these signals shows that the VMD algorithm captures the center frequency close to the fundamental frequency defined for each glottal cycle of emotional speech utterance through its modes. This center frequency characteristic of the corresponding mode signal helps in the accurate estimation of epoch locations from the emotional speech signal. The performance evaluation of the proposed method is carried out on six different emotions taken from the German emotional speech database with simultaneous electroglottographic signals. Experimental results on clean emotive speech signals show that the proposed method provides identification rate and accuracy comparable to that of the best performing algorithm. Besides, the proposed method provides better

✉ E. A. Gopalakrishnan
ea_gopalakrishnan@cb.amrita.edu

G. Jyothish Lal
jyothishlal@gmail.com

D. Govind
d_govind@cb.amrita.edu

[1] Centre for Computational Engineering and Networking (CEN), Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

reliability in epoch estimation from emotive speech signals degraded by the presence of noise.

Keywords Epoch estimation · Glottal closure instants · Excitation source · Emotional speech signal · EGG signal · Variational mode decomposition

1 Introduction

Emotions in speech signals are reflected as the subtle variations in the excitation source parameters and vocal tract parameters [42]. Both these parameters contribute equally toward the characterization of various emotions. Nevertheless, there is a special emphasis on the analysis and recognition of emotion using the source parameters in the literature [7,19,23,28,29,31,38,42]. This is mainly due to the availability of reference electroglottographic signal [4] and well-established tools for the estimation of source parameters [35,53].

Instantaneous pitch [3,29,50], strength of excitation [23,28,38] and glottal flow parameters [5,6] are reported as the major emotion-dependent source parameters. Among these parameters, the instantaneous pitch is widely used for the analysis and synthesis of the emotional speech signal. For instance, Bulut et al. reported that the statistical measures derived only from instantaneous pitch show significant emotion class discrimination characteristics [3]. Bulut et al. also report that the instantaneous pitch is more important than the average pitch during emotional speech synthesis. Besides, the instantaneous pitch contour plays a significant role in the analysis stage of applications such as emotion recognition [7,23,29,38] and emotion conversion [5,6,16,20].

The instantaneous pitch contour for a given speech signal is derived as the inverse of the time interval between successive epoch locations (or glottal closure instants) [35]. This in turn demands the accurate estimation of epochs, which are the major source of excitation during the vibration of vocal folds [11,12]. Furthermore, analysis of other source features such as the strength of excitation and glottal flow parameters also requires the accurate estimation of epochs from the emotional speech signal [5,19,38,41]. Hence, the objective of the present work is the estimation of epoch locations from the speech for the emotion analysis.

The estimation of epochs from the speech signal is a challenging task due to the interaction of the vocal tract response [11]. There exist many efficient algorithms which can provide an accurate estimation of epochs from the speech signal by removing the vocal tract influence to a maximum extent. These methods are discussed briefly in the next subsection.

1.1 Existing Methods for Epoch Estimation

The methods proposed for the estimation of epochs from the speech signal employ different criteria for the identification or localization of epochs. The first type includes methods that rely on the residual signal extracted from the speech signal using linear prediction (LP) analysis for epoch estimation [11,39]. The LP residual signal shows

large values of error as discontinuities around epoch locations. However, the bipolar nature of peaks in the LP residual creates ambiguities in locating epochs [1]. Therefore, the Hilbert envelope (HE) of the LP residual was proposed by Ananthapadmanabha et al. [1] for unambiguous epoch estimation, exploiting its unipolar nature. However, the use of prediction error for epoch estimation is found to be less effective since the LP residual is often influenced by the vocal tract system [11]. This is because the inverse filter does not remove the vocal tract response completely.

The other criterion includes zero crossings of the phase slope function derived from the LP residual or the wavelet transform [36,46], properties of impulse-like excitation [35], singularity exponents [27] and the structure of the glottal flow derivative [30]. The dynamic programming phase slope algorithm (DYPSA) [36] uses the phase slope function of the LP residual for identification of the locations of epoch candidates, as negative zero crossings. Besides, the algorithm employs a phase slope projection technique to recover the undetected epoch locations. True epoch locations are then obtained by N-best dynamic programming. Later, 'yet another GCI/GOI algorithm' (YAGA) [46] was proposed by modifying DYPSA. In contrast to DYPSA, YAGA identifies the epoch location by applying the phase slope function on the wavelet transform of the source signal. The epoch identification rate is improved in YAGA by a GCI refinement process, which is not performed in DYPSA. The zero frequency filtering (ZFF)-based method proposed by Murty *et al.* [35] exploits the nature of impulse excitation during glottal closures. That is, the discontinuities due to impulse excitation are reflected across all frequencies including the zero frequency. Hence, the speech signal is passed through two cascaded zero frequency resonators. The resonator output is then trend removed to obtain the zero frequency filtered signal (ZFFS). The trend removal operation is performed by subtracting the mean over 1–2 times the average pitch period of the speech signal. The locations of positive zero crossings of the ZFFS are identified as the epoch locations. In speech event detection using the residual excitation and a mean-based signal (SEDREAMS) algorithm [12], the first step is to obtain a mean-based signal from the speech signal. Again, the window length is fixed based on the average pitch period for the computation of mean-based signal. Then, this mean-based signal is used to determine the intervals where an epoch is present. Finally, the peak in the LP residual is examined in that interval to identify the epoch. However, this in turn requires prior estimation of the polarity of the speech signal [17,25], to decide about the sign of peaks corresponding to epochs. The micro-canonical multi-scale formalism (MMF) [27] relies on the estimation of a multi-scale parameter called singularity exponents for detection of epoch locations. The MMF shows that epoch location corresponds to samples with lower singularity exponents. The glottal closure/opening instant estimation forward-backward algorithm (GEFBA) [30] estimates the epoch locations only in voiced regions of the speech signal. The GEFBA exploits the structure of the glottal flow derivative using simple time-domain criteria.

1.2 Drawbacks of the Existing Methods in the Context of Emotive Utterances

Most of the aforementioned methods identify more than one epoch candidates in one glottal cycle, which is followed by a candidate selection procedure. Therefore,

the choice of thresholds or window size fixed for localization of true epochs may directly affect the reliability of epoch estimation. For example, the fixation of window length based on the average pitch period creates an issue of missing epoch or spurious epoch in the zero frequency filtering approach. In summary, the performance of epoch estimation in the speech signal is mainly dependent on factors such as vocal tract resonances, size of analysis window length, algorithmic thresholds, polarity and uncontrollable variations in pitch. Nevertheless, the aforesaid factors are not reported to show any significant degradation in the performance of the epoch estimation in neutral speech [24]. However, all these factors contribute to the degradation in the performance of the epoch estimation in emotional speech signal [18,24]. Researchers have come up with studies concentrating on the robustness of epoch estimation techniques to additive noise and reverberation [27,30]. However, attempts focusing on epoch estimation from the emotional speech signals are limited.

1.3 Methods Proposed Exclusively for Emotive Speech Signals

In the literature, there exists only the modification of the ZFF method (m-ZFF) for the estimation of epochs from the emotional speech signal. Besides emotive speech, various other types of speech signals such as singing [26], laughter [13,32,45] and telephonic voices [8,21] are also analyzed based on modified ZFF method. For instance, Kumar *et al.* [32] used m-ZFF for the estimation of excitation source information (instantaneous pitch and epoch strength) for the analysis and characterization of laughter signals. Later, Thati et al. in [45] modified the estimated excitation source features for the synthesis of laughter signals. Also, Kadiri et al. in [26] has studied the effect of wider pitch range in singing voice using the m-ZFF method for the extraction of GCIs. Furthermore, Govind et al. proposed a m-ZFF approach [18] for epoch extraction from the emotional speech by re-filtering the ZFF signal using a low pass filter. Even though the method provides fair results for epoch estimation from emotional speech, it introduces many artifacts due to block processing [16]. Recently, Kadiri et al. proposed a method [24] based on the multi-scale product (MSP) of the single frequency filtered signal for deriving impulse-like events from the emotional speech signal. Then, prominent epochs are identified from derived impulses using the m-ZFF approach. Nevertheless, the performance evaluation results of these approaches show still scope for improvement.

1.4 Motivation and Formulation of the Proposed Method

The state-of-the-art methods approximate the resonance effect from the vocal tract system on the glottal excitation signal as a linear filter model. However, this kind of approximation is not appropriate for dealing with the highly nonlinear source filter interaction during the production of emotional speech signals. Consequently, it affects the performance of epoch estimation. Hence, it is more appropriate to analyze the speech signal using techniques meant for nonlinear signal processing. This has motivated us to explore the possibilities of a new adaptive time series decomposition technique called variational mode decomposition (VMD) for analyzing the non-stationary speech signals.

The discontinuities due to impulse excitation at epochs occur with a fundamental frequency defined for each glottal cycle [33]. These variations can be analyzed by decomposing the given emotional speech signal around the fundamental frequency defined for each glottal cycle. Among the three well-known adaptive signal decomposition techniques such as empirical wavelet transform (EWT) [15], empirical mode decomposition (EMD) [22] and variational mode decomposition (VMD) [10], VMD has been extensively used in areas of biomedical signal processing, speech signal processing and seismic signal processing [34,47,51]. The advantage of using VMD is that it captures the relevant center frequencies, ensuring good frequency separation [10]. Moreover, VMD is efficient for identifying various discontinuities present in a non-stationary signal [33,43]. In Lal et al. [33] and Deshpande et al. [9], the authors propose the estimation of GCIs from the electroglottographic signal using VMD. Furthermore, VMD algorithm has been applied on the neutral speech signal in an iterative manner for voice/unvoiced detection and estimation of the instantaneous pitch frequency [47,48]. Experimental results from Upadhyay et al. show that the iterative application of the VMD algorithm separates the fundamental frequency (F_0) component from the neutral speech signal. Upadhyay et al. do not use the epoch information for the estimation of the instantaneous fundamental frequency. However, there is no guarantee that the vocal tract system generates similar speech waveforms for each impulse-like excitation [53]. Also, we cannot assure any periodicity in the impulsive excitation at epochs. Hence, it is more advantageous to use an epoch-based approach for the estimation of instantaneous fundamental frequency.

