IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 June 2024, accepted 1 July 2024, date of publication 9 July 2024, date of current version 25 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3425394

RESEARCH ARTICLE

Significance of Variational Mode Decomposition for Epoch Based Prosody Modification of Speech With Clipping Distortions

M. RAMA RAJESWARI, D. GOVIND, SURYAKANTH V. GANGASHETTY, (Member, IEEE), AND AKHILESH KUMAR DUBEY

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh 522302, India

Corresponding author: D. Govind (d_govind@kluniversity.in)

This work was partially supported from the project titled ''Speech Quality Control'' (Reference Number: 11(1)/2022-HCC(TDIL), dated: 18.02.2022) which is one of the consortium members of ''Speech Technologies in Indian Languages'', Under the scheme ''National Language Translation Mission (NLTM): Bhashini''. In this regard, the authors are thankful to consortium members of IIT Madras and Ministry of Electronics and Information Technology (MEITY), Government of India.

ABSTRACT Clipping is one of the non-linear distortions commonly introduced due to microphone saturation during speech recording. Present work focuses on the effect of clipping in the task of prosody modification. Since, F_0 contour and duration are the important prosodic parameters, the present work studies the effect of clipping in the manipulation of F_0 and duration of a given speech. Epoch based prosody modification is considered as the popular method to generate waveforms with good perceptual quality by scaling F_0 contour and duration of the given speech by fixed scaling factors. Therefore, present work studies the effect of waveform clipping on the perceptual quality of prosody modified speech. Deviations in the estimation of epochs (which are used as the analysis pitch marks) and method used for generating the waveform are the two ways wherein perceptual quality in epoch based prosody modification can be compromised. The work proposed in this paper examines, effect of clipping on the aforesaid stages of epoch based prosody modification affecting the perceptual quality of the generated speech. Zero frequency filtering (ZFF), a simple and popular method, is chosen as the epoch estimation algorithm for epoch based prosody modification presented in the paper. Based on comparative epoch estimation performance analysis carried out by introducing various amplitude clipping levels, epoch identification rates are confirmed to be unchanged, irrespective of the level of clipping distortions present. However, due to saturation in the waveform samples, the waveform generation stage of the prosody modification was observed to be affected to the level which was proportional to the clipping distortions present in the signal. A variational mode decomposition (VMD) based signal approximation of the prosody modified speech is proposed to reduce the non-linear effect due to clipping. At the gross level, the re-estimated speech signal obtained from the VMD modes observed to have improved the perceptual quality of the pitch and duration modified speech. The improved perceptual quality of VMD based re-estimation of prosody modified speech was confirmed from subjective and NIST-STNR based objective assessments. Further, VMD based refinement is proposed as an alternative to local mean subtraction for trend removal in conventional ZFF of speech for the accurate epoch estimation. Comparative performance analysis carried out on CMU arctic database, confirmed improvement in the identification accuracy for the epochs estimated by using VMD based trend removal in ZFF algorithm.

INDEX TERMS Clipping distortion, epochs, perceptual quality, prosody modification, variational mode decomposition, zero frequency filtering.

I. INTRODUCTION

Development of data driven End-to-End speech systems marked phenomenal advancement in the field of speech and

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

language processing. Performances of these systems depend solely on the large amount of data collected. Among the large repository of databases prepared by crawling from various internet sources, utterances with non-linear distortions such as clipping are commonly present [1]. Therefore, the work presented in the paper focuses to explore the effect of clipping distortions in the analysis and processing of voice quality parameters of prosody such as F_0 and duration in speech signals.

Clipping distortions in speech occur due to saturation of microphones as a result of improper calibration of automatic gain control (AGC), rise in loudness of speakers during recording and generation of speech files without properly normalizing the waveforms. According to Harvella and Stern, amplitude clipping (clip by amplitude) for a speech signal is mathematically expressed as given in Eq. 1 [2].

$$s_c[n] = \begin{cases} s[n], & s[n] \leq |\tau| \\ |\tau\|, sgn(s[n]), & s[n] > |\tau| \end{cases} \quad (1)$$

According to Eq. 1, samples are assigned a constant value for all the amplitudes of $s[n]$ above τ. For a speech signal normalized between $[-1, 1]$, the percentage of clipping is given by the factor $(1 - \tau) \times 100$.

Figure 1, shows the speech waveform and amplitude histograms for various clipping amplitudes. A significant change in the distribution of amplitudes (Figure 1((g)-(j))) affect signal characteristics as compared to that of the given speech signal having no clipping distortion (Figure 1(f)). It has to be noted that Figure 1 demonstrates the clip by amplitude as reported by Hansen et al. in [1]. However, clip by sample measured in (%) is computed as the ratio of the samples contaminated by clipping to the total number of samples present in the given speech utterance. In the Figure 1, the clip by amplitude level by Eq. 1 for clipping amplitude of 0.99 (0.1%) corresponds to 0.01% clip by sample level, approximately. Similarly, extreme clip by sample level of 13% corresponds a clip by amplitude level of 99% (clipping amplitude of 0.01).

There are many works reported for restoring clipped samples of speech signals in the context of audio inpainting [3], [4]. By subjecting the speech signal to auto regressive (AR) modeling, Janssen et al. [5] proposed a method to restore contaminated samples due to clipping. By the higher order AR modeling of speech, authors could estimate samples distorted by amplitude and temporal clipping which provided minimum error in a least square (LS) sense [5]. Although, there were earlier attempts to restore the distorted speech samples in speech by localized estimation of time and frequency, optimization based approaches were popularized by the unpublished work by Selesnick based on LS optimization [6]. LS method proposed by Selesnick, estimates the missing samples in a waveform by optimizing the second order derivative of the smoothness of the estimated samples of each frame. As an extension to the estimation of missing samples, the speech de-clipping was achieved by minimizing the third derivative of mean square energy of estimated samples. Exploiting the sparse coding representation for compressed sensing, Adler et al. proposed an audio inpainting algorithm for speech declipping [3]. The speech declipping was formulated as an inverse problem of audio frame reconstruction using the constrained orthogonal matching pursuit algorithm. The sparse modeling was carried out using Gabor dictionaries. Motivated by the audio inpainting work proposed by Adler et al., Kritic et al., proposed an iterative hard thresholding (IHD) based approach which was used for faster convergence and generating the dictionary of sparse code vectors for estimating the sparse representation of the input speech frame. Declipping was then considered as the inverse problem to reconstruct the signal from the partial observation of signal samples from the given speech frame [7]. Harvilla and Stern reported the effect of clipping on the speech recognition performance [2]. A significant drop in word error rates were observed as the clipping amplitude thresholds were increased. Further, the work by Harvilla et al. proposes a speech declipping algorithm by combining Selesnick's LS based unconstrained sample estimation and IHT based sparse coding representation proposed by Harvilla and Stern [2]. In the proposed work, the speech samples were estimated to be above the clipping threshold, τ which was known a priori. A constrained blind amplitude reconstruction algorithm was then used for the estimation of clipped samples for the inverse sparse reconstruction in the least square sense.

In addition to the aforesaid optimization based approaches for speech declipping, there are a few works on audio inpainting recently reported in the literature which are based on deep neural networks, particulary on generative modeling [4], [8], [9]. However, most of the works on audio inpainting using generative modeling focused predominantly on the temporal clipping where audio samples were missed for a long gap of 100 ms to a few seconds using various generative adversarial network (GAN) architectures [4], [8]. The recent work by Hansen et al. studied the effect of clipping distortion severities on performances of speech system by taking speaker identification task as the case [1]. Here, the speech amplitude clipping was introduced in the range of 0 to 15% (clipping by samples). The first part of the study showed the effect of clipping on speech quality assessment parameters, non-reference NIST supported signal to noise ratio (STNR), waveform distortion assessment (WADA-SNR), PESQ and blind source separation (BSS) based measure which are treated as the objective measures that reflect the comparative mean opinion scores obtained by conducting subjective evaluation [10]. As reported by Hansen et al., when clipping levels were increased, the speech quality parameter measures were found to be reduced proportionately. Further, presence clipping observed to have a significant adverse impact on the equal error rates (EER) of state of the art speaker identification systems (SID). Excluding the distorted speech files while building the speaker models showed no degradation in EER. Finally,

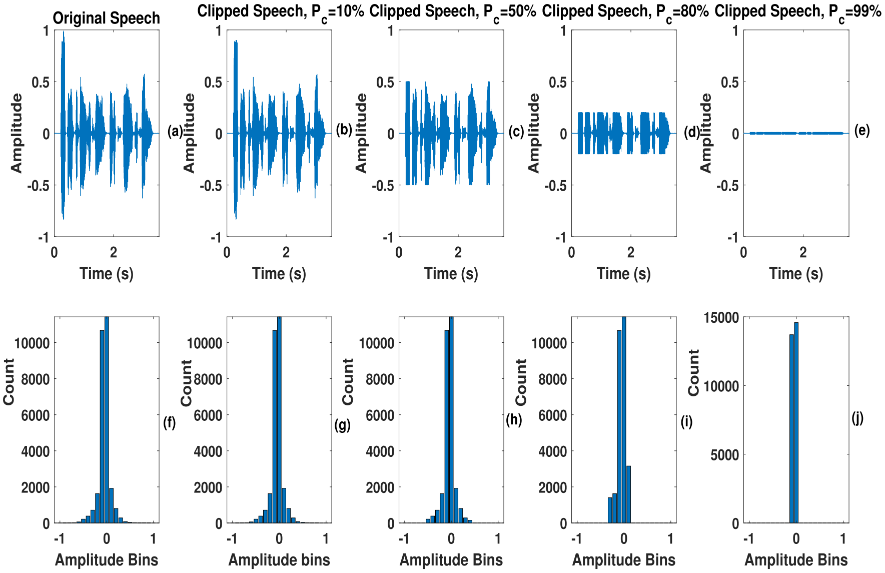

FIGURE 1. Waveform and amplitude histograms of original speech ((a) and (f)), speech with clip by amplitude levels of 0.1 ((b) and (g)), 0.5 ((c) and (h)), 0.8 ((d) and (i)) and 0.99 ((e) and (j)).

the work concluded with the remarks that there were no noticeable differences in the performances of systems when the clipped samples were under 1%.