In contrast to Upadhyay et al. [47] and Lal et al. [33], the proposed method tries to estimate epochs from emotive speech signals whose characteristics are completely different from neutral speech signals and EGG signals. Thus, the novelty of the proposed work is the effective utilization of the VMD algorithm in capturing the glottal source characteristics of emotive speech utterances for the estimation of epochs. Precisely, the proposed method tries to decompose the emotional speech signal to a sub-signal (mode) similar in structure to that of the excitation signal. The important characteristic of the desired mode is that its center frequency of oscillation should be close to the fundamental frequency (F_0) defined for each glottal cycle. Finally, we use this center frequency characteristic of the sub-signal for the estimation of epochs.

The rest of the paper is organized as follows. In Sect. 2, we describe the methodology for estimation of epochs from emotional speech signals. Section 3 discusses the database used, empirical experiments conducted for fixing the tuning parameter of VMD, performance evaluation of the proposed method and performance comparison results with other popular methods. Finally, Sect. 4 draws the conclusion and future directions.

2 Proposed Method for Epoch Estimation Using VMD

In the proposed method, we perform an iterative decomposition of the emotional speech signal using VMD. The desired VMD mode signal is then analyzed for identification of epoch location. Firstly, a brief description of the VMD algorithm is given below.

2.1 VMD Algorithm

VMD is a non-recursive and adaptive decomposition technique for any kind of non-stationary signal. It decomposes the non-stationary signal into a set of sub-signals or modes, with the number of components specified in prior [10]. Each of these decomposed modes has a compact support around a corresponding center frequency. VMD algorithm identifies these modes by minimizing the sum of the bandwidth of the modes. However, it enforces a constraint that the original signal should be obtained by summing up the decomposed modes. The procedure for identifying the mode is as follows. For each mode,

1. The one-sided frequency spectrum of the signal is obtained by using Hilbert transform.
2. The frequency spectrum is shifted to baseband region by multiplying an exponential tuned to the estimated center frequency.
3. The bandwidth is estimated through H1 Gaussian smoothness of the demodulated signal, i.e., the squared L2-norm of the gradient.

The mathematical representation of the procedure is given below.

$$\min_{x_k,\omega_k}\left\{\sum_k\left\|\frac{\partial}{\partial_t}\left[\left(\delta(t)+\frac{j}{\pi t}\right)*x_k(t)\right]e^{-j\omega_k t}\right\|_2^2\right\}\quad s.t\qquad\sum_{k=1}^{K}x_k(t)=x(t)\quad(1)$$

where $\frac{\partial}{\partial_t}[.]$ denotes the partial derivative of a function. Further, $x_k(t)$ corresponds to kth component of the signal $x(t)$ having center frequency (ω_k) and K represents the total number of modes. The analytical signal corresponding to $x_k(t)$ is obtained by convolution operation with $\left(\delta(t)+\frac{j}{\pi t}\right)$ [Hilbert transform]. Here, $j=\sqrt{-1}$ and $\delta(t)$ is the unit impulse function whose value is zero everywhere except at the origin (where it is infinity). The new signal formed has a unilateral spectrum which is shifted to the baseband by mixing with $e^{-j\omega_k t}$ tuned to mode's center frequency ω_k. Finally, the bandwidth of the mode is estimated based on the squared L2-norm of the gradient. Precisely, the formulation tries to find the K central frequencies and the corresponding modes $x_k(t)$.

Now, this optimization procedure is converted into an unconstrained one as follows.

$$\begin{aligned}\mathcal{L}(x_k,\omega_k,\lambda):=\alpha\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*x_k(t)\right]e^{-j\omega_k t}\right\|_2^2\\+\left\|x(t)-\sum_{k=1}^{K}x_k(t)\right\|_2^2+\left\langle\lambda(t),x(t)-\sum_k x_k(t)\right\rangle\end{aligned}\quad(2)$$

In Eq. 2, $\mathcal{L}$ represents the augmented Lagrangian , λ is the Lagrangian multiplier, and α is the bandwidth control parameter.

The above unconstrained problem is solved using alternate direction method of multipliers (ADMM) [10]. The ADMM solves one variable at a time assuming all

other variables are known. Firstly, the update for $x_k\ (t)$ is obtained by absorbing the last inner product term $\left\langle \lambda(t), x(t) - \sum_k x_k(t) \right\rangle$ into the second term $\left\| x(t) - \sum_{k=1}^{K} x_k(t) \right\|_2^2$. Therefore,

$$x_k^{n+1} = \arg\min_{x_k(t)} \alpha \sum_k \left\| \frac{\partial}{\partial_t} \left[\left(\delta(t) + \frac{j}{\pi t} \right) * x_k(t) \right] \mathrm{e}^{-j\omega_k t} \right\|_2^2 + \left\| x(t) - \sum_{k=1}^{K} x_k(t) + \frac{\lambda}{2} \right\|_2^2 \tag{3}$$

Equation 3 is solved in the spectral domain by noting the fact that norm in the time domain is same as that in the frequency domain. The solution for updated mode is obtained as follows.

$$\hat{X}_k^{n+1}(\omega) = \frac{\hat{x}(\omega) - \sum_{i \neq k} \hat{X}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \tag{4}$$

where $\hat{x}(w)$, $\hat{X}_i(w)$, $\hat{\lambda}(\omega)$ and $\hat{X}_k^{n+1}(\omega)$ represent the Fourier transforms of $x(t)$, $x_i(t)$, $\lambda(t)$ and ${x_k}^{n+1}(t)$, respectively. Similarly, the update for center frequency is obtained by solving the first term of Eq. 2 in the spectral domain. The updated central frequency is as follows.

$$\hat{\omega}_k^{n+1} = \frac{\int_0^\infty \omega \left| \hat{X}_k(\omega) \right|^2 \mathrm{d}\omega}{\int_0^\infty \left| \hat{X}_k(\omega) \right|^2 \mathrm{d}\omega} \tag{5}$$

The complete algorithm of VMD can be found in [10].

In order to illustrate the effectiveness of VMD in decomposing a multi-component signal, we have simulated a synthetic signal resembling the characteristics of a voiced segment of a speech signal. A voiced speech signal can be represented as an amplitude and frequency modulated (AM–FM) signal in the low frequency region (50–500 Hz) as follows [47].

$$F_{\mathrm{LFR}}(n) = \sum_{k=1}^{N} a_k(n) \cos\left(2\pi k f_0[n] n + \theta_k[n]\right) \tag{6}$$

where $f_0[n]$, $a_i(n)$, $\theta_k[n]$ and N represents the time varying frequency, time varying amplitude and phase of the kth harmonic of $f_0[n]$ and the number of harmonics, respectively. Here, we simulate a signal containing frequency components of 200 and 400 Hz. The time varying amplitude parameters are fixed as 1 and 0.5, respectively, and neglected the phase part. The sampling frequency used is 8 kHz. Further, we added white Gaussian noise (SNR of 10 dB) to the signal. The noisy simulated signal and its linear magnitude spectrum are shown in Fig. 1a, b. This input signal has

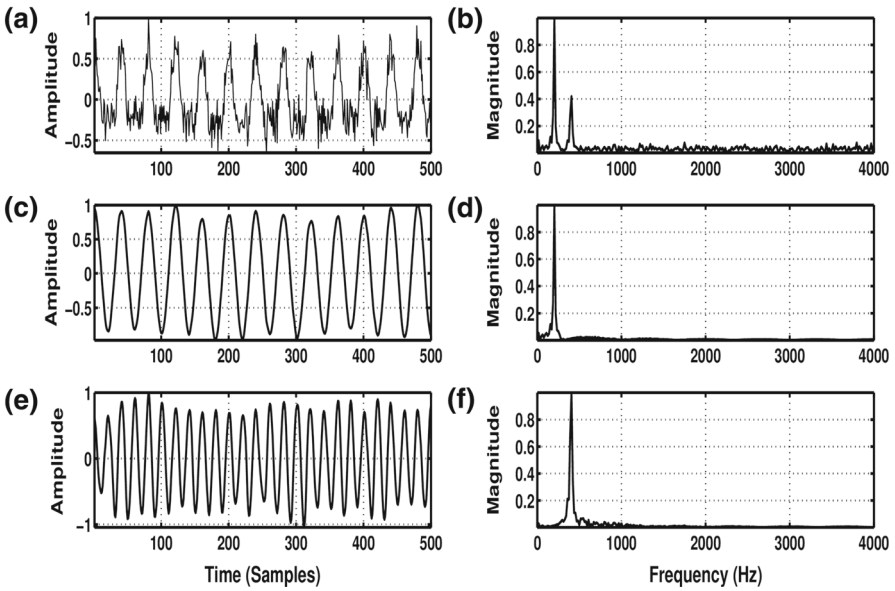

Fig. 1 Variational mode decomposition of a multi-component synthetic signal. Waveform and linear magnitude spectrum of **a**–**b** synthetic signal, corresponding **c**–**d** mode 1 component, **e**–**f** mode 2 component

been decomposed into two modes using the VMD algorithm. The modes and the corresponding linear magnitude spectrum obtained after the decomposition are shown in Fig. 1c, d and 1e, f. The center frequencies of the two modes obtained are 199.75 and 401.06 Hz, respectively. This confirms the effectiveness of VMD in separating the frequency components from the signal.

In VMD, the decomposition of a particular mode to a compact center frequency is largely dependent on two tuning parameters such as the number of modes K and bandwidth control parameter α. The control parameter K is fixed based on the number of sub-signals or components required, while α is fixed based on the center frequency of interest [52]. In theory, α is inversely proportional to the bandwidth of the components of the original signal. Further, the number of modes K controls the energy distribution among modes. A combination of very small α and a very few modes result in sharing of components among themselves. The sharing of mode components is termed as mode mixing [10,33]. Mode mixing occurs when the center frequency of the neighboring modes is very near to each other. Also, a combination of α with superfluous modes K would lead to redundant VMD information [52]. A smaller value of α makes the bandwidth of the filter wider. This tends to add more background noise to the results of VMD. Conversely, a narrow bandwidth makes distorted VMD results [52]. Furthermore, a combination of accurate α and K will include all the frequency components of the input in the results of VMD. Hence, proper selection of these two parameters is essential for assuring the accuracy of the results of VMD.

The emotion specific source features such as the location of glottal closures (epochs) and strength of excitation of a speech signal have its roots embedded in the glottal

waveform itself. Hence, the glottal waveform alone is sufficient for the estimation of epoch locations. However, this glottal excitation signal is filtered by the vocal tract system to produce the speech signal [11]. Moreover, there will be an increase in the fundamental frequency and energy of higher harmonics due to the rapid vibration of the vocal folds. Hence, it is required to separate the excitation characteristic from the influence of higher harmonics for the reliable estimation of epoch locations. The decomposition should be in such a way that one of the modes should preserve the excitation characteristics. The other mode corresponds to higher-frequency oscillations, which will be discarded. Therefore, we fix the number of modes K as two for the decomposition. Precisely, the center frequency of one of the modes should be near to the fundamental frequency (F_0) defined for each glottal cycle. However, VMD being a non-recursive algorithm, a single iteration might not be sufficient to bring the center frequency of a mode close to the fundamental frequency defined for each glottal cycle. Therefore, we apply VMD iteratively on the emotional speech signal until the center frequency of a mode is near to the fundamental frequency. The average F_0 of the emotional speech signal is obtained using the *fxrapt* algorithm [2,44].

The determination of α is challenging in the sense that the excitation characteristics of the various emotional speech signals are entirely different. In this work, we fix α for the first iteration of VMD based on the center frequency of interest. That is, we selected α such that the deviation of the center frequency from the average F_0 is the least. The results of the empirical evaluation are discussed in Sect. 3.1. Further, the value of α for successive iterations is fixed such that the gross error and mean absolute deviation are minimized in the estimation of the instantaneous pitch from emotional speech signals. Again, the results of pitch evaluation experiments are discussed in Sect. 3.1. Precisely, we used the empirically obtained optimal α combination (100,000, 10,000) [100,000 for the first iteration, 10,000 for successive iterations] for the estimation of epochs from the emotional speech signal.

2.2 Procedure for Epoch Estimation

The flow diagram of the proposed method is given in Fig. 2. The procedure is as follows.

1. Apply VMD on the emotional speech signal with K and α set to 2 and 100,000.
2. Select the mode with lesser center frequency and discard the other mode. Center frequency less than 80 Hz is also discarded since the human pitch ranges from 80 to 400 Hz.
3. If CF_{sm} is less than or equal to the average F_0 of the emotive speech signal, the VMD iteration is stopped. The selected mode signal is taken as the VMD output signal and proceeds to step 5.
4. If CF_{sm} is greater than the average F_0, apply VMD iteratively on the mode having the lesser center frequency ($K = 2$ and $\alpha = 10{,}000$). The iteration is stopped if CF_{sm} is less than or equal to the average F_0. Now, the selection of a particular mode or combination of modes as the VMD output signal is fixed based on CFD_{mm} between the two modes. If CFD_{mm} is greater than the threshold, choose the mode with lower center frequency as the VMD output signal. If CFD_{mm} is less than

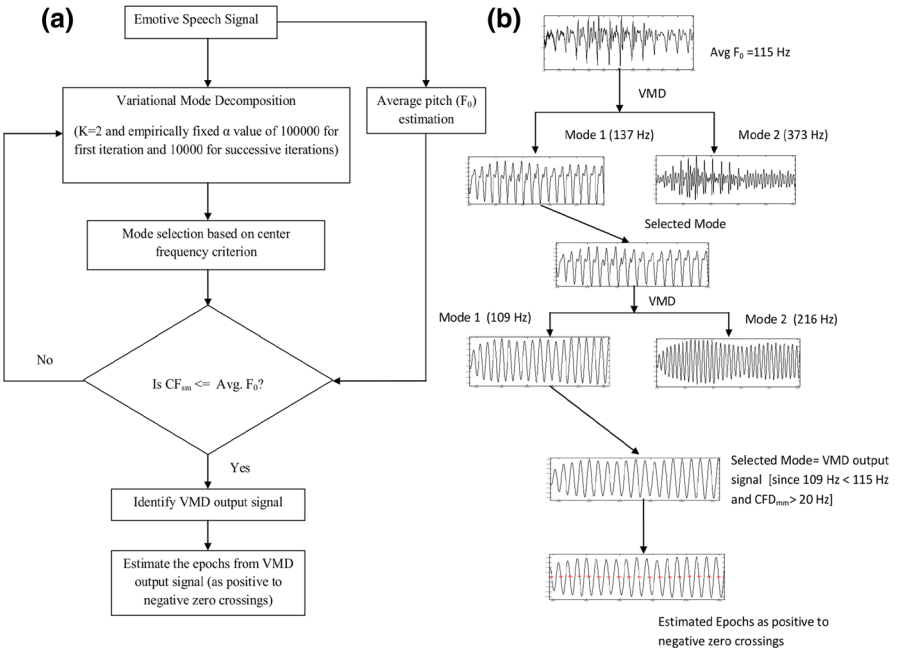

Fig. 2 **a** Flow diagram of the proposed method. CF_{sm} indicates the center frequency of the selected mode signal, CFD_{mm} denotes the absolute value of the difference between the center frequency two modes. Threshold is fixed as (1/4)th of the minimum pitch. **b** Waveform representation of the flow graph

or equal to the threshold, choose the combination of modes as the VMD output signal.

5. The positive to negative zero crossings of the VMD output signal are hypothesized as epoch locations.

The parameters CF_{sm}, CFD_{mm} and threshold are defined as follows.

- CF_{sm} indicates the center frequency of the selected mode signal.
- CFD_{mm} denotes the absolute value of the difference between the center frequency of the two modes.
- The threshold is fixed by computing CFD_{mm} between the two modes. We keep the threshold at (1/4)th of the minimum pitch, which is 20 Hz.

Figure 2b demonstrates the flow graph using waveforms obtained during each step. Here, we can observe that center frequency converges to the average F_0 in the second VMD iteration. The selected mode signal is identified as the VMD output signal, and its positive to negative zero crossings are hypothesized as epoch locations.

3 Experimental Results and Discussion

In this study, we perform the following different experiments with regard to epoch estimation from the emotional speech signal.

1. Experiments for determining the optimal value for α of VMD.
2. Experiments for evaluating the performance of the epoch estimation in emotional speech signals.
3. Experiments for the performance comparison with the state-of-the-art methods.

Firstly, we provide a brief description of the speech material and ground truth used for conducting the aforesaid experiments.

Database and ground truth The proposed method has been evaluated on the German emotional speech corpus (EMO-DB) having the simultaneous recording of electroglottogram (EGG) signals [4]. The database comprises of six basic emotions such as boredom, sad, disgust, fear, anger and happiness along with corresponding neutral versions [4]. It includes approximately 100 speech utterances (10 test sentences per emotion) spoken by 10 professional German actors (5 males and 5 females) with simultaneous EGG recordings. The recordings were initially sampled at 48 KHz and later down sampled to 16 KHz [4].

The ground truth for evaluating the performance of epoch estimation in the emotional speech signal is obtained manually from the corresponding DEGG signals. We used the Wavesurfer tool [49] for creating manual reference epochs. The labeling is done by observing the locations corresponding to significant negative peaks in the DEGG signal. Besides manual reference epochs, we collected algorithmic reference epochs based on the method proposed in Lal et al. [33]. In Lal et al., we show that epochs can be estimated more accurately and reliably from the EGG signal using the VMD algorithm. Thus, even if manual references are not available, one can use the complementary algorithmic references obtained using the method proposed in Lal et al. for evaluating the performance of epoch estimation.

Figure 3 shows an illustration of epoch estimation from the emotional EGG signal using the method proposed in Lal et al. [33]. Here, Fig. 3a, b depicts a voiced region of the EGG signal corresponding to anger speech and its first-order derivative (DEGG). Figure 3c–e shows the three modes obtained from VMD. From the decomposition results, it is observed that the positive to negative zero crossings [marked 'x' (blue)] of the second mode coincide with the locations of prominent positive peaks in the DEGG signal [marked as thick red lines in Fig. 3d]. This phenomenon occurs because the center frequency of the second mode coincides with the fundamental frequency of oscillation (F_0) in the EGG signal. Therefore, the positive to negative zero crossings of the second mode (Fig. 3d) correspond to epoch locations. This phenomenon cannot be seen in other modes because their center frequencies are far apart from F_0).

3.1 Determination of the Optimal α Value by Empirical Evaluation

The optimal α value pair for VMD iterations is determined empirically based on the experiments conducted on emotional speech signals taken from the EMO-DB. Initially, we search for the best α value (for the first iterations of VMD) which can minimize the deviation in center frequency from the average F_0. For successive iterations, we fix the α value such that the best performance is attained in the estimation of instanta-

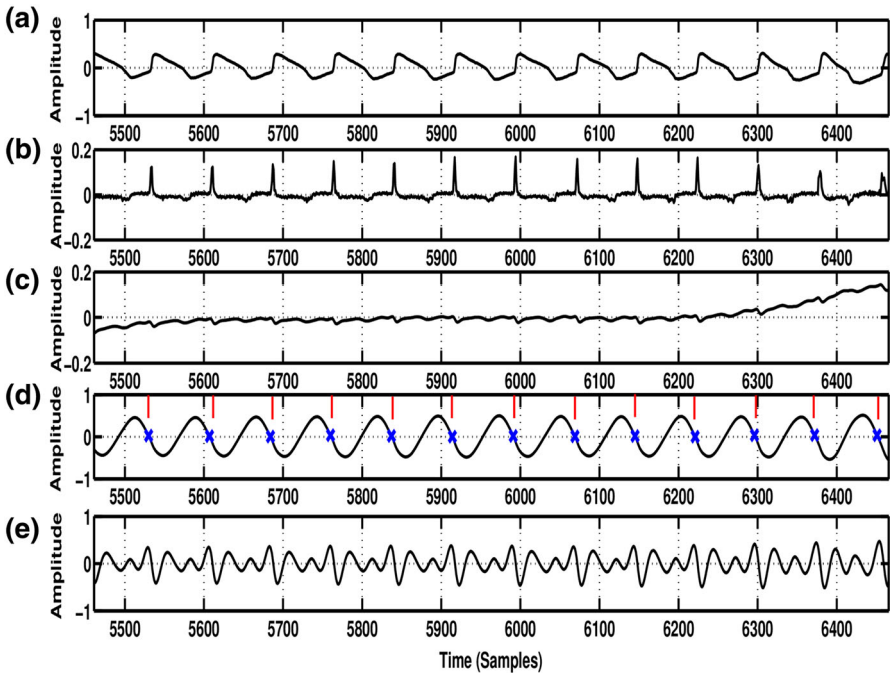

Fig. 3 Illustration of epoch estimation from the emotional EGG using VMD. **a** Voiced segment of anger EGG signal, corresponding **b** DEGG signal, **c** mode 1 component, **d** mode 2 component, **e** Mode 3 component. Epoch locations corresponding positive peaks in the DEGG signal are marked using thick red line (Color figure online)

neous pitch values. The standard performance measures for pitch evaluation are given below [40,53].