Similar to the works by Hansen et al. and Harvilla et al. where the effect of speech clipping was studied by considering speaker identification and speech recognition, respectively as the cases, the present work explores the effect of clipping in prosody manipulation and attempts to improve the perceptual quality of the prosody modified speech. Prosody manipulation is one of the stages in many speech synthesis applications such as synthesis of emotions [11], speaker anonymization [12] for privacy preservation and so on. Modifying the pitch (F_0) contour and duration of the given speech signal is termed as prosody manipulation [13] [14], [15], [16], [17], [18]. Pitch synchronous overlap (PSOLA) applied in time domain (TD- PSOLA), frequency domain (FD-PSOLA) and on linear prediction (LP) redual (LP-PSOLA) are the popular methods for achieving prosody modified speech [19], [20]. Considering the simple and computationally efficient algorithmic implementation, epochs estimation using ZFF of speech is used to extract the analysis pitch marks for the studies presented in this paper. Other well performing epoch estimation methods include dynamic programming projected phase slope algorithm (DYPSA) [21] [22], Integrated linear prediction residual (ILPR) [23], Speech event detection using residual excited and mean based signal (SEDREAMS) [24] and so on. By keeping the prosody modification as the task, the objectives of the work presented in the paper are formulated to address the following issues:

- To check the effect of clipping distortions on he accuracy of epochs estimate for prosody modification
- To device methodologies to reduce the effect of clipping distortion in pitch marks estimation and waveform generation stages of prosody modification

To reduce the effect of clipping on the perceptual quality of the prosody modified speech, restoration of the clipped samples are essential. The clipped samples can be restored by computing the best estimates of predominant variations at the time-frequency scale around dominant frequencies. There are a few signal processing tools which compute the time-frequency estimated for the restoration of the waveforms. Among those methods popular one is empirical mode decomposition (EMD) based signal approximation by estimating intrinsic mode functions (IMF) [25]. EMD computes IMFs iteratively by estimating the signal extrema from the amplitude envelopes. In the case of signal contamination by clipping, the envelopes are distorted and therefore the IMFs computed from the signal extrema may not be reliable. As an alternative, variational mode decomposition (VMD) based method for computing IMFs is introduced which out performed EMD in the presence signal distortions due to noise [26]. In VMD, IMFs representing the AM-FM components are used to reconstruct the signal to reduce the clipping distortion. The unique aspect of VMD which motivated us to propose for the enhancement of clipped speech is the estimation of signal samples of IMFs by constraining the signal variations around the prominent frequencies compared to IMF signal mode decomposition in EMD. In EMD, IMFS are estimated iteratively computing extreme points obtained from amplitude envelope of the given speech signal [27].

Therefore, the proposed novelties of the paper are listed below:

- Studies carried on the effect of clipping on epochs estimation from clipped speech
- Proposing a refined epoch estimation approach for clipped speech signals using VMD
- An attempt to improve the perceptual quality of pitch and duration modified clipped speech signals using VMD

The proposed organization of the paper is given below: Section II describes the formulation of variational mode decomposition for estimating the IMFs in speech signals. Studies on the effect of clipping in estimation of epochs are given in Section III. An alternate refinement is proposed for conventional ZFF method for the robust estimation of epochs is provided in Section IV. A VMD based method for reducing perceptual distortions due to clipping has been described in Section V. Section VI summarizes the work with remarks on the permissible levels of clipping so that distortions introduced in prosody modified speech are perceptually unnoticeable.

II. VARIATIONAL MODE DECOMPOSITION FOR ESTIMATING INTRINSIC MODE FUNCTIONS OF SPEECH

The VMD decomposes the given speech signal into real valued signals which are termed as intrinsic modes and are conveniently called as modes. Mode decomposition is carried out by minimizing the bandwidth of frequency variations around the center frequency of each mode. Therefore, estimating modes and their center frequencies are formulated as quadratic programming optimization problem which is shown in Eq. 2.

$$\min_{(s_k,\omega_k)} \sum_{k=1}^{K} \|\partial_t[(\delta(t) + \frac{j}{\pi t}) * s_k(t)]e^{-j\omega_k t}\|_2^2$$
$$subject\ \ to: \ \ s(t) = \sum_{k=1}^{K} s_k(t) \tag{2}$$

where, K is the number of modes which is specified by the user and ω_k is the center frequency of each mode $s_k(t)$. The objective function of the quadratic programming problem shown in Eq 2, is set to minimize the bandwidth of each mode $s_k(t)$ around the center frequency ω_k. The mode band width optimization problem is constrained to reconstruct the signal by adding all modes. In Eq 2, the bandwidth of each mode, k, is computed by deriving the analytic signal and modulating the one sided spectrum to the baseband with center frequency, ω_k. In this way, each mode is compact with respect to the bandwidth around its center frequency. The required modes and the center frequencies are estimated by solving the quadratic programming optimization using quadratic penalty term and Lagrangean multipliers, λ, for adding the reconstruction constraint. Eq. 3 combines the constraints with the mode bandwidth objective function using Lagrangean multipliers. The parameter variable α in Eq. 3 is the quadratic penalty on the bandwidth constraint.

$$\begin{aligned} L(s_k, \omega_k, \lambda) = \alpha \sum_{k=1}^{K} \|\partial_t[(\delta(t) + \frac{j}{\pi t}) * s_k(t)]e^{-j\omega_k t}\|_2^2 \\ + \|s(t) - \sum_{k=1}^{K} s_k(t)\|_2^2 \\ + \langle\lambda(t), s(t) = \sum_{k=1}^{K} s_k(t)\rangle \end{aligned} \tag{3}$$

Eq. 3 has been solved using alternate direction multipliers method (ADMM) to estimate modes from the given signal $s(t)$ [28], [29]. Subproblems to update each mode,s_k and ω_k are solved in the frequency domain by applying Fourier Parseval's relation [26]. The expression for mode update in the spectral domain is given in Eq. 4.

$$\begin{aligned} s_k^{n+1} = \arg \min_{\hat{s_k}, s_k \epsilon X} \alpha\|j\omega[(1 + sgn(\omega + \omega_k).\hat{s_k}(\omega + \omega_k)\|_2^2 \\ + \|\hat{s}(\omega) - \sum_{k} \hat{s_k}(\omega) + \frac{\hat{\lambda(\omega)}}{2}\|_2^2\} \end{aligned} \tag{4}$$

$$\omega_k^{n+1} = \arg \min_{\omega_k} \int_0^\infty (\omega - \omega_k)^2 |\hat{s}_k(\omega)|^2 \underline{d}\omega\} \tag{5}$$

Eq. 4 and Eq. 5 iteratively estimate each mode,s_k (Eq. 4 and central frequency of each mode ω_k, where n corresponds to the sample index. Both the equations ensure that for every iteration, bandwidths of estimated modes, $\hat{s_k}$, with the estimated central frequency $\hat{\omega_k}$, are minimized. Stopping criteria for the iterative ADMM is set when the absolute difference of the bandwidths in the successive iterations falls below ϵ.

A. VMD FOR APPROXIMATING CLIPPED SAMPLES

As described by Dragomiretskiy et al., according to the requirements, constraints of the optimization problem given in Eq 3 can be modified. In the context of restoration of clipped sample amplitudes, the equality constraints for the prefect reconstruction of the signal from the modes need not be enforced. This is true with many of the VMD based signal denoising applications reported earlier [26], [30]. For the estimation of approximate sample amplitude values during the estimation VMD modes, the update of Langrangean multipliers (λ) for combining the perfect reconstruction constraints from the modes are turned off. Hence, λ term given in the Eq. 4 vanishes and the resulting estimated modes provide approximate sample amplitude for the clipped amplitudes in the original signal.

III. EFFECT OF CLIPPING IN EPOCH ESTIMATION

The focus of the present work is to first analyze the effect of clipping on epoch estimation. Epochs represent analysis pitch marks, the accuracy with which pitch marks are estimated is crucial in generating prosody modified speech with better perceptual quality. In earlier works, we have

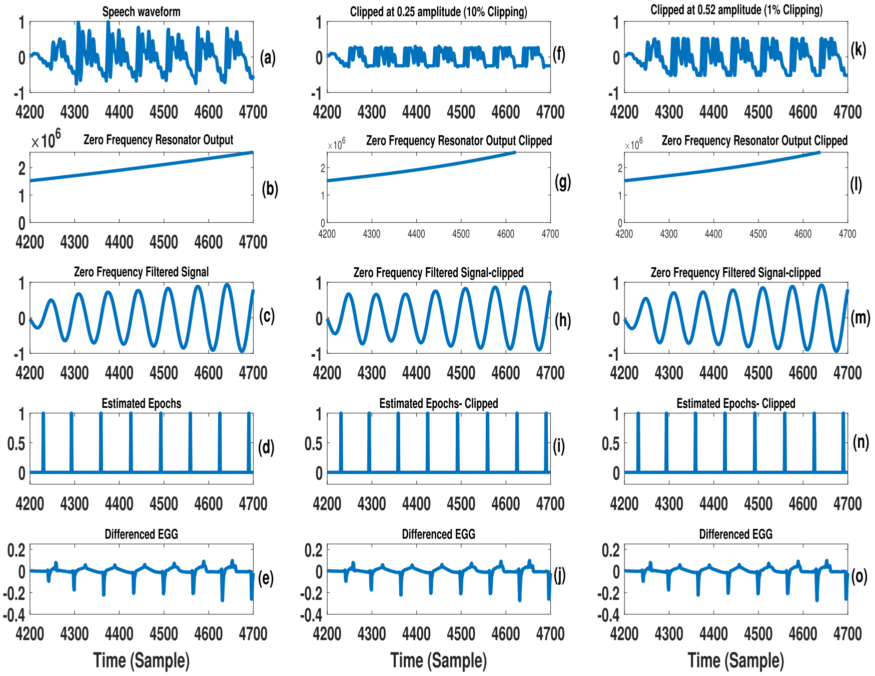

FIGURE 2. Effect of clipping in epoch estimation. The waveform, output of the zero frequency resonator, zero frequency filtered signal, estimated epochs and differenced EGG representing reference epochs of original speech ((a)-(e)), clipped speech above 0.25 absolute amplitude ((f)-(j)) and clipped speech above 0.52 absolute amplitude ((k)-(o)).

shown how deviations of estimated epochs adversely affect perceptual quality of the prosody modified speech [31], [32]. Therefore, it is essential to check the effect of clipping in epoch estimation performance and device methodologies for the improvement. The ZFF algorithm is chosen as the method for estimating epochs from clipped speech signals [33], [34], [35].