1. *Mean absolute error (MAE)* Mean of the absolute value of the difference between the estimated and reference pitch values.

$$\text{MAE}, \ \bar{e} = \frac{1}{N}\sum_{i=1}^{N}|e(m_i)| \quad \text{where} \ \ e(m_i) = P_i(r) - P_i(e) \tag{7}$$

   In Eq. 7, $P_i(r)$ and $P_i(e)$ represent the reference and estimated pitch value of the ith voiced frame and N is the total number of voiced frames.
2. *Standard deviation (SD)* Standard deviation of the difference between the estimated and reference pitch values.

$$\text{SD}, \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}e^2(m_i) - \bar{e}^2} \tag{8}$$

3. *Gross error (GE)* The percentage of voiced frame with an estimated pitch value that deviates from the reference pitch value by more than 20%.

$$\mathrm{GE} = \frac{D_p}{N} \times 100 \tag{9}$$

where D_p is the number of voiced frames with an epoch deviation greater than 20%.

In our experiments, we used the gross error and mean absolute error as measures for fixing the best α value for successive VMD iterations.

Experiments conducted We find that a lower α value includes high-frequency oscillations in the estimated modes from the emotional speech signal. This is true from a theoretical point of view since α is inversely proportional to the bandwidth of modes. Precisely, if we select a lower value of α (< 5000), it will make the bandwidth of the filter wider. Therefore, the estimated modes will contain high-frequency oscillation than the fundamental frequency of oscillation of the emotive utterance. This in turn affects the epoch estimation performance. Thus, a higher value of α is required to have a frequency band in the decomposed modes close to the fundamental frequency range (80–400 Hz) of an adult human being. Moreover, Yang et al. report that the value of α is fixed based on the center frequency of interest [52].

Here, we compare the influence of a lower α value and higher α value in capturing the center frequency close to the average fundamental frequency. The experiments are performed on the ten sentences of EMO-DB, spoken in the anger emotion by a female speaker. We varied the α value from a low value (5000) to a high value (100,000) for the first iteration of VMD on each utterance. Then, we measured the deviation in the center frequency of the selected mode from the average F_0 for each alpha value considered. Further, we computed the average deviation in center frequency (denoted as $\mathrm{CF}_{\mathrm{error}}$) of the ten utterances. The results of the empirical studies are given in Table 1. From the results, it is evident that the deviation error is on the higher side for a lower α value. The deviation error reduces after $\alpha = 50{,}000$ and attains the least value when $\alpha = 100{,}000$. Hence, one can choose an α value anywhere in the range between 50,000 and 100,000 for capturing the required center frequency from the emotive utterance. In this work, we have used an α value of 100,000 for the first iteration of VMD on the emotional speech signal. However, if we use a higher α value for every iteration of VMD, the correct center frequencies of modes will not be captured. Therefore, we conducted the pitch evaluation experiments for obtaining the optimal α combination for the iterative procedure. The experiments are conducted on a test sentence ('Das will sie am Mittwoch abgeben,' meaning 'She will hand it in on Wednesday') spoken in all emotions by 10 speakers. Reference pitch values are obtained by taking the inverse

Table 1 Deviation in center frequency from the average F_0 ($\mathrm{CF}_{\mathrm{error}}$) averaged over the ten emotive utterances

α Value	$\mathrm{CF}_{\mathrm{error}}$ (Hz)
5000	550.10
10,000	363.95
20,000	225.55
50,000	60.27
75,000	54.09
100,000	45.21

of the time interval between two successive ground truth epoch locations (manually labeled epochs). Then, pitch evaluation is performed based on measures such as gross error and mean absolute deviation.

During the first experiment, we kept the same α value of 100,000 for every iteration of VMD on the selected mode signal. Then, using the proposed method discussed in Sect. 2.2, epoch locations are identified. Further, pitch values are estimated by taking the inverse of the difference between two successive epoch locations. Mathematically, it is expressed as follows.

$$F_i(t) = \frac{1}{[e_l(\mathrm{t}+1) - e_l(\mathrm{t})]} \tag{10}$$

where $F_i(t)$ represents the instantaneous pitch or fundamental frequency and $e_l(t)$ represents the beginning of the pitch period.

Finally, we computed the gross error and mean absolute error between the estimated and the reference pitch values. During the second experiment, we kept α value as 100,000 only for the first iteration of VMD on the emotional speech signal. For successive iterations of VMD on the selected mode signal, α value is changed to 75,000. During the third experiment, the α value is changed to 50,000 for all successive VMD iterations on the selected mode signal after the first iteration. For the fourth, fifth and sixth experiment, α value pair used is (100,000, 25,000), (100,000, 10,000) and (100,000, 1000), respectively.

Furthermore, we have conducted the same pitch evaluation experiments for a lower value of α equal to 2000, without iteration. It is observed that the performance measures such as gross error and mean absolute error are very high when compared with the optimal α combination (100,000, 10,000).

Thus, the application of VMD in a non-iterative manner with a lower value of α is not found to be effective in improving epoch estimation performance. Further, we checked the feasibility of α equal to 2000 in an iterative manner. Again, the error measures are found to be on the higher side. Moreover, the number of VMD iteration required for a lower α value is always more than that of the proposed α value pair. For instance, the maximum number of iterations recorded for $\alpha = 2000$ in the pitch evaluation experiment is eight. However, for $\alpha = [100{,}000, 10{,}000]$, the number of iterations went up to a maximum of three only. Table 2 gives the gross error and mean absolute deviation obtained for the various experiments. From the results, it is evident

Table 2 Empirical results of various experiments conducted on emotional speech signals for fixing α

Alpha (α)value pair	GE (%)	MAE (Hz)
(100,000, 100,000)	12.74	21.74
(100,000, 75,000)	9.57	16.00
(100,000, 50,000)	8.38	15.12
(100,000, 25,000)	7.96	14.44
(100,000, 10,000)	**7.35**	**14.38**
(100,000, 1000)	8.23	18.03
(2000 for all iterations)	9.37	17.65
(2000 non-iterative)	37.58	100.15

Bold values indicate the least GE and MAE

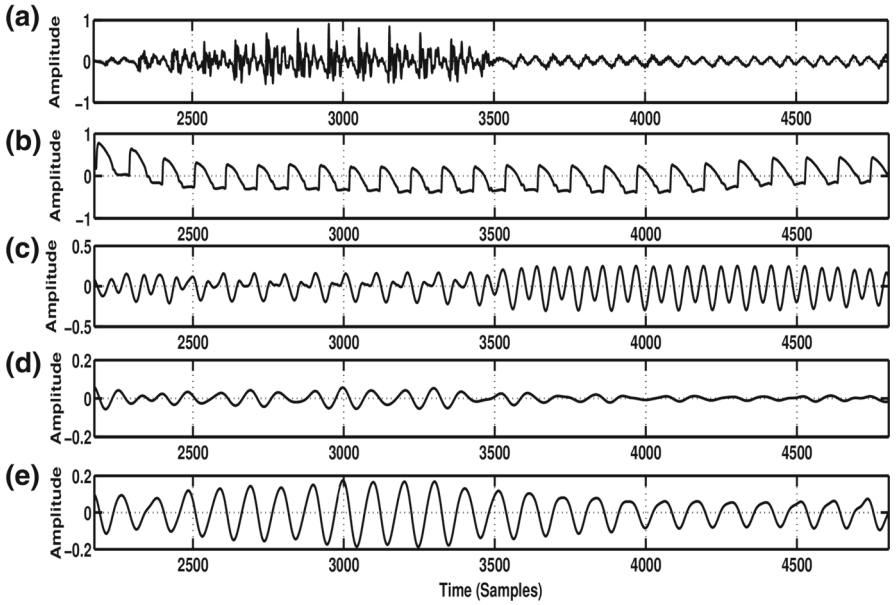

Fig. 4 Influence of α value on capturing correct frequency of oscillation. **a** Voiced segment of anger speech signal, corresponding **b** EGG signal, **c** selected mode after the first iteration of VMD using $\alpha = 100{,}000$, **d** VMD output signal using $\alpha = 100{,}000$ for successive iterations, **e** VMD output signal using $\alpha = 10{,}000$ for successive iterations

that when the α value pair is (100,000, 10,000), the gross error and mean absolute error are the least.

The influence of α value pair (100,000, 100,000) and (100,000, 10000) on capturing the fundamental frequency of oscillation in the emotional speech signal is demonstrated in Fig. 4. A voiced segment of the anger speech signal and the corresponding EGG signal are given in Fig. 4a, b. Figure 4c depicts the selected mode component (mode with lesser center frequency) after the first iteration of VMD with α value set to 100,000. By visual inspection of Fig. 4c, it is observed that the selected mode is not near to the fundamental frequency of oscillation in the glottal wave. Hence, the iteration continues on the selected mode. During the first experiment, we obtained the VMD output signal using an α value of 100,000 for successive VMD iterations. During the second experiment, we used a lower α value of 10,000 for successive VMD iterations. Figure 4d shows the VMD output signal obtained using an α value of 100,000 (from second iteration onwards). From Fig. 4d, it is evident that the fundamental frequency of oscillation is not captured in the output signal. In contrast, the VMD output signal obtained using an α value of 10,000 (Fig. 4e) clearly captures the fundamental frequency of oscillation in the glottal wave.