In ZFF method, the differenced speech signal is integrated four times. Two times integration of a sequence corresponds to filtering through resonator located at 0 Hz which is termed as zero frequency resonator (ZFR). Speech is then filtered through a cascade of two resonators to obtain a polynomially growing/decaying sequence with a better roll-off. To compute variations due to epochs, the local mean substraction is carried out using a window length equivalent to average pitch period of the given signal. The local mean subtracted signal is named as zero frequency filtered signal (ZFFS). The positive zero crossings of the ZFFS are hypothesized as the epochs in speech. Eventhough, there are a few studies reported earlier which addressed issue of trend removal of the ZFR output by local mean subtraction using fixed length windows to derive ZFFS [16], [36], [37]. Presence of spuriously estimated epochs were observed in speech segments with rapid F_0 variations such as in the case of emotions, laughter, vocal singing and so on. However, for clean speech signals with moderate noises, ZFFS derived using fixed window length provided reliable epoch estimation as compared to other existing epoch extraction alogorithms. Less computational complexity, reliable performance and reduced tuning parameters are the factors that makes ZFF method a default choice for reliable epoch estimation [38], [39].

The subplots in Figure 2((a)-(e)) show the voiced segment of clean speech (a), corresponding ZFR output (b), ZFFS (c), estimated epochs (d) and the corresponding differenced EGG segment (e) showing the groundtruth epoch locations as the prominent negative peak. It has to be noted that the epoch locations estimated from speech plotted in the subplot (d) coincide with the negative peaks of the differenced EGG.

The performance measures used to assess an epoch estimation algorithm are the identification rate (IDR), miss rate (MR), false alarm rate (FAR) and epoch identification accuracy (IDA) [21], [24]. Epoch identification rate is computed as the number of estimated candidate epochs identified within larynx cycle defined by the reference epochs. Missing rate gives the count of reference larynx cycles where there are no estimated epochs. False alarm is where there are more than one epochs estimated within the larynx cycle. Finally, epoch identification accuracy gives the standard deviation of the sample deviations of candidate epochs with respect to reference epochs of the given larynx cycle. Table 1 provide the epoch estimation performance of ZFF method when evaluated using CMU-Arctic Database [40]. The Arctic database has simultaneous speech and EGG recordings obtained for 1132 phonetically balanced utterances. Three speakers (two male (JMK and BDL) and one female (SLT)) of CMU-Arctic database are considered for the epoch performance evaluation. All

TABLE 1. Epoch estimation performance of ZFF of speech for 3 Speakers in CMU-Arctic database with reference epochs derived from corresponding EGG recordings.

Spkr	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Tot. Ref. Epochs
BDL	99.43	0.03	0.53	0.29	203650
JMK	99.8	0.05	0.15	0.47	133089
SLT	98.93	0.03	1.04	0.32	320193
Combined	99.39	0.04	0.58	0.36	656932

the speakers speak US-accented English language. Each utterance in the Arctic database is stereo recorded in a clean noise free anechoic studio with speech and EGG signals in left and right channels, respectively. As reported by Murty et al., ZFF showed superior performance compared to DYPSA, Hilbert envelope and group delay (GD) epoch estimation algorithms.

Various amplitude thresholds were put ranging from (0.1 to 0.9) to the speech waveform which was normalized in the range [−1,1].

To check the effect of clipping in the estimation of epochs, speech samples were contaminated at various levels by clip by amplitude for various values of τ. Different levels of clipping distortions were introduced according to Eq. 1, as shown in earlier section. Eventhough, there are differences in measuring clipping levels in speech signals, clipping is introduced by setting the amplitude threshold. Harvilla et al. compute the clipping percentages based on the threshold [2]. However, Hansen et al. computed the clipping percentage as the percentage of samples in the speech signal which were affected by the amplitude threshold. For instance, an amplitude threshold, $\tau = 0.1$ results in a clipping percentage of around 10% when computed from the amplitude histogram of the speech samples [1]. In the present work, both clip by amplitude and clip by sample methods are used for introducing clipping distortions.

To study the effect of clipping in epoch estimation, the clipping threshold τ varied in steps from 0.1 to 1. The τ variations inturn result in variation of 0%-15% samples affected by clipping distortion (where 0% being no samples affected by clipping distortion for a $\tau = 1$). Subplots in the Figure 2 ((f)-(i) &(k)-(l)), show the variation in the segments of ZFR output, ZFFS and estimated epochs. In comparison with the original unclipped signal segments plotted in Figure 2 ((a)-(e)), there are no changes observed in the characteristics of ZFR and ZFFS segments. Further, it has to be noted that there are no significant deviations introduced in estimated epochs from the waveform with clipping distortions. To check the statistical consistency of visual analysis of a voiced segment for various clipping levels, the performance measures were computed for all utterances of CMU-Arctic database with three speakers (BDL, JMK and SLT) by varying *tau* from 0.1 to 1 in the steps of 0.1. Figure 3 plots IDR and IDA measures versus clipping amplitude τ. The analysis of Figure 3 reinforces the inference drawn from the visual analysis of the epoch estimation. However, Figure 3 indicates that the epoch IDR and IDA showed marginal drop when the clipping threshold was kept very low at $\tau = 0.1$ which corresponds to 10-15% clip by sample percentage of the speech utterance. The IDR and IDA measures are remained same as the performance obtained for the unclipped version of CMU arctic database shown in the Table 1. Based on the comparative performance analysis, clipping distortions were not affecting the epochal information present in speech. This also hints us the possibility of reliable estimation of excitation source features that can be computed from the epochs such as strength of excitation (SoE), instantaneous pitch and so on.

$$x(t) = 6t^2 + cos(10\pi t + 10 * pi * t^2) + \begin{cases} cos(60\pi t), & t \leq 0.5 \\ cos(80\pi t - 10\pi), & t > 0.5 \end{cases} \tag{6}$$

Figure 4 plots the modes of non-stationary signal $x(t)$ as indicated in the Eq 6. It has to be noted that the nonlinear trend in the signal has been captured by mode 1 as plotted in Figure 4 (b). The remaining non-stationary components are captured by other two modes as indicated in the subplots (c) and (d). Motivated by this example of extracting non-linear functional components of the given signal, trend in the ZFR output can be estimated using VMD. The mode representing the non-linear trend can be discarded and the subsequent reconstructed signal can have ZFFS characteristics.

IV. REFINED VMD BASED TREND REMOVAL IN ZERO FREQUENCY FILTERING BASED EPOCH ESTIMATION

In ZFF method,the output obtained by passing speech signal through a cascade two ZFRs is subjected to local mean substraction using fixed length windows for deriving ZFFS. However, in the case of speech signals having rapid pitch variations adaptive windows have to be used for estimating epochs reliably. For instance, if the pitch of a speech segment is high, local mean substraction with longer window causes missing of epochs where as use of short window length introduces spurious epochs [16], [36]. There are many post processing methods proposed earlier to adaptively estimate the window lengths and smoothing the ZFFS signal obtained from the conventional ZFF method to remove spurious zero crossings. In this section, the effectiveness of VMD for the trend removal from ZFR output is proposed to estimate the variations due to epochs.

Motivation: As described by Dragomiretskiy et al., VMD can be used to decompose the given non-stationary sequence into its constituent modes [26]. We are regenerating the same example which is given in [26] as a motivation for the trend removal of ZFR using VMD. As an example, Eq. 6 shows a synthetically generated signal having non-linear and non-stationary components.

Figure 5 demonstrates the trend removal from ZFR output using VMD. Similar to the example described earlier, the nonlinear polynomial growth/decay trend has been represented by one of the modes when the ZFR output is subjected to VMD. Figure 5(b) plots the growing

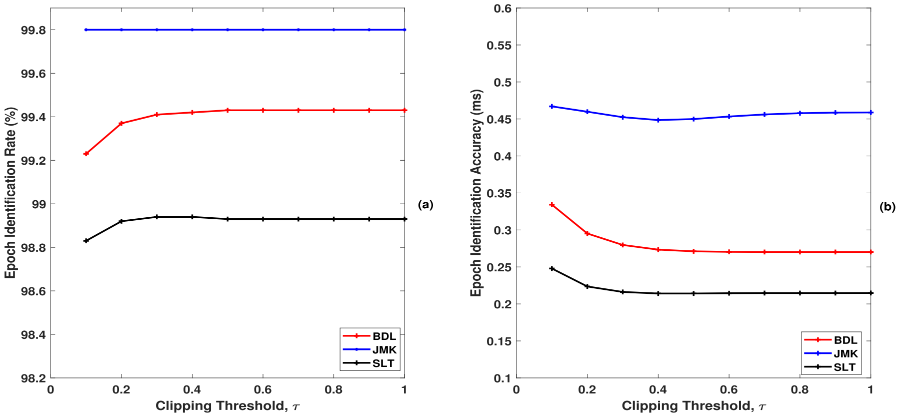

FIGURE 3. Variations in epoch IDR and IDA for different clipping amplitude thresholds,τ. (a) The epoch identification rate and (b) epoch identification accuracy.