3.2 Performance Evaluation of Epoch Estimation in Emotional Speech Signals Using the Proposed Method

In order to illustrate the proposed method for epoch estimation in the emotional speech signal, an anger speech signal is taken. Then, the signal has been decomposed into

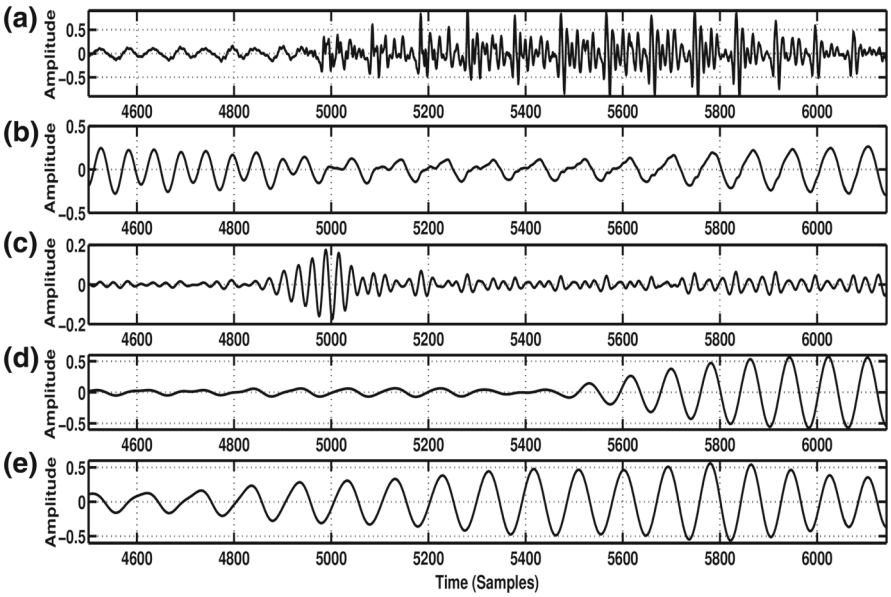

Fig. 5 Emotional speech signal decomposition using VMD. **a** Voiced segment of anger speech signal, corresponding **b–c** mode 1 and mode 2 after the first VMD iteration, **d–e** mode 1 and mode 2 after the final VMD iteration

two modes by keeping α as 100000. The results of the decomposition are given in Fig. 5, where (a) represents a voiced segment of anger speech, and (b)–(c) represents two modes obtained with center frequencies 230 and 570 Hz, respectively. The average fundamental frequency calculated using the *fxrapt* algorithm is approximately 193 Hz. Hence, the VMD iteration ($\alpha = 10{,}000$) continues on the mode with lesser center frequency (Fig. 5b). The modes obtained after the fourth iteration (with center frequencies 198 and 184 Hz) are shown in Fig. 5d, e, respectively.

Now, the VMD iteration has been halted since the center frequency of one of the modes has fallen below the average fundamental frequency . Finally, based on step 4 of the procedure for epoch estimation described in Sect. 2.2, the combination of modes is taken as the VMD output signal. Figure 6 plots the linear magnitude spectrum corresponding to modes selected after the first iteration and that of the VMD output signal. By visual inspection of all subplots, the spectrum of the selected mode after the first iteration (Fig. 6f) shows spectral peaks beyond the fundamental frequency. In contrast, the spectrum of the VMD output signal shows only the spectral peaks corresponding to the fundamental frequency.

Precisely, the spectral peak in Fig. 6h resembles with the spectral peaks corresponding to the fundamental frequency in the glottal waveform (Fig. 6d). Analysis of the VMD output signal shows rapid changes around the positive to negative zero crossings. From Fig. 7, it is evident that the time instants corresponding to these rapid changes represent the epoch locations. Here, Fig. 7a depicts the same segment of anger speech signal used in Fig. 5a. Figure 7b,c plots the corresponding EGG and DEGG waveforms, respectively. The reference epoch locations labeled using Wavesurfer are

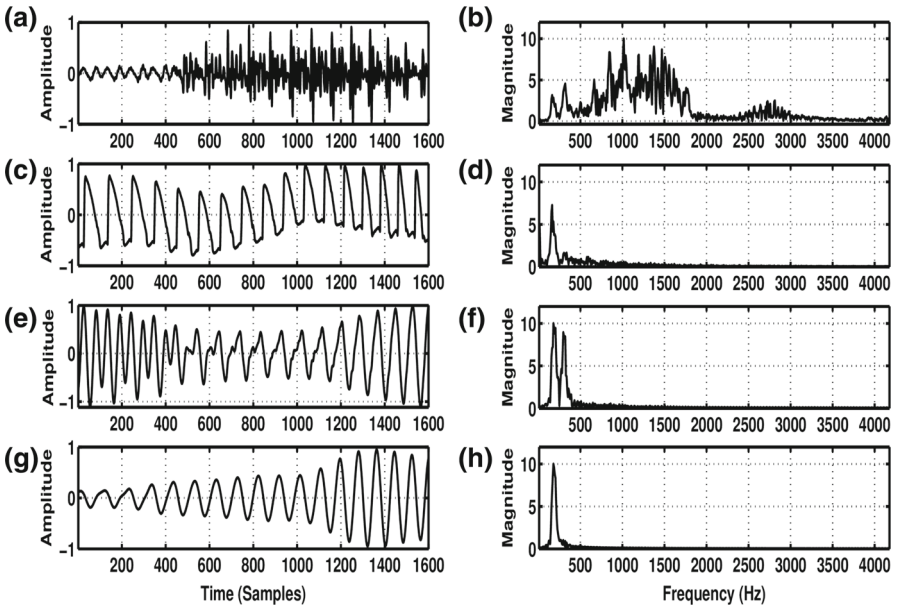

Fig. 6 Linear magnitude spectrum corresponding to modes selected after the first and final iteration of VMD on anger speech segment. The waveform and corresponding linear magnitude spectrum of **a**–**b** voiced segment of anger speech signal, **c**–**d** EGG segment corresponding to anger speech, **e**–**f** mode selected after the first VMD iteration, **g**–**h** VMD output signal

indicated in the DEGG signal as '$\times$'(magenta). Besides, the reference epoch locations estimated from the EGG signal using VMD are marked as '+'(blue) in the corresponding selected mode component (mode 2) [Fig. 7d]. It is observed that the positive to negative zero crossings (marked '*o*' (red)) of the VMD output signal (Fig. 7e) closely coincide with reference epoch locations shown in Fig. 7c,d. Hence, the time instants corresponding to positive to negative zero crossings are identified as epoch locations.

VMD is suitable for the extraction of noise robust component since it follows the Wiener filter structure [47]. However, we found that the signal-to-noise ratio should be a minimum of 5 dB for reliable estimation of epochs as positive to negative zero crossings. This is validated by measuring the reliability of the proposed method for emotive speech with additive noises at SNR levels from 0 to 30 dB. Firstly, we briefly describe the measures for testing the reliability and accuracy of the proposed method.

Performance measures Performance evaluation is performed on the voiced regions of the speech signal by defining the larynx cycle as in [36].

If the rth reference epoch occurs at e_r, then larynx cycle is defined as the range of samples $(1/2)(e_{r-1} + e_r) \leq n \leq (1/2)(e_r + e_{r+1})$. Based on the larynx cycle, two sets of measures are defined for evaluating the reliability and accuracy of the proposed method. The first set includes the following.

1. *Identification rate (IDR)* The percentage of larynx cycle for which exactly one epoch is detected.
2. *Miss rate (MR)* The percentage of larynx cycle for which no epoch is detected.

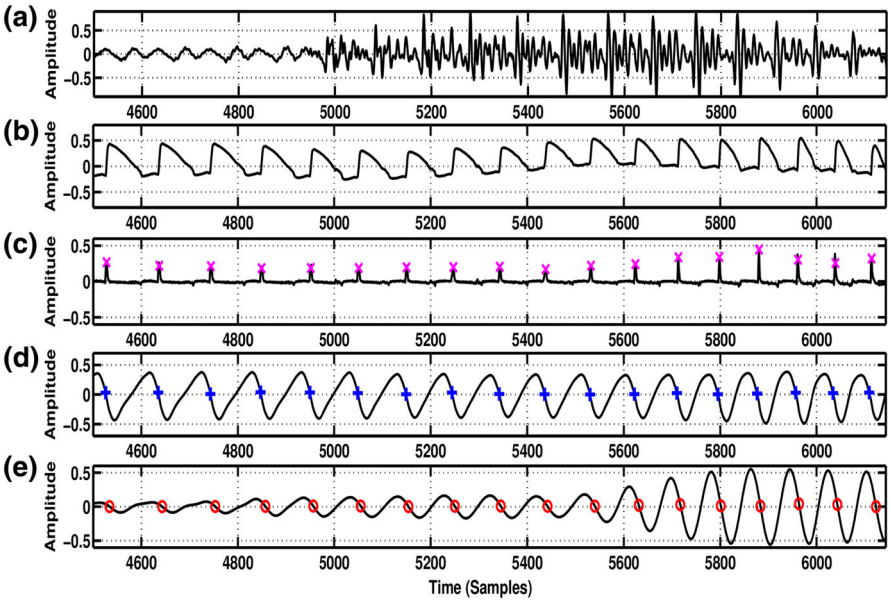

Fig. 7 Illustration of epoch estimation in emotional speech signal using proposed method. **a** Voiced segment of anger speech signal, corresponding **b** EGG segment, **c** DEGG signal with manually labeled reference epochs indicated using '×' (magenta), **d** VMD output signal from EGG with reference epochs indicated using '+' (blue), **e** VMD output signal. Estimated epochs are marked using '*o*'(red) (Color figure online)

3. *False alarm rate (FAR)* The percentage of larynx cycle for which more than one epoch is detected.

The IDR, MR and FAR quantify the reliability of epoch estimation. The second set includes the following.

1. *Identification error* ζ The timing error between the reference epoch and the estimated epoch in larynx cycle for which one epoch was identified.
2. *Identification accuracy* (IDA in '*ms*') The standard deviation of identification error ζ.
3. *Accuracy to* ± 0.25 *ms* (IDA to ± 0.25 ms in '%') The percentage of larynx cycles for which exactly one epoch is identified and ζ is within ± 0.25 ms.

IDA in '*ms*' and '%' quantifies the accuracy in the estimation of epochs. For IDA in '*ms*', lower value indicates higher accuracy. Further, IDA in '%' is measured as follows.

$$\text{IDA in '\%'} = \frac{\text{Number of epochs having '}\zeta\text{' within } \pm\, 0.25\,\text{ms}}{\text{Total number of correctly identified epochs}} \times 100 \qquad (11)$$

3.2.1 Assumption on Signal-to-Noise Ratio

We identified the positive to negative zero crossings of the VMD output signal as epochs. However, VMD method is sensitive to noise. Hence, we conducted the fol-

Table 3 Empirical results of the test for robustness to additive noise

SNR (dB)	IDR (%)	CF_{error} (Hz)
30	93.77	5.96
25	93.77	6.01
20	93.34	6.71
15	93.33	8.38
10	92.58	9.80
5	92.00	10.84
0	86.77	26.64

CF_{error} indicates the average deviation in center frequency from the average F_0 of each utterance

lowing experiments for checking the conditions under which such an assumption can be made.