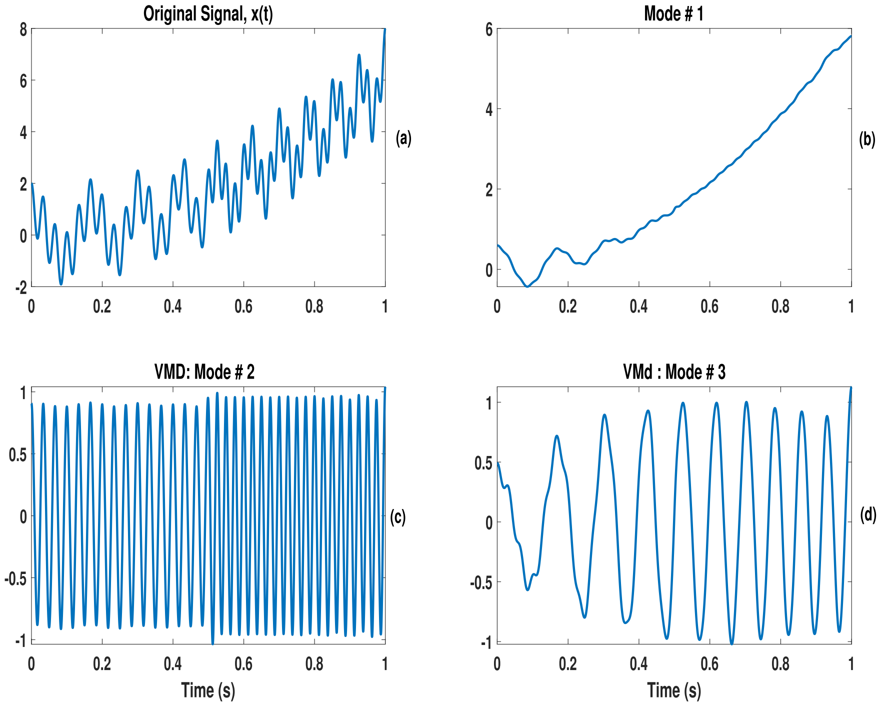

FIGURE 4. Constituent components of a non-linear and non-stationary signal. (a) The signal $x(t)$, (b) the mode representing the trend in the signal represented by the component, $6t^2$, (c) second mode representing the non stationary component $cos(10\pi t + 10 * pi * t^2)$ and (d) $\begin{cases} cos(60\pi t), & t \leq 0.5 \\ cos(80\pi t - 10\pi), & t > 0.5 \end{cases}$. (example given by Dragomretskiy et al. in [26].

non-linear trend captured by the mode 1 of ZFR output plotted in Figure 5(a). Subplots ((c)-(e)) show the subsequent first modes obtained from VMD of the signal reconstructed by discarding the mode signal plotted in Figure 5(b). The process of trend removal is depicted in the following block diagram shown in Figure 6. The output sequence from the ZFR is subjected to initial VMD to derive five modes. Since the predominant trend in the ZFR sequence is the polynomially growing/decaying function, the first mode captures the trend. Therefore, the subsequent processing has been carried out by discarding the estimated trend. The first level trend removed sequence is differenced (an operation equivalent to pre-emphasis filtering) prior to further subjecting the sequence to VMD. A series of 3 decompositions are carried on the sequence obtained from the first level of VMD. The modes which are discarded or included on each decomposition have been indicated in Figure 6 red and green colors, respectively. The differenced first mode from the 4^{th} level of VMD is considered as the trend removed sequence and is equivalent to ZFFS derived

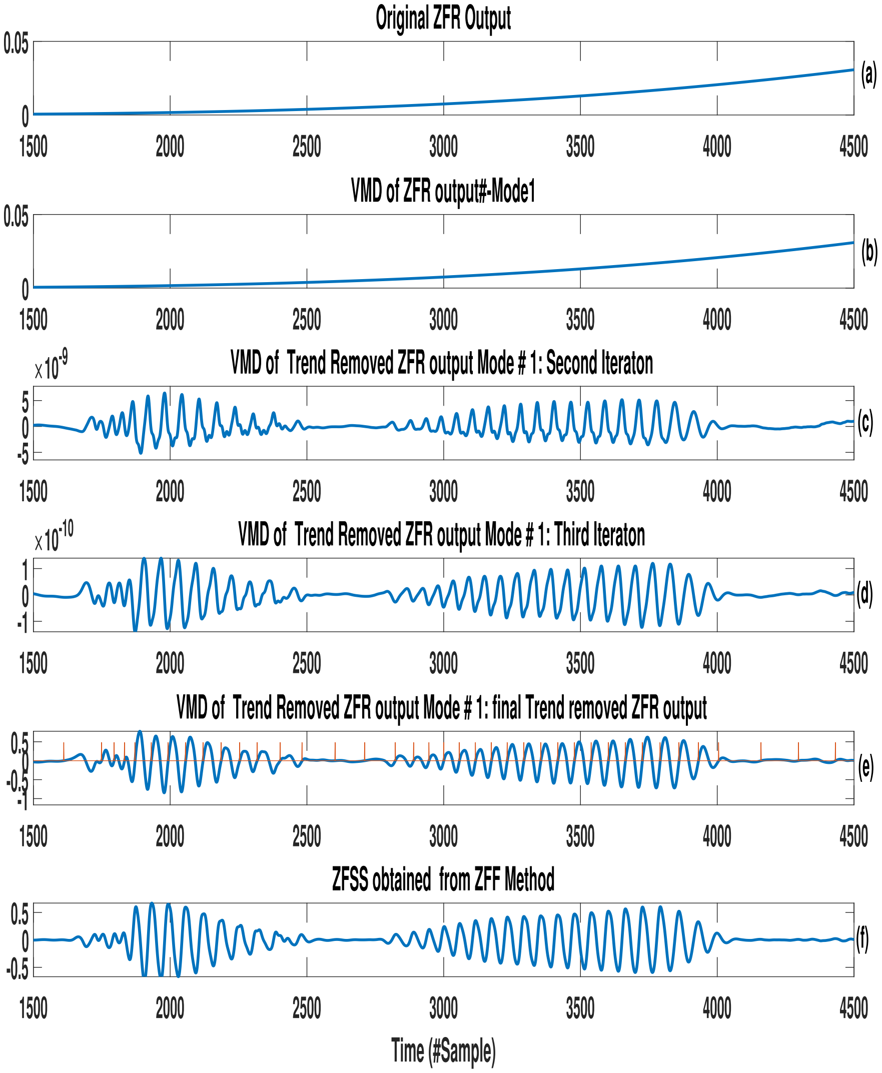

FIGURE 5. Proposed VMD based trend removal from ZFR output. (a) ZFR output obtained as the output of two ZFRs in cascade (b) the non-linear polynomially growing trend captured by the VMD (c) VMD Mode 1 obtained from the reconstructed signal by discarding the non-linear trend function in (b) (second VMD iteration) (d) VMD mode 1 obtained by subsequent decomposition of mode in (c), (e)mode 1 obtained from VMD of mode in (d) treated as the final trend removed ZFR output whohse zero crossings are hypothesized as epochs (plotted in red color) and (e)corresponding ZFFS degment obtained by local mean subtraction in the conventional ZFF method.

from the local mean subtraction of ZFR output in ZFF method. The positive zero crossings of the trend removed sequence are hypothesized as the epochs in the given speech signal.

Figure 7 plots the probability density function of epoch deviations for various epoch deviations,η (in ms). The distribution of epochs for various epoch deviation, η has been computed for three speakers in CMU arctic database. From the Figure 7, it has to be observed that more epochs have been estimated with reduced epoch deviation for ZFFS obtained using proposed VMD based trend removal (black colored line plot) than the conventional ZFFS obtained by local mean subtraction (red colored line plot) in ZFF method. The peak values of the density function of proposed method are located close to reference epoch locations as compared to the conventional ZFF method. However, for the SLT speaker, both the density curves obtained from both the methods are coinciding which indicates the comparable epoch identification accuracies. Other epoch estimation performance measures such as IDR, MR and FAR were found to be similar for both trend removal methods.

Figure 9 plots IDA obtained for proposed VMD based trend removal for various clipping amplitudes. It has to be note that for extreme clipping ($\tau \leq 0.2$) a significant rise in the epoch deviation was observed. However, as obtained for the conventional ZFF, the deviations are

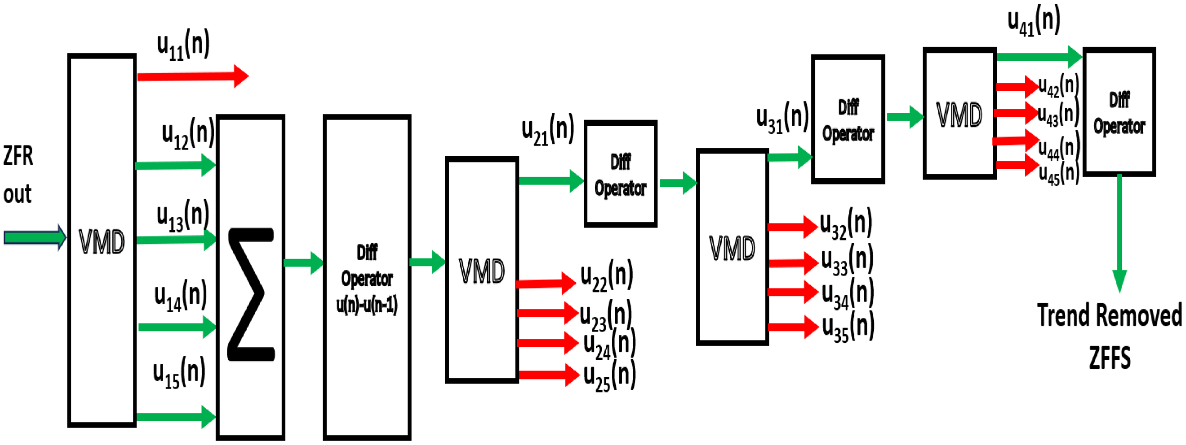

FIGURE 6. Block diagram for the proposed trend removal method from the sequence of ZFR output using VMD. Red color arrows indicate the discarded VMD modes and green color arrows indicate modes considered for subsequent processing. The trend removed ZFR provides the ZFFS.

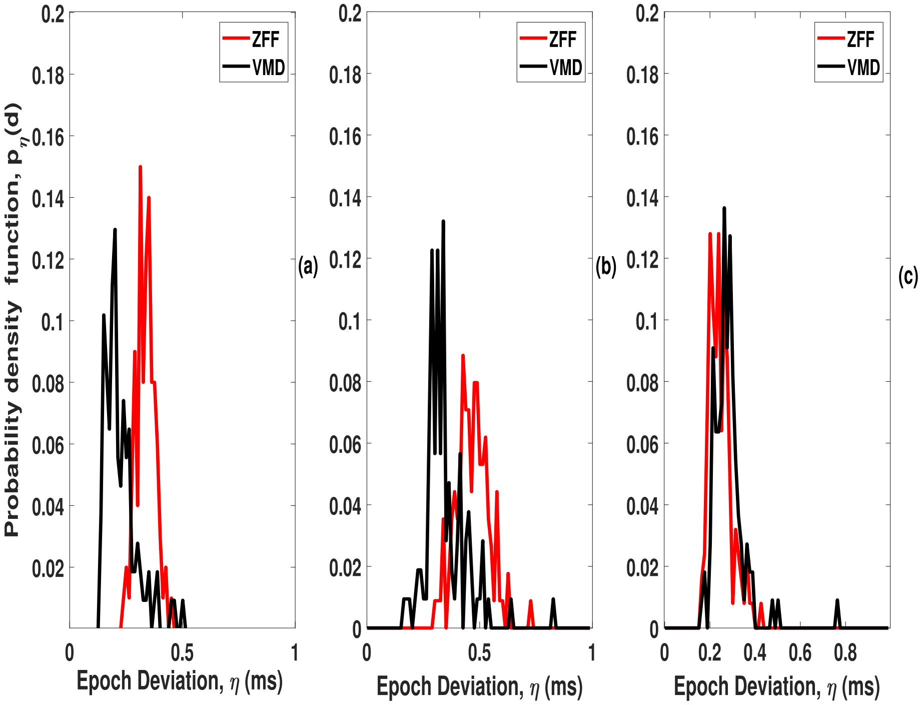

FIGURE 7. Probability density function of epoch deviations by considering the epoch deviation as the random variable for (a) BDL, (b) JMK and (c) SLT speakers of CMU arctic database.

noticeable for larger percentages of clipped samples in the waveform. For moderate clipping levels (clip by sample <10%), IDA measures more or less remain intact. Apart from IDA, epoch IDR, MR and FAR showed same values as obtained for clean speech independent of the clipping levels present. The google drive link to the MATLAB programs implementing the proposed epoch estimation algorithm using VMD is provided in the following link: https://forms.gle/7y5bVSFcDST3HEv8A.

For the comparative analysis of the proposed trend removal for ZFF using VMD, we have compared with the conventional ZFF method. However, Table 2 provides an extended performance comparison of the proposed refinement to ZFF algorithm with well known techniques such as group delay (GD), DYPSA and Hilbert envelope of linear prediction (LP) residual based epoch estimation method. Table 2 shows that the conventional ZFF significantly performs better than other methods in terms of IDR, MR, FAR and IDA. However, the

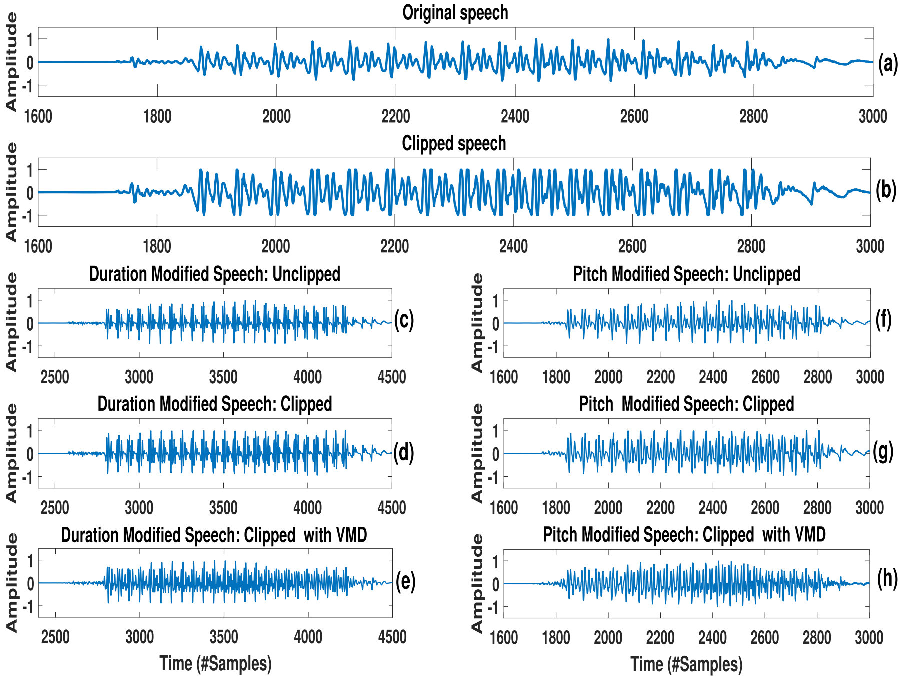

FIGURE 8. (a) A segment of original speech (b) corresponding clipped segment (clipping above 0.5 amplitude), (c) duration modified original speech segment (for the duration scale factor of 1.5),(d) duration modified segment for the clipped speech, (e) duration modified clipped speech segment with VMD based declipping, (f) pitch modified speech segment for the original speech segment (with pitch period scaling factor 0.8), (g) pitch modified clipped speech segment and (h) declipped pitch modified speech using VMD.

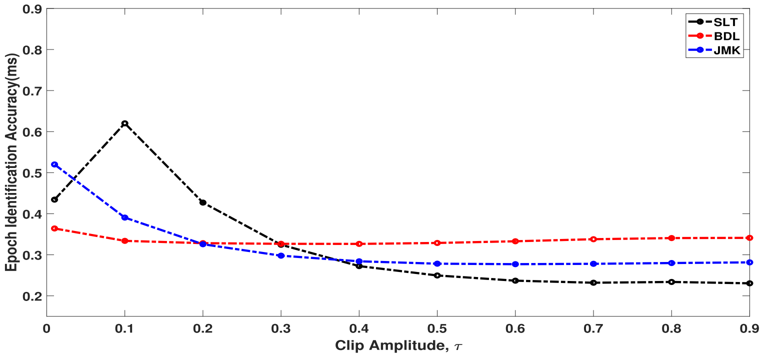

FIGURE 9. The IDA computed for various clipping amplitudes (τ) ranging from 0.01 to 0.9.

proposed VMD based refined trend removal method in ZFF further improves the epoch estimation performance.

V. RESTORATION OF CLIPPED SPEECH SAMPLES USING VARIATIONAL MODE DECOMPOSITION FOR PROSODY MODIFICATION

Restoration of the speech samples are proposed to be achieved by considering predominant variations around the central frequency of various VMD modes for reconstruction. The restoration of samples which are contaminated by clipping are re-estimated through Equations 4 and 5 of VMD. In the prosody modification task, synthesis pitch marks are derived according to the pitch and duration scaling factors from the sequence of analysis pitch marks (epochs) estimated from the original speech. Based on the experimental studies presented in Section III, the effect of clipping on the estimation of epochs is not significant. Since, the goodness of synthesis pitch marks depends only on the estimated epochs and prosody modification scale factors, the level of clipping distortions present in the original speech samples has no influence on the second stage of prosody modification. However, the third stage of waveform synthesis involves copying or resampling the samples in the original pitch cycles to fill the modified pitch periods (obtained from the second stage) of the sequence. Hence, the contamination of samples due to clipping causes significant reduction of

TABLE 2. Performance comparison of proposed method with conventional ZFF, Hilbert envelope of linear prediction residual (HE-LPR), DYPSA and group delay (GD) methods in CMU arctic database.

Method	IDR (%)	MR (%)	FAR (%)	IDA(ms)
Conventional ZFF of Speech [33]	99.39	0.04	0.58	0.36
HE-LPR [11]	96.21	01.47	02.32	0.60
DYPSA [21]	98.05	0.62	1.34	0.35
GD Algorithm [41]	94.48	4.07	1.45	0.45
Proposed Method	99.43	0.03	0.54	0.24

perceptual quality of the generated prosody modified speech. To reduce the effect of clipping on the prosody modified speech samples, the generated waveform is subjected to VMD. The refined waveform is reconstructed by using the VMD modes having predominant frequency variations. The methodology is similar to the VMD based signal denoising reported in various studies [26], [30], [42].

Figure 10 represents the steps involved in the VMD based enhancement of prosody modified speech. As per the block diagram following are the steps involved in the proposed VMD based speech enhancement to reduce perceptual distortion introduced due to clipping:

- Modes are derived from the prosody modified clipped speech using VMD
- Speech signal is reconstructed by discarding higher modes which predominantly capture the noise components due to distortions

Figure 8 plots the pitch and duration scaled segments of original and clipped versions of speech. The clipped segment is generated by saturating the amplitude values above 0.5 as shown in Figure 8(b). The duration modified speech segment for a duration modification of 1.5 is plotted as a subplot (c) in Figure 8 which is a stretched version of original speech segment plotted in subplot(a). Figure 8 (d) and (g) show the duration and pitch modified segments of clipped segment (subplot(b)), respectively. From the plots, clipping distortions can be observed in the amplitude envelopes of pitch and duration modified segments as compared to corresponding unclipped versions plotted in Figure 8 (c) and (f). Figure 8 (e) and (h) show the proposed VMD based reconstruction applied on the duration and pitch modified speech segments, respectively. Compared to clipped pitch and duration modified segments, VMD based processing of prosody modified speech reduced the envelope distortion to some extent. To improve the perceptual quality of waveforms which are plotted in Figure 8 (e) and (h), first 4 VMD modes are added by discarding the 5^{th} mode. Following section presents the empirical studies on the effect of number of VMD modes used for declipping the pitch and duration modified speech.