The experiments are conducted on a test sentence ('Das will sie am Mittwoch abgeben,' meaning 'She will hand it in on Wednesday') spoken in all emotions by a female speaker. In the first experiment, we calculated the average IDR value and average deviation in center frequency (from average F_0 of each utterance) for the clean emotive signals. The results obtained are 93.91% and 5.90 Hz, respectively. In the subsequent experiments, we added white Gaussian noise at different SNR level to the emotive utterances. Again, we calculated the measures (average IDR and average deviation in center frequency) for each SNR value considered. The results obtained are given in Table 3. From the table, it is clear that the identification rate reduces by more than 2% for SNR values below 5 dB. Also, the deviation in center frequency from the average F_0 (denoted as CF_{error}) is found to be more for SNR levels below 5 dB. Therefore, we conclude that the identification of epoch locations as positive to negative zero crossings is robust to white Gaussian noise for SNR level as low as 5 dB.

3.2.2 Performance Evaluation Using Manual Reference and VMD-Based Reference

Now, the performance evaluation of the estimation of epochs in the emotional speech signal is evaluated across six basic emotions (boredom, disgust, fear, anger, sorrow and happiness) taken from German emotional speech corpus (EMO-DB). The results of the performance evaluation are given in Table 4. We provide results of the evaluation based on both the manual reference and VMD-based reference. From the results, we can observe that the IDR values are lesser in highly aroused emotions such as anger and happiness. This is due to the reduced strength of excitation in the speech signal corresponding to these emotions. This in turn might have reduced the energy associated with modes, leading to spurious epochs (as indicated by higher FAR values). However, the IDR values of emotions with lesser loudness levels (boredom, sad and disgust) are on the higher side. Furthermore, the IDR values obtained based on manual references and VMD-based references are found to show small deviation for the same emotion.

For example, the IDR of boredom for VMD reference is lower than manually labeled and the IDR of happy for VMD reference is higher than manually labeled. This inconsistent variation is due to the following two reasons.

Table 4 Performance evaluation of the proposed method for epoch estimation in emotional speech signals

Emotions	Using manually labeled reference					Using VMD-based reference				
	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDA to ±0.25 ms (%)	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDA to ±0.25 ms (%)
Boredom	95.02	0.82	4.16	0.52	66.36	93.81	0.77	5.42	0.57	68.25
Sad	95.34	1.65	3.01	0.46	68.24	96.28	0.57	3.15	0.48	66.33
Disgust	95.13	1.20	3.67	0.53	64.96	94.89	0.94	4.17	0.50	63.41
Fear	95.07	1.25	3.68	0.59	57.12	93.81	1.77	4.42	0.68	55.22
Anger	90.10	3.45	6.45	0.70	54.17	90.43	3.69	5.88	0.67	54.08
Happy	90.12	2.13	7.75	0.71	57.65	91.18	2.85	5.97	0.71	55.85
Average	**93.46**	1.75	4.79	0.59	**61.42**	**93.40**	1.76	4.83	0.60	**60.52**

The average performance in terms of IDR and IDA is given in bold

- The epoch in the voicing offset regions of the EGG signal are clearly identified using VMD [22]. The manual method fails to identify any epoch at the end of the voiced segment where the DEGG signal is almost zero.
- Again, manual reference epoch creation is prone to human error.

Therefore, the number of reference epochs considered for performance evaluation in both cases differs. This in turn results in a small difference (around 1–1.5%) in the performance in terms of IDR.

Among the accuracy measures such as IDA in '*ms*' and IDA in '%', the latter seems to the best measure for discussing the accuracy of the proposed method. This is because it identifies the percentage of epochs estimated within ± 0.25 ms of the reference epochs. From the results, it appears that the accuracy of epoch estimation decreases with the increase in arousal levels. The accuracy of boredom, sad and disgust is on the higher side compared to anger, fear and happy. A similar trend can be observed from the IDA measure in '*ms*.' That is, IDA (*ms*) values are lower [indicating high accuracy] for low aroused emotions and vice versa. Further, the difference in accuracy of manual and VMD-based reference is due to the difference in IDR obtained for the same emotion category. That is, for lower IDR, the chance of an increase in accuracy is higher.

The general observation of the outcome is that the average reliability and the accuracy of the proposed method are almost equal for both types of reference used. The results confirm the effectiveness of using VMD-based reference epochs from the EGG signal for the performance evaluation of the proposed method.

3.3 Performance Comparison of the Proposed Method with Existing Methods

The performance of the proposed method for epoch estimation in emotional speech signals is compared with popular methods such as ZFF, SEDREAMS, DYPSA, MMF, GEFBA and modified ZFF. All these methods are evaluated in the voiced regions of the emotional signal based on manually labeled reference epochs and epochs estimated from EGG signals using the VMD algorithm [33]. The comparative results obtained for manual and algorithmic references are shown in Tables 5 and 6, respectively.

The IDR of all methods is found to be decreasing as the level of arousal increases in emotions. That is, the reliability is less in anger and happy when compared to boredom and sad. Methods such as SEDREAMS and GEFBA are found to be performing well on boredom and sad emotions. However, these methods show a reduced IDR performance in other emotions. This is due to the rapid changes in the glottal excitation characteristics of high aroused emotions. Among the methods compared, DYPSA and MMF are found to show the least IDR performance in all emotions. The standard ZFF method gives a higher performance than SEDREAMS, GEFBA, DYPSA and MMF. However, its performance is lower than that of the modified ZFF approach. This is because of the fact that the ZFF method uses a single window length based on the average pitch period for getting the trend removed signal. The chances of spurious or missed epoch estimation are higher in the zero frequency filtered signal when using a fixed window length. In contrast, the proposed method uses the average fundamental frequency only to check whether the iteration has brought modes of decomposition

Table 5 Performance comparison results of epoch estimation in emotional speech signals using manually labeled reference

Emotion (# Files)	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDA to ±0.25 ms (%)
Boredom (112)	SEDREAMS	95.36	2.12	2.51	0.59	56.87
	DYPSA	93.59	2.67	3.74	0.78	49.58
	MMF	89.09	5.57	5.33	0.84	50.12
	GEFBA	95.72	2.73	1.55	0.55	54.63
	ZFF	98.62	0.75	0.63	0.34	**67.40**
	m-ZFF	**98.68**	0.64	0.68	0.62	64.51
	VMD	95.02	0.82	4.16	0.52	66.36
Disgust (105)	SEDREAMS	91.14	4.02	4.84	0.82	50.63
	DYPSA	89.58	5.21	5.21	0.92	54.16
	MMF	84.53	12.06	3.42	1.03	46.70
	GEFBA	90.39	6.28	3.33	1.09	49.64
	ZFF	95.31	2.09	2.61	0.61	55.80
	m-ZFF	**97.25**	1.40	1.36	0.51	**66.28**
	VMD	95.13	1.20	3.67	0.53	64.96
Fear (124)	SEDREAMS	86.44	5.14	8.42	1.01	44.75
	DYPSA	83.38	8.93	7.69	1.13	40.25
	MMF	68.34	28.00	3.66	1.20	35.93
	GEFBA	87.41	7.33	5.26	1.20	39.98
	ZFF	91.57	3.54	4.89	0.87	51.58
	m-ZFF	94.62	2.72	2.66	0.40	**58.93**
	VMD	**95.07**	1.25	3.68	0.59	57.12
Anger (136)	SEDREAMS	84.26	5.19	10.55	1.22	44.51
	DYPSA	82.44	10.94	6.62	1.23	45.11
	MMF	58.96	37.51	3.53	1.33	37.89
	GEFBA	87.52	8.34	4.14	1.37	47.20
	ZFF	84.20	4.99	10.80	1.16	49.34
	m-ZFF	**90.82**	6.08	3.10	0.45	**57.43**
	VMD	90.10	3.45	6.45	0.70	54.17
Happy (115)	SEDREAMS	86.67	4.76	8.57	1.21	47.66
	DYPSA	83.90	9.66	6.44	1.21	45.83
	MMF	59.98	36.54	3.47	1.28	39.44
	GEFBA	86.98	9.11	3.91	1.32	48.25
	ZFF	88.34	4.64	7.02	1.10	50.87
	m-ZFF	**90.80**	6.13	3.07	0.48	**57.76**
	VMD	90.12	2.13	7.75	0.71	57.65
Sad (120)	SEDREAMS	93.51	4.31	2.17	0.72	59.39
	DYPSA	90.34	3.70	5.96	1.01	53.15
	MMF	86.63	5.60	7.77	1.07	43.40
	GEFBA	90.04	9.17	0.79	0.36	**68.64**

Table 5 continued

Emotion (# Files)	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDA to ±0.25 ms (%)
	ZFF	93.73	2.92	3.35	0.55	68.18
	m-ZFF	**95.57**	1.92	2.51	0.56	66.10
	VMD	95.34	1.65	3.01	0.46	68.24
Average	SEDREAMS	89.57	4.26	6.18	0.93	50.64
	DYPSA	87.21	6.85	5.94	1.05	48.01
	MMF	74.59	20.88	4.53	1.12	42.25
	GEFBA	89.68	7.16	3.16	0.98	51.39
	ZFF	91.96	3.16	4.88	0.77	57.20
	m-ZFF	**94.62**	3.15	2.23	0.50	**61.84**
	VMD	**93.46**	1.75	4.79	0.59	**61.42**

Best results in IDR and IDA for each emotion category are given in bold

close to the fundamental frequency of oscillation defined for each glottal cycle. The center frequency of decomposed mode is controlled only by the tuning parameters of VMD. Proper selection of tuning parameters helps the VMD output signal to oscillate at the fundamental frequency defined for each glottal cycle in the emotional speech signal. This in turn improves the performance of the proposed method in terms of reliability. Precisely, the IDR of the proposed method is found to be higher than that of the five standard methods (SEDREAMS, GEFBA, DYPSA, MMF and ZFF) in highly aroused emotions.

Further, it is found that the m-ZFF method gives better reliability in epoch identification across various emotions. This improved performance in m-ZFF is due to the local pitch period oscillations in the ZFF signal. The proposed method is found to give a close match in IDR performance to that of the m-ZFF method, especially in emotions such as anger, happy, fear and sad. This is because of the fact that the VMD output signal also oscillates close to the fundamental frequency. The average reliability of the proposed method is found to be comparable with that of the m-ZFF approach.