Assessment of VMD based declipping on the perceptual quality of the prosody modified speech has been carried out using (1) objective evaluations and (2) subjective evaluations.

A. OBJECTIVE EVALUATION

A non-reference widely accepted speech distortion measure known as Signal-To-Noise Ratio supported by NIST (NIST-STNR) has been used to measure the level of amplitude distortions present in the prosody modified speech [43], [44]. NIST-STNR estimates the distortion measures based on the histogram energy distribution of clusters in voiced speech regions and noise/silence regions. Compared to PESQ objective measure, NIST-STNR scores quantify the amplitude distortions irrespective of the perceptual quality variations due to scaling of prosodic features. Therefore, for the comparative performance analysis of the prosody modified speech, we prefer to use the NIST-STNR measure over PESQ measure.

For the objective assessment of the pitch and duration modified speech signals, the NIST-STNR measures are computed for various pitch modification factors such as 0.5, 0.8, 1.2 and 2, and clip by sample levels 0.1, 1 and 10 %. Speech utterances, one file each selected from male and female speakers of phonetically balanced CMU-arctic database, are used for objective and subjective assessment. Table 3 provides NIST-STNR values obtained for various duration modification factors and percentages of clipping distortions. Higher values of STNR indicate reduced distortions present in the signal. The appropriate clipping amplitudes were selected from the cumulative amplitude distribution of samples in the given speech signal for introducing required levels of clipping distortions.

In Table 3, STNR values are obtained for different duration scale factors 0.5 (extreme time-down scaling), 0.8 (moderate time down-scaling), 1.2 (moderate time up-scaling) and 2.0 (extreme time up-scaling). For duration modification, irrespective of the scale factors and clip distortion levels, STNR showed significantly lower values for clipped speech as compared to unclipped duration modified speech signals. Variations in the STNR values obtained for the original and clipped version of the same speech files reinforce the results reported by Hansen et al. in [1]. Equivalent variations can be observed while comparing the STNR values obtained for duration modified signals of original and clipped versions. For instance, duration modified speech with 10% clipping distortions showed lower STNR value (54.5 for File1) as compared to its unclipped duration modified signal (52 for

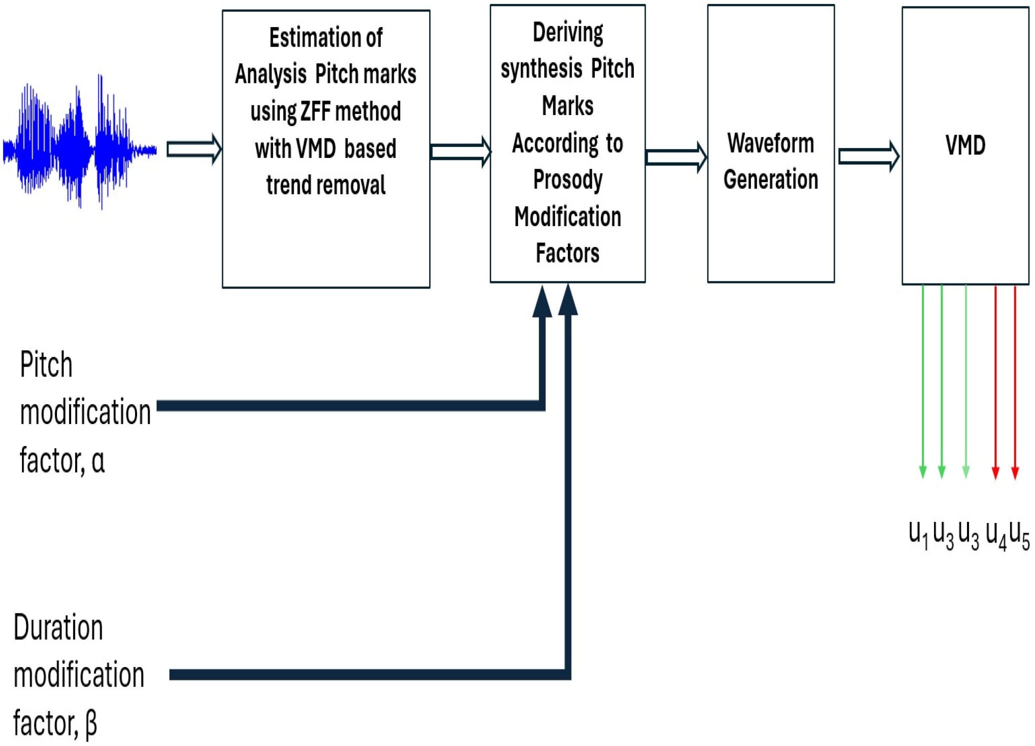
Estimation of Analysis Pitch marks using ZFF method with VMD based trend removal
Deriving synthesis Pitch Marks According to Prosody Modification Factors
Waveform Generation
VMD
Pitch modification factor, α
Duration modification factor, β
$u_1 u_3 u_3 u_4 u_5$

FIGURE 10. Proposed enhancement of prosody modified speech using VMD. The red color lines (u_4 and u_5) indicates the higher VMD IMFs to be discarded and the green colored lines indicate the modes to be retained.

TABLE 3. NIST-STNR measures estimated for duration modified speech for various clipping levels.

	Signal Type	Duration Modification Factors							
		0.5		0.8		1.2		2.0	
Clip. Level 0.1 %		File1	File2	File 1	File2	File1	File 2	File 1	File 2
	Orig.Dur. Mod	54.5	34.5	56.25	38.25	56.75	37.5	57.25	37.25
	Clip. Dur. Mod	54.5	34.5	56.25	36.75	56.75	36.75	56.75	37.25
	Declipped with 3 VMD modes	67.5	52	67.25	48.75	67	50.25	67.5	49
	Declipped with 4 VMD modes	67.75	48.75	68	49.75	68.75	50	69.25	49
	Declipped with 5VMD modes	68.5	47	70.25	47.25	70.25	46.5	70.25	46
	Orig. Signal (File1: 55.25, File2: 36)								
	Clipped Signal (File1: 57.75, File2: 35.5)								
Clip Level 1%		File 1	File 2	File1	File 2	File 1	File 2	File 1	File 2
	Orig.Dur. Mod	54.5	34.5	56.25	38.25	56.75	37.5	57.25	37.25
	Clip. Dur. Mod	54.5	34.25	56	37.5	56.25	37.25	57	36.75
	Declipped with 3 VMD modes	64	26.25	63.25	51.25	65.25	49.25	65.25	50
	Declipped with 4 VMD modes	63.75	51.75	65.5	51.75	65.25	49.5	65.5	49.5
	Declipped with 5VMD modes	65.5	46.5	65.25	47.5	65.25	47.25	65.5	46.5
	Orig. Signal (File1: 55.25, File2: 36)								
	Clipped Signal (File1: 55.5, File2: 35)								
Clip Level 10%		File 1	File 2	File1	File 2	File 1	File 2	File 1	File 2
	Orig.Dur. Mod	54.5	34.5	56.25	38.25	56.75	37.5	57.25	37.25
	Clip. Dur. Mod	52	34	53.5	32.5	53.75	33	54.75	33.5
	Declipped with 3 VMD modes	61.5	43.75	61.5	45.5	62.25	45.5	62.5	46
	Declipped with 4 VMD modes	62	45.75	62.25	46	62	45.5	62.75	45.75
	Declipped with 5VMD modes	62.5	42.75	62	46	62.25	46	62.75	45.25
	Orig. Signal (File1: 55.25, File2: 36)								
	Clipped Signal (File1: 52.75, File2: 32.5)								

File 2). By observing the overall results in Table 3, the proposed VMD based declipping showed improved STNR values under all levels of clipping distortions. According to Table 3 signals reconstructed using 3 VMD modes

TABLE 4. NIST-STNR measures estimated for pitch modified speech for various clipping levels.

	Signal Type	Pitch Modification Factors							
Clip. Level 0.1 %		**0.5**		**0.8**		**1.2**		**2.0**	
		File1	File2	File 1	File2	File1	File 2	File 1	File 2
	Orig.Pitch. Mod	56	37.75	57.5	36.5	55	39.25	57.25	36.5
	Clip. Pitch. Mod	56.25	36.75	57.75	37.75	56.5	37.5	56.75	36.75
	Declipped with 3 VMD modes	28.25	57	68.5	53.5	67.25	48	50.75	44
	Declipped with 4 VMD modes	46.25	59	68	54.5	68.75	39.5	50	33.5
	Declipped with 5VMD modes	35	60	70	45.75	70	46.25	37.25	36.5
	Orig. Signal (File1: 55.25, File2: 36)								
	Clipped Signal (File1: 57.75, File2: 35.5)								
Clip Level 1%		File 1	File 2	File1	File 2	File 1	File 2	File 1	File 2
	Orig.Pitch. Mod	56	37.75	57.5	36.5	55	39.25	57.25	36.5
	Clip. Pitch. Mod	55.75	37	57.75	36.25	56.5	36.5	56.5	36.25
	Declipped with 3 VMD modes	49.5	57.75	66.75	54.5	66.25	48.5	53.5	44.5
	Declipped with 4 VMD modes	30.75	60	68	54.75	68	39.25	38	36.75
	Declipped with 5VMD modes	33.75	60.25	69.25	47.75	69.5	46	49.75	42
	Orig. Signal (File1: 55.25, File2: 36)								
	Clipped Signal (File1: 55.5, File2: 35)								
Clip Level 10%		File 1	File 2	File1	File 2	File 1	File 2	File 1	File 2
	Orig. Pitch. Mod	56	37.75	57.5	36.5	55	39.25	57.25	36.5
	Clip. Pitch. Mod	53.25	34.25	54.75	33.75	54.75	33.75	55	33
	Declipped with 3 VMD modes	42.25	24.75	56.75	49	52.75	43	51.25	28.75
	Declipped with 4 VMD modes	43.25	25.5	56.75	49	50.75	42.75	35.75	39.5
	Declipped with 5VMD modes	32	25.5	56	49.5	52.5	42	50.25	34.25
	Orig. Signal (File1: 55.25, File2: 36)								
	Clipped Signal (File1: 52.75, File2: 32.5)								

obtained slightly better mean STNR for moderate duration modification factors.