Now, the identification accuracy (in terms of IDA in '*ms*' and IDA in '*%*') also shows a similar trend with respect to arousal level. That is, the accuracy of all methods is on the higher side for boredom and sad when compared to anger and happy. Among the seven methods, the standard ZFF method shows better identification accuracy in boredom and sad.

The accuracy of the proposed method is found to be slightly less than the standard ZFF method in boredom and sad emotion. However, the proposed method outperforms the standard ZFF method in other emotions. Further, SEDREAMS, DYPSA and MMF show reduced epoch identification accuracy in all the emotion categories. The GEFBA has shown slightly better identification accuracy in sad emotion (for manual reference). However, this increase in accuracy is due to the decrease in IDR value. In contrast, the proposed method has shown almost equivalent IDA performance with better identification rate in sad emotion.

Table 6 Performance comparison results of epoch estimation in emotional speech signals using VMD-based reference

Emotion (# Files)	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDA to ±0.25 ms (%)
Boredom (112)	SEDREAMS	93.16	2.72	4.11	0.83	54.17
	DYPSA	91.03	3.13	5.84	0.97	49.16
	MMF	81.47	12.47	6.06	1.05	52.20
	GEFBA	93.55	2.71	3.75	0.84	52.04
	ZFF	96.60	0.95	2.46	0.50	**68.43**
	m-ZFF	**97.13**	1.36	1.51	0.61	65.83
	VMD	93.81	0.77	5.42	0.57	68.24
Disgust (105)	SEDREAMS	89.76	4.00	6.23	0.94	53.57
	DYPSA	89.16	4.94	5.90	1.04	52.57
	MMF	83.77	11.82	4.40	1.11	47.76
	GEFBA	90.01	5.74	4.25	1.16	49.71
	ZFF	94.92	1.88	3.20	0.66	55.38
	m-ZFF	**96.67**	1.57	1.76	0.50	63.11
	VMD	94.89	0.94	4.17	0.50	**63.41**
Fear (124)	SEDREAMS	86.77	5.12	8.11	1.11	42.14
	DYPSA	83.16	9.23	7.61	1.22	41.04
	MMF	67.00	28.84	4.17	1.30	38.19
	GEFBA	86.98	7.74	5.28	1.25	39.71
	ZFF	90.19	4.24	5.57	0.98	50.66
	m-ZFF	93.33	4.01	2.65	0.40	**57.40**
	VMD	**93.81**	1.77	4.42	0.68	55.22
Anger (136)	SEDREAMS	83.96	4.95	11.09	1.26	44.80
	DYPSA	81.95	10.78	7.27	1.27	43.67
	MMF	58.88	37.09	4.03	1.42	38.13
	GEFBA	87.57	7.76	4.66	1.36	46.06
	ZFF	84.17	4.72	11.11	1.20	48.51
	m-ZFF	**90.95**	5.78	3.27	0.45	**58.59**
	VMD	90.43	3.69	5.88	0.67	54.08
Happy (115)	SEDREAMS	86.69	4.69	8.62	1.24	46.00
	DYPSA	83.51	9.69	6.80	1.24	45.58
	MMF	59.35	36.71	3.94	1.35	38.56
	GEFBA	86.99	8.78	4.23	1.34	46.20
	ZFF	88.78	4.43	6.78	1.15	50.17
	m-ZFF	90.08	6.48	3.44	0.48	**57.95**
	VMD	**91.18**	2.85	5.97	0.71	55.85
Sad (120)	SEDREAMS	93.12	3.46	3.41	0.79	58.13
	DYPSA	90.80	2.69	6.51	1.10	53.03
	MMF	86.70	4.83	8.47	1.11	44.31

Table 6 continued

Emotion (# Files)	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDA to ±0.25 ms (%)
	GEFBA	91.85	6.74	1.41	0.46	65.13
	ZFF	95.91	1.42	2.67	0.42	**69.87**
	m-ZFF	**96.94**	0.34	2.71	0.54	65.66
	VMD	96.28	0.57	3.15	0.48	66.33
Average	SEDREAMS	88.91	4.16	6.93	1.03	49.80
	DYPSA	86.60	6.74	6.65	1.14	47.51
	MMF	72.86	21.96	5.18	1.22	43.19
	GEFBA	89.49	6.58	3.93	1.07	49.81
	ZFF	91.76	2.94	5.30	0.82	57.17
	m-ZFF	**94.18**	3.26	2.56	0.50	**61.42**
	VMD	**93.40**	1.76	4.83	0.60	**60.52**

Best results in IDR and IDA for each emotion category are given in bold

The m-ZFF approach is found to give better identification accuracy across highly aroused emotions (anger, fear, happy and disgust) when compared to other methods. The better result in m-ZFF is attributed to its nature of extracting impulsive excitations directly from the emotive utterance. Furthermore, the accuracy of the proposed method is found to be higher than the m-ZFF method by 1–3% in low aroused emotions (boredom and sad). However, the m-ZFF outperforms the proposed method in highly aroused emotions. The deviation in accuracy is around 1–3%.

In summary, we can conclude that the proposed method is superior to five other methods except the m-ZFF method (in terms of identification accuracy) in highly aroused emotions . The average identification accuracy of the proposed method is found to be comparable with that of the m-ZFF approach.

Figure 8 depicts the histogram of epoch timing error averaged over the emotional database for the proposed method and the m-ZFF approach. The peaks in the distribution are mostly concentrated near the origin. The proposed method has a similar histogram as that of the m-ZFF approach.

Even though the m-ZFF approach provides a slightly better epoch estimation performance than the proposed method, it suffers from the following disadvantages [16].

1. m-ZFF uses block processing, which introduces unwanted spectral leakage during the post-processing of the ZFF signal.
2. The local pitch (F_0) value obtained for each frame is crucial in the estimation of epochs. A small change in the estimated F_0 will degrade the performance.

The proposed method holds an advantage over m-ZFF in the sense that it does not use any kind of block processing. The proposed approach processes the entire emotional speech signal at once to estimate the epoch locations. Hence, any artifacts due to block processing and windowing are avoided. Further, the miss rate in the proposed method is found to be lesser than that of the m-ZFF approach.

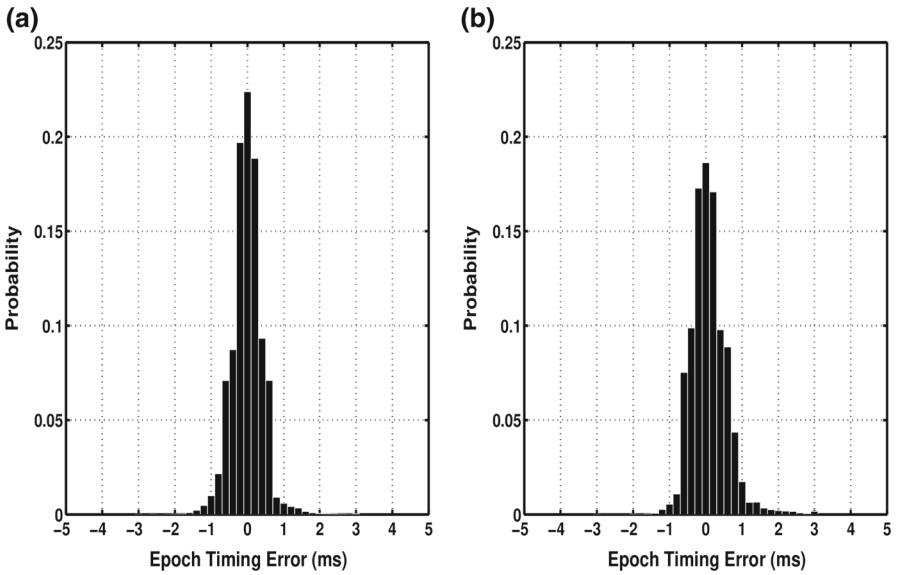

Fig. 8 Histogram of the epoch timing error averaged over six different emotions. **a** Proposed method (IDA to ± 0.25 ms is 61.42%), **b** m-ZFF method (IDA to ± 0.25 ms is 61.84%)

3.3.1 Performance Comparison of the Proposed Method with m-ZFF Method for Degraded Emotive Speech Signals

We evaluated the robustness of the proposed method and the m-ZFF approach for different noise degradations. The evaluation is done by calculating the IDR measure for the total database in the presence of three additive noises (white, babble and pink) taken from NOISEX database [37] at SNR levels of 0 dB and 10 dB. The average results obtained for different noise degradations are given in Table 7. From the results, it is evident that the proposed method gives a significantly higher identification rate than that of the m-ZFF for each of the noise degradations considered. The improved performance of the proposed method is attributed to the selection of noise robust VMD output signal for epoch estimation. VMD embeds Wiener filtering to update the modes directly in the frequency domain [10]. This enables the extraction of noise robust modes [14]. In contrast, the m-ZFF method refines only the conventional ZFF signal by block processing and re-filtering. Therefore, the major issues associated with ZFF method (such as speech contaminated with interference from other speakers, spurious impulse-like sequences and so on) [35] prevails in the m-ZFF method also. This results in a reduced epoch estimation performance of the m-ZFF method for degraded emotive speech signals.

4 Conclusion

This paper proposes a novel method for the estimation of epoch locations from the emotive utterance. The proposed approach benefits from the effectiveness of the VMD

Table 7 Performance comparison in terms of IDR for the m-ZFF and the proposed method over the total database in noise-degraded conditions at SNR levels of 0 and 10 dB

Best results in IDR are highlighted in bold

Noise	SNR (dB)	Proposed IDR (%)	m-ZFF IDR (%)
White	0	**88.27**	77.39
	10	**90.61**	84.92
Babble	0	**82.12**	71.42
	10	**88.03**	78.19
Pink	0	**86.71**	68.98
	10	**87.86**	76.81

algorithm in decomposing the emotional speech signal into modes with correct center frequencies. Finally, the decomposed modes are analyzed for the estimation of epochs.

The major contributions of the proposed work are:

- Effective utilization of the VMD algorithm in capturing the glottal source characteristics of the emotive speech utterances.
- Reliable estimation of epoch locations from clean and noise-degraded emotional speech signal using center frequency criterion of VMD.