Table 4 presents STNR values obtained for pitch modification for the extreme and moderate levels of pitch scaling. In contrast with the duration modification, STNR values obtained for pitch modification showed variations for the male (File1) speaker in particular. Lower STNR values are obtained for proposed VMD based enhancement for extreme levels of pitch modification factors (0.5 and 2). The reason for the reduced STNR is the distortions introduced at the time of manipulation of pitch cycles during the pitch modification. However, for moderate pitch scaling factors, STNR obtained for VMD based signal approximation provided higher values as compared to the clipped pitch modified speech. This trend was observed for both the files consistently. However, the speech file corresponds to the female speaker (File2) showed less variation in STNR values as compared to Male speaker (File 2). The general rise in the pitch contours in females causes reduced dynamic range of clipped samples present in shorter epoch intervals is the reason for the lowered distortions. Where as dynamic range of number of samples per pitch cycle modified is higher as compared to that of female speakers. STNR values obtained from male speakers (File1) for extreme pitch modification (pitch modification factors (2.0)) showed reduced variations as compared to extreme pitch period down scaling factor (0.5). Further, excluding the STNR variation obtained for extreme pitch scale factors, VMD based signal approximation provided improved perceptual quality for moderate levels of pitch scaling.

B. SUBJECTIVE EVALUATION

A perceptual evaluation study has been conducted for assessing the perceptual quality of the pitch and duration modified speech with various level of clipping distortions. The clip by sample level of 0.1% has been omitted as signals with such distortions are perceptually indistinguishable as compared to unclipped original signals. For a comparative analysis with respect to the objective measures, the same files (Male and Female) and prosodic scale factors are used for subjective evaluations as well. People in varied age group of 20-60 years who are aware of various distortions present in speech signal are chosen as the subjects for the perceptual evaluation of pitch and duration modification. Pitch and duration modified files were generated for 4 modification factors (0.5,0.8,1.2 and 2), 3 clipping levels (1% and 10%), 3 VMD approximation signals (using 3,4 and 5 modes). A set of 74 stimuli (2 files × 4 Scale Fact. × 3 clip level (0%,1% and 10 %) × 3 + 2 Orig. Files) were generated each for duration and pitch modification. The subjects were asked to rate each file presented to them in five point scale according to the level of distortions present. Subjects were explained with the relevance of each scale used for rating as given in the Table 5. The process of rating the files was demonstrated through a pilot study to subjects who participated in the subjective evaluation. The files used for pilot study were different from that used in the actual perceptual evaluation. The filenames of each of the generated files were encoded to avoid the chances of getting biased opinions from the subjects towards a particular method or a set of methods.

TABLE 5. Description of the opinion score based ratings used in the perceptual evaluation.

Rating	Description	Description of Perceptual Quality
5	Distortions are perceptually Unnoticeable distortion	Excellent
4	Distortions present but perceptually indistinguishable	Very good
3	Distortions are noticeable & perceptually distinguishable	Good
2	Perceptually Significant level distortions	Fare
1	Annoying levels of distortions present	Poor

TABLE 6. Mean opinion scores obtained for duration modified speech.

	Signal Type	**Pitch Modification Factors**			
Clip. Level 1 %		**0.5**	**0.8**	**1.2**	**2.0**
	Orig.Dur. Mod	3.75 > 97%	4.63 > 99%	4.63 > 99%	4.04 > 95%
	Clip. Dur. Mod	3.6 <90%	4.21>98%	4.40>95%	3.51<90%
	Declipped with 3 VMD modes	3.54<80%	4.58> 99%	4.58>99%	3.67<90%
	Declipped with 4 VMD modes	3.58<80%	4.75>98%	4.29>99%	3.63<80%
	Declipped with 5 VMD modes	3.46<80%	4.5>99%	4.13>99%	3.5<80%
Clip Level 10%		**0.5**	**0.8**	**1.2**	**2**
	Orig.Dur. Mod	3.75 > 97%	4.63 > 99%	4.63 > 99%	4.04 > 95%
	Clip. Dur. Mod	2<80%	2.2<80%	1.7<95%	1.6<80%
	Declipped with 3 VMD modes	1.95>99%	2.38>99%	1.79>99%	1.67<80%
	Declipped with 4 VMD modes	1.71<80%	2.13>99%	1.79>99%	1.71<80%
	Declipped with 5 VMD modes	1.75<80%	1.92>99%	2.04>99%	2.08<90%

TABLE 7. MOS ratings obtained for pitch modfication.

	Signal Type	**Pitch Modification Factors**			
Clip. Level 1 %		**0.5**	**0.8**	**1.2**	**2.0**
	Orig.Pitch. Mod	5>95%	4.5<99%	4.3>99%	2.75<90%
	Clip. Pitch. Mod	4.75<95%	4.51>95%	4.1>99%	2.3<80%
	Declipped with 3 VMD modes	4.75>95%	4.91>99%	4.85>99%	3.31<90%
	Declipped with 4 VMD modes	4.31<90%	4.34>97.5%	4.32>99%	3.32<90%
	Declipped with 5VMD modes	3.8<90%	4.84>99%	4.61>99%	2.84<80%
Clip Level 10%		**0.5**	**0.8**	**1.2**	**2**
	Orig.Pitch Mod	5>95%	4.5>99%	4.3>99%	2.75<90%
	Clip. Pitch. Mod	1.51<80%	1.53<80%	1.25<80%	1.25<80%
	Declipped with 3 VMD modes	1.51<80%	1.53>80%	2.12>95%	1.84<80%
	Declipped with 4 VMD modes	2.1<80%	1.84<95%	1.34<90%	1.25<80%
	Declipped with 5VMD modes	2.01<80%	1.75<95%	1.54<95%	1.75<80%

Table 6 shows the mean opinion score (MOS) ratings obtained for each method by taking the average of opinion scores received. It has to be noted that for moderate modification factors (0.8 and 1.2), VMD based declipping obtained an improved MOS ratings. Confidence intervals computed from the mean and standard deviation of scores obtained from all subjects for a given level of clipping distortion and modification factors are included in the MOS Table. Lower the confidence level indicates higher variance in the ratings provided by the subjects. Even though, there are improvements in MOS ratings for other extreme duration modification scale factors, the statistical confidence of the scores were less than 80% compared to that of 99% confidence levels computed from opinion scores obtained for moderate factors. Among the VMD based declipping, inclusion of 3, 4 & 5 modes were not having much perceptual relevance as the MOS ratings obtained for each of the approximated files provided similar scores with higher standard deviation.

Table 7 shows the MOS ratings obtained for pitch modified files. Similar to perceptual evaluation conducted for duration modification, the same scaling factors were used for pitch modification. Essentially, the original pitch contours were varied from half to double the pitch scale. Similar to the observation made from the perceptual studies of duration modification, the VMD based speech approximation hasn't received any gain in the perceptual quality for extreme scale factors (0.5 and 2.0) as compared to that of clipped speech. However, pitch manipulation by moderate scale factors (0.8 and 1.2) provided improved MOS ratings. The STNR objective measures also supported the relative improvement in MOS ratings for moderate scale factors in pitch modification.

VI. SUMMARY AND CONCLUSION

In the paper, we have presented studies on the effect of clipping on pitch and duration (prosody) modification of speech signals. Accurate estimation of epochs as the analysis pitch marks in epoch based prosody modification is one of the crucial stages in the epoch based prosody modification which is followed by the stage of deriving synthesis pitch marks according to prosodic scale factors. The final stage of prosody modification is the waveform reconstruction to get the prosody modified speech. The study presented in this paper explores how clipping distortions present in the speech signals at various levels affect different stages of prosody modification. The major contribution of the present work is the proposed usefulness of signal reconstruction by using estimated IMFs to improve the perceptual quality of the clipped waveforms. The reconstruction of prosody modified speech was affected by clipping distortions at the moderate level showed improved perceptual quality when reconstructed using VMD modes. The effectiveness of VMD as a declipping algorithm for the task of prosody modification has been confirmed from the improved STNR objective measures and subjective evaluation based MOS ratings obtained for various clipping levels and moderate prosody modification factors.

Since the accuracy of estimated epochs is another crucial factor in determining the perceptual quality of prosody modified speech, the work presented in the paper showed a study on the effect of clipping on deviations of the estimated pitch marks. We have used ZFF based epoch estimation as the analysis pitch mark estimation algorithm for prosody modification. Based on the experimental investigations carried out as a part of this study, the performance of epoch estimation remained intact for moderate levels of clipping distortions present in the signals. This result was expected as the periodicity of the signal remains unaffected even in the presence moderate levels of clipping distortions. However, to further improve the epoch identification accuracy, local mean substraction operation of ZFR output sequence in ZFF method has been replaced by VMD. Due to local mean subtraction over fixed window lengths corresponding to average pitch period, introduced more deviations in the estimated epochs. As an alternative, effectiveness of VMD in capturing non-linear polynomial growth/decay function from the given sequence is exploited to estimate the ZFFS in ZFF method. The estimated epochs using VMD as the trend removal method in ZFR output provides epoch estimation with reduced epoch deviations with respect to the reference epochs estimated from EGG. To Summarize the contributions of the present work:

- Effectiveness of VMD based approximation to improve the perceptual quality of prosody modified speech for moderate scale factors in speech signals with clipping distortions
- VMD based trend removal has been proposed as an alternative to local mean subtraction in conventional ZFF for the estimation of epochs in speech
- Based on the experimental studies, the accuracy of estimated epochs are found to be least affected by clipping distortions present in the original speech

Even though, reconstruction of prosody modified speech using combinations of VMD modes provides better perceptual quality based on STNR and MOS ratings, computational complexities involved when the given clipped speech signal is subjected to VMD iteratively. Similar increase in computational complexity is observed when VMD is used for trend removal for epoch estimation in ZFF. Further, VMD provides promising results as a signal approximation algorithm for moderate pitch and duration modification factors. However, for extreme prosodic scale factors, statistical consistencies of VMD based declipping were observed from STNR and MOS ratings. Similarly, for clipping levels of more than 10% of waveform samples, VMD based trend removal module showed more deviation than the conventional local mean subtraction in ZFF. VMD showed more dependency on fidelity parameters and bandwidth constraints to improve the epoch identification accuracy for the signals with more than 10% of the total samples affected by clipping. From our experimental studies carried out on clipped speech signals for prosody modification, application of VMD as a trend removal in ZFF method should be restricted to only moderate levels of clipping ($\leq$ 10%) for better estimation of epochs. Since VMD is applied iteratively for epoch estimation, the time complexity of the proposed VMD based epoch estimation is on a higher side which is the major limitation of the proposed epoch estimation method. However, the computational complexity of the VMD based clipped speech enhancement is similar to that of the EMD based approach.

REFERENCES

[1] J. H. L. Hansen, A. Stauffer, and W. Xia, "Nonlinear waveform distortion: Assessment and detection of clipping on speech data and systems," *Speech Commun.*, vol. 134, pp. 20–31, Nov. 2021.

[2] M. J. Harvilla and R. M. Stern, "Least squares signal declipping for robust speech recognition," in *Proc. Interspeech*, 2014, pp. 2073–2077.

[3] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbe, "Audio inpainting," *IEEE Trans. Audio, Speech Language Process.*, vol. 20, no. 3, pp. 922–933, Mar. 2012.

[4] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "GACELA: A generative adversarial context encoder for long audio inpainting of music," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 1, pp. 120–131, Jan. 2021.

[5] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 2, pp. 317–330, Apr. 1986.

[6] I. Selesnick, "Least squares with examples in signal processing," Polytech. Inst., New York Univ., Tech. Rep. Accessed: Aug. 31, 2023.

[7] S. Kitic, L. Jacques, N. Madhu, M. P. Hopwood, A. Spriet, and C. De Vleeschouwer, "Consistent iterative hard thresholding for signal declipping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 5939–5943.

[8] P. P. Ebner and A. Eltelt, "Audio inpainting with generative adversarial network," 2020, *arXiv:2003.07704*.

[9] A. Marafioti, N. Holighaus, P. Majdak, and N. Perraudin, "Audio inpainting of music by means of neural networks," 2018, *arXiv:1810.12138*.

[10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2001, pp. 749–752.

[11] D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *Proc. Interspeech*, Aug. 2011, pp. 2969–2972.

[12] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The VoicePrivacy 2020 challenge: Results and findings," *Comput. Speech Lang.*, vol. 74, Jul. 2022, Art. no. 101362.

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

[14] M. R. Rajeswari, D. Govind, S. V. Gangashetty, and A. K. Dubey, "Improved epoch based prosody modification by zero frequency filtering of Gabor filtered telephonic speech," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2023, pp. 1–5.

[15] D. Govind, R. Vishnu, and D. Pravena, "Improved method for epoch estimation in telephonic speech signals using zero frequency filtering," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Oct. 2015, pp. 11–15.

[16] D. Govind, S. Mahanta, and S. R. M. Prasanna, "Significance of duration in the prosodic analysis of assamese," in *Proc. Speech Prosody*, 2012, pp. 494–497.

[17] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proc. INTERSPEECH*, 2006, pp. 1137–1140.

[18] J. P. Cabral, "Transforming prosody and voice quality to generate emotions in speech," M.S. thesis, L2F, Spoken Lang. Syst. Lab, Lisboa, Portugal, 2006.

[19] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, nos. 5–6, pp. 453–467, Dec. 1990.

[20] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, no. 2, pp. 175–205, Feb. 1995.

[21] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[22] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Application of the DYPSA algorithm to segmented time scale modification of speech," in *Proc. EUSIPCO*, Aug. 2008, pp. 1–5.

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

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

[25] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[26] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, Feb. 2014.

[27] R. Sharma, L. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner, and S. R. M. Prasanna, "Empirical mode decomposition for adaptive AM-FM analysis of speech: A review," *Speech Commun.*, vol. 88, pp. 39–64, Apr. 2017.

[28] D. P. Bertsekas, *Constrained Optimization and Lagrangean Multiplier Methods* (Computer Science and Applied Mathematics). Academic, 1982.

[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–22, 2010.

[30] O. Banerjee, D. Govind, A. K. Dubey, and S. V. Gangashetty, "Significance of dimensionality reduction in CNN-based vowel classification from imagined speech using electroencephalogram signals," in *Proc. Int. Conf. Speech Comput. (SPECOM)*, 2022, pp. 44–55.

[31] N. Adiga, D. Govind, and S. R. M. Prasanna, "Significance of epoch identification accuracy for prosody modification," in *Proc. Int. Conf. Signal Process. Commun. (SPCOM)*, Jul. 2014, pp. 1–6.

[32] S. L. Priya and D. Govind, "Significance of epoch identification accuracy for neutral to emotion conversion," in *Proc. Int. Symp. Signal Process. Intell. Recognit. Syst. (SPIR)*, 2018, pp. 1–12.

[33] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[34] K. Sri Rama Murty, B. Yegnanarayana, and M. Anand Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 469–472, Jun. 2009.

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

[36] G. Bapineedu, B. Avinash, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of lombard speech using excitation source information," in *Proc. Interspeech*, 2009, pp. 1091–1094.

[37] K. T. Deepak and S. R. M. Prasanna, "Epoch extraction using zero band filtering from speech signal," *Int. J. Circuits, Syst. Signal Process.*, vol. 34, pp. 2309–2333, Dec. 2015.

[38] S. R. M. Prasanna, D. Govind, K. S. Rao, and B. Yegnanarayana, "Fast prosody modification using instants of significant excitation," in *Proc. Speech Prosody*, May 2010, pp. 1–4, Paper 925.

[39] K. S. S. Srinivas and K. Prahallad, "An FIR implementation of zero frequency filtering of speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 9, pp. 2613–2617, Nov. 2012.

[40] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, USA, 2004, pp. 223–224.

[41] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[42] P. Singh and G. Pradhan, "Variational mode decomposition based ECG denoising using non-local means and wavelet domain filtering," *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 4, pp. 891–904, Dec. 2018.

[43] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. Int. Conf. Acoustic, Speech Signal Process. (ICASSP)*, 1995, pp. 153–156.

[44] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, 2008, pp. 2598–2601.

M. RAMA RAJESWARI received the B.Sc. degree from Andhra University, in 2001, the M.C.A. degree from Indira Gandhi National Open University (IGNOU), New Delhi, in 2008, the M.Tech. degree in engineering from Jawaharlal Nehru Technological University Kakinada (JNTUK), Andhra Pradesh, India, in 2014, and the AMIE degree from the Institute of Engineers (IEI), Kolkata, India, in 2018. She is currently pursuing the Ph.D. degree in speech processing with the Department of Computer Science and Engineering, Koneru Lakshmaih University (KL University), Andhra Pradesh.

D. GOVIND was born in Kerala, in 1983. He received the bachelor's degree in technology from the Government College of Engineering Kannur, Kerala, in 2004, the master's degree in technology with a specialization in computer vision and image processing from Amrita Vishwa Vidyapeetham University, Coimbatore, Tamil Nadu, in 2007, and the Ph.D. degree in speech processing from Indian Institute of Technology Guwahati, in 2013. He is currently a Professor with the Department of Computer Science and Engineering, K. L. University, Andhra Pradesh.

SURYAKANTH V. GANGASHETTY (Member, IEEE) received the Ph.D. degree in neural network models for recognition of consonant-vowel units of speech in multiple languages from IIT Madras, in 2005. He is currently a Faculty Member with K. L. University, Guntur, Andhra Pradesh, India. Before joining K. L. University, he was a Faculty Member with IIIT Hyderabad, Telangana, from 2006 to 2020. Previously, he was a Senior Project Officer with the Speech and Vision Laboratory, IIT Madras. He was a Faculty Member at BIET Davangere, Karnataka, from 1991 to 1999. He was also a Visiting Research Scholar with Oregon Graduate Institute (OGI), Portland, OR, USA, for three months, in Summer 2001. He has done his Postdoctoral studies (PDF) with Carnegie Mellon University (CMU), Pittsburgh, PA, USA, from April 2007 to July 2008. He is the author of about 150 papers published in national and international journals, conferences, and edited volumes. His research interests include speech processing, neural networks, machine learning, natural language processing, and artificial intelligence. He is a Life Member of the CSI, IE, IUPRAI, ASI, IETE, ORSI, and ISTE. He has reviewed papers for reputed journals and conferences. He was the Local Organizing Chair for the INTERSPEECH-2018 Conference, Hyderabad, India, in September 2018.

AKHILESH KUMAR DUBEY received the Ph.D. degree from Indian Institute of Technology, Guwahati. In July 2021, he joined the Department of Computer Science and Engineering, K. L. University, Andhra Pradesh, as an Associate Professor. His research publications are in reputed conferences and journals. His publications include prestigious conferences in the area of speech processing, such as INTERSPEECH organized by the International Speech Communication Association, and journals, such as *Journal of the Acoustical Society of America* and *Speech Communication*. His research interest includes speech signal processing.