We show that the application of the VMD algorithm iteratively on the emotional speech signal helps to capture the required center frequency of the mode. The center frequency of the selected mode is found to be near to the fundamental frequency of the glottal excitation signal. This is significant in the sense that the epochs occur with a fundamental frequency defined for each glottal cycle. The center frequency characteristic of the corresponding mode is utilized for reliable and accurate estimation of epoch locations. Epoch locations are hypothesized as positive to negative zero crossings of the VMD output signal.

We evaluated the performance of the proposed method in terms of identification rate and identification accuracy on emotional speech signals taken from the German emotional database. Further, we compared the effectiveness of the proposed method with the state-of-the-art epoch estimation methods. Performance comparison results show that the proposed method is almost as reliable and accurate as the m-ZFF approach for clean speech signals. Besides, we show that the proposed method outperforms m-ZFF in the presence of additive noise degradations. Therefore, the proposed method can be used as a better approach toward epoch estimation in emotive speech degraded with additive noise. Moreover, the proposed method can be used as a tool for accurate emotion analysis by deriving instantaneous pitch contours from epochs estimated.

Furthermore, like other methods, the reliability of the proposed method is found to be lower in highly aroused emotions such as anger and happiness. Future work will address this limitation, and we are formulating suitable modifications to proposed work to resolve the issue. The reduced strength of excitation in these emotions might have reduced the energy associated with decomposed modes. This in turn resulted in spurious epoch estimation. We need to explore more to address this issue.

Acknowledgements The authors gratefully acknowledge Amrita Vishwa Vidyapeetham for supporting the first author in pursing his Ph.D. The authors would like to thank Dr. K.P. Soman and Ms. M. Neethu (Amrita Vishwa Vidyapeetham) for lucidly explaining the concept of VMD .

References

1. T. Ananthapadmanabha, B. Yegnanarayana, Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Signal Process. **27**(4), 309–319 (1979)
2. M. Brookes, VOICEBOX: speech processing toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. Accessed 30 May 2017
3. M. Bulut, S. Narayanan, On the robustness of overall f0-only modifications to the perception of emotions in speech. J. Acoust. Soc. Am. **123**, 4547–4558 (2008)
4. F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech, in *Interspeech*, pp. 1–4 (2005)
5. J.P. Cabral, L.C. Oliveira, Emo voice: a system to generate emotions in speech, in *Interspeech*, pp. 1798–1801 (2006)
6. J.P. Cabral, L.C. Oliveira, Pitch-synchronous time-scaling for prosodic and voice quality transformations, in *Interspeech*, pp. 1137–1140 (2005)
7. F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech. Spoken Language, in *ICSLP 96*, pp, 1970–1973 (1996)
8. K.T. Deepak, S.R.M. Prasanna, Epoch extraction using zero band filtering from speech signal. Circuits Syst. Signal Process. **34**(7), 2309–2333 (2015)
9. P. Deshpande, M.S. Manikandan, Effective glottal instant detection and electroglottographic parameter extraction for automated voice pathology assessment. IEEE J. Biomed. Health Inf. **PP**(99), 1–11 (2017)
10. K. Dragomiretskiy, D. Zosso, Variational mode decomposition. IEEE Trans. Signal Process. **62**(3), 531–544 (2014)
11. T. Drugman, P. Alku, A. Alwan, B. Yegnanarayana, Glottal source processing: from analysis to applications. Comput. Speech Lang. **28**(5), 1117–1138 (2014)
12. T. Drugman, T. Dutoit, Glottal closure and opening instant detection from speech signals, in *Interspeech*, pp. 2891–2894 (2009)
13. S.R. Dumpala, K.V. Sridaran, S.V. Gangashetty, B. Yegnanarayana, Analysis of laughter and speech-laugh signals using excitation source information, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 975–979 (2014)
14. Z. Gao, X. Wang, J. Lin, Y. Liao, Online evaluation of metal burn degrees based on acoustic emission and variational mode decomposition. Measurement **103**, 302–310 (2017)
15. J. Gilles, Empirical wavelet transform. IEEE Trans. Signal Process. **61**(16), 3999–4010 (2013)
16. D. Govind, Epoch based dynamic prosody modification for neutral to expressive conversion, Ph.D Thesis, http://gyan.iitg.ernet.in/handle/123456789/363. Accessed 10 July 2017
17. D. Govind, P. Hisham, D. Pravena, Effectiveness of polarity detection for improved epoch extraction from speech, in *National Conference on Communication (NCC)*, pp. 1–6 (2016)
18. D. Govind, S.R.M. Prasanna, Epoch extraction from emotional speech, in *International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5 (2012)
19. D. Govind, S.R.M. Prasanna, Expressive speech synthesis: a review. Int. J. Speech Technol. **16**(2), 237–260 (2013)
20. D. Govind, S.R.M. Prasanna, B. Yegnanarayana, Neutral to target emotion conversion using source and suprasegmental information, in *Interspeech*, pp. 2969–2972 (2011)
21. D. Govind, R. Vishnu, D. Pravena, Improved method for epoch estimation in telephonic speech signals using zero frequency filtering, in *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 11–15 (2015)
22. N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Royal Soc. Lond. A Math. Phys. Eng. Sci. **454**, 903–995 (1988)
23. S. R. Kadiri, P. Gangamohan, S.V Gangashetty, B. Yegnanarayana, Analysis of excitation source features of speech for emotion recognition, in *Interspeech*, pp. 1324–1328 (2015)
24. S.R. Kadiri, B. Yegnanarayana, Epoch extraction from emotional speech using single frequency filtering approach. Speech Commun. **86**, 52–63 (2017)

25. S.R. Kadiri, B. Yegnanarayana, Speech polarity detection using strength of impulse-like excitation extracted from speech epochs, in *ICASSP)*, pp. 5610–5614 (2017)
26. S.R. Kadiri, B. Yegnanarayana, Analysis of singing voice for epoch extraction using zero frequency filtering method, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4260–4264 (2015)
27. V. Khanagha, K. Daoudi, H. Yahia, Detection of glottal closure instants based on the microcanonical multiscale formalism. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1941–1950 (2014)
28. S.G. Koolagudi, S. Devliyal, B. Chawla, A. Barthwal, K.S. Rao, Recognition of emotions from speech using excitation source features. Procedia Eng. **38**, 3409–3417 (2012)
29. S.G. Koolagudi, R. Reddy, K.S. Rao, Emotion recognition from speech signal using epoch parameters, in *International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5 (2010)
30. A.I. Koutrouvelis, G.P. Kafentzis, N.D. Gaubitch, R. Heusdens, A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(2), 316–328 (2016)
31. S.R. Krothapalli, S.G. Koolagudi, Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. **16**(2), 181–201 (2013)
32. K.S. Kumar, M.S.H. Reddy, K.S.R. Murty, B. Yegnanarayana, Analysis of laugh signals for detecting in continuous speech, *Interspeech*, pp. 1591–1594 (2009)
33. G.J. Lal, E.A. Gopalakrishnan, D. Govind, Accurate estimation of glottal closure instants and glottal opening instants from electroglottographic signal using variational mode decomposition. Circuits Syst. Signal Process. **37**(2), 810–830 (2018)
34. A. Mert, ECG feature extraction based on the bandwidth properties of variational mode decomposition. Physiol. Meas. **37**(4), 530–543 (2016)
35. K.S.R. Murty, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1602–1613 (2008)
36. P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans. Audio Speech Lang. Process. **15**(1), 34–43 (2007)
37. Noisex-92, www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html. Accessed 9 Dec 2017
38. S.R.M. Prasanna, D. Govind, Analysis of excitation source information in emotional speech, in *Interspeech*, pp. 781–784 (2010)
39. A.P. Prathosh, T.V. Ananthapadmanabha, A.G. Ramakrishnan, Epoch extraction based on integrated linear prediction residual using plosion index. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2471–2480 (2013)
40. L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, C.A. McGonegal, A comparative performance study of several pitch detection algorithms. IEEE Trans. Audio Speech Lang. Process. **24**(5), 399–418 (1976)
41. K.S. Rao, B. Yegnanarayana, Prosody modification using instants of significant excitation. IEEE Trans. Audio Speech Lang. Process. **14**, 972–980 (2006)
42. K.R. Scherer, Vocal affect expressions: a review and a model for future research. Psychol. Bull. **99**, 143–165 (1986)
43. K.P. Soman, P. Prabaharan, S. Athira, K. Harikumar, Recursive variational mode decomposition algorithm for real time power signal decomposition. Procedia Technol. **21**, 540–546 (2015)
44. D. Talkin, A robust algorithm for pitch tracking, in *Speech Coding and Synthesis*, ed. by W.B. Kleijn, K.K. Paliwal (Elsevier, New Providence, 1995), pp. 495–518
45. S.A. Thati, K.S. Kumar, B. Yegnanarayana, Synthesis of laughter by modifying excitation characteristics. J. Acoust. Soc. Am. **133**(5), 3072–3082 (2013)
46. M.R.P. Thomas, J. Gudnason, P.A. Naylor, Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. IEEE Trans. Audio Speech Lang. Process. **20**(1), 82–91 (2012)
47. A. Upadhyay, R.B. Pachori, Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition. J. Frankl. Inst. **352**, 2679–2707 (2015)
48. A. Upadhyay, R.B. Pachori, A new method for determination of instantaneous pitch frequency from speech signals, in *IEEE Signal Processing and Signal Processing Education Workshop*, pp. 325–330 (2015)
49. WAVESURFER, https://www.speech.kth.se/wavesurfer. Accessed 6 Mar 2017
50. C.E. Williams, K. Stevens, Emotions and speech: some acoustic correlates. J. Acoust. Soc. Am. **52**, 1238–1250 (1972)

51. Y.J. Xue, J.X. Cao, D.X. Wang, H.K. Du, Y. Yao, Application of the variational-mode decomposition for seismic time–frequency analysis. IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. **9**(8), 3821–3831 (2016)
52. W. Yang, Z. Peng, K. Wei, P. Shi, W. Tian, Superiorities of variational mode decomposition over empirical mode decomposition particularly in time–frequency feature extraction and wind turbine condition monitoring. IET Renew. Power Gener. **11**, 443–452 (2016). https://doi.org/10.1049/iet-rpg.2016.0088
53. B. Yegnanarayana, K.S.R. Murty, Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio Speech Lang. Process. **17**(4), 614–624 (2009)