This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Foreground Speech Segmentation and Enhancement using Glottal Closure Instants and Mel Cepstral Coefficients

K. T. Deepak and S. R. M. Prasanna, *Senior Member, IEEE*

***Abstract*—In this work, the speech signal recorded from the desired speaker close to microphone in natural environment is regarded as *foreground speech* and rest of the interfering sources as *background noise*. The proposed work exploits speech production features like Glottal Closure Instants (GCIs) in time domain and vocal tract information in spectral domain to segment the desired speaker's speech and to further enhance it. The foreground speech is perceptually enhanced using the auditory perception feature in mel frequency domain using mel cepstral coefficients (MCC) and its inversion using mel log spectrum approximation (MLSA) filter. The focus is on enhancing the production and perceptual features of foreground speech rather than relying on modeling the interfering sources. The speech data is collected in different natural environments from different speakers in order to evaluate the proposed method. The enhanced speech signals derived at three different stages of the proposed method are evaluated with state-of-the-art methods in terms of subjective and objective measures. The proposed method provides improved performance compared to the considered state-of-the-art methods. In terms of the proposed objective measure *Foreground to Background Ratio (FBR)*, the enhancement approach presented in this work gives an average improvement of 12 dB as opposed to existing spectral subtraction based method which provides 3 dB. Moreover, subjective evaluation using 24 different subjects corroborates the objective test results.**

***Index Terms*—Foreground segmentation, zero band filter (ZBF), glottal closure instants (GCI), formant peaks, MCC, MLSA, speech enhancement.**

I. Introduction

In most of the cases, the speech signal recorded in natural environment is degraded by other interfering acoustic sources. The natural environment typically refers to an office or laboratory environment with relatively high acoustic background noise like call centers. The degradation can be of different levels and consists of one or more interfering sources. Speech enhancement is still a challenging task when signal is recorded using a single sensor in natural environment [1]. The task is complicated if the interfering sources are from other background speakers. The perceptual quality improvement in terms of intelligibility and reduction in the background noise to enable comfortable speech communication is of paramount importance [2]. In such cases, it is of interest to segment the desired speaker's speech from rest of the interfering background noise first and then enhance the desired speaker's speech.

K. T. Deepak and S. R. M. Prasanna are with the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India e-mail: ({deepakkt, prasanna}@iitg.ernet.in).

The general approach of most methods is to model the additive noise only components from noisy speech signal and suppress them to obtain the enhanced speech [3], [4]. The noise modeling depends on efficient segmentation of additive background noise from rest of the speech regions. The suppression of noise usually takes place in spectral domain through subtraction. However, such spectral subtraction methods introduce distortions in enhanced speech because of overestimating the noise spectrum in the form of undesired musical tones. Many methods were proposed in order to overcome this problem [5], [6]. The human auditory perceptual cues were considered to improvise the spectral subtraction methods [7], [8]. The objective is not to completely suppress the additive noise. Rather, the purpose is to utilize the masking properties of human auditory system and live with certain amount of residual noise that are below the masking threshold. The advantage of such algorithms over conventional methods is that, they do not introduce spectral distortions in the form of audible musical noise due to over subtraction. However, utilizing complex auditory phenomena like psychoacoustic models for the sake of speech enhancement can make such methods to be more complex.

There is some evidence that humans perceive speech by capturing the features from high signal to noise ratio (SNR) regions and extrapolate those features to low SNR regions, both in temporal and spectral domains [9]. The idea is to enhance the high SNR regions further relative to low SNR regions for enhancing the noisy speech signal. This is equivalent to the phenomenon of Lombard effect, where, speaker emphasizes the production apparatus to increase SNR of the produced speech when background noise in feedback path increases [10]. There are some methods proposed in literature that basically utilize the high SNR regions like instants of significant excitation of speech signal in the temporal domain to enhance speech [11]. The idea is to emphasize instants of significant excitation which are predominantly glottal closure instants (GCIs) relative to other regions of LP residual signal. The GCIs are located using Hilbert envelope of LP residual (HELP), using which, a temporal weighting function is derived so to enhance GCI positions relative to other regions. The synthesized speech signal from such temporal enhancement may not completely eliminate the background noise, nevertheless, the distortion caused by such enhancement methods are minimal. The recently proposed method in [12] utilizes temporal enhancement as the preliminary stage of enhancement and subsequently uses spectral enhancement to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

eliminate the remaining residual noise. The advantage of this method is that it causes lesser spectral distortion.

There is evidence that when speech signal is recorded in natural environment, the speech production characteristics tend to vary depending on the levels of interfering sources [13]. Therefore, speech enhancement techniques may have to focus on exploiting such unique characteristics of speech signal. The present work focuses on practical scenario when the speech signal is recorded in different natural environments. In a typical recording scenario (head mounted microphone or mobile phones held close to ears), the desired speaker is closer to microphone relative to other interfering sources. In this work it is assumed that the proximity of desired speaker is closest to microphone compared to the distance between other sound sources and the microphone. The speech signal so recorded from desired speaker is termed as *foreground speech* and rest of the interfering sources are termed as *background noise*. Due to close proximity of speaker to microphone and also the modified speech production characteristics due to acoustic feedback, there are significant differences in the nature of signal for foreground speech and the background noise [14], [15]. The core idea being that not all foreground speech regions are affected equally by interfering background noise. In particular, the instants of significant excitation in temporal domain and formant locations in the spectral domain remain robust to background noise. Hence, such regions form high SNR regions of speech signal that can be utilized for speech enhancement. Furthermore, the proposed method utilizes the subtle aspects of human auditory feature to enhance the foreground speech. The merits of proposed method lies in exploiting the speech production features such as excitation source and vocal tract information along with auditory perception feature and study the benefits of each in terms of speech enhancement.

The rest of the paper is organized as follows: Section II illustrates the overall block diagram of the proposed method and explains foreground segmentation in detail. The Sections II-A and II-B describe excitation source based features and vocal tract articulatory feature, while, Section II-C explains the combination of features for foreground speech segmentation. Section III-A describes the details of excitation source based enhancement, while, Section III-B explains the details of formant based enhancement using a block diagram. The Section III-C explains the details of analysis and synthesis sub modules using mel cepstral coefficients (MCC) and mel log spectral approximation (MLSA) filter, respectively. Section IV describes the details of the evaluation procedure and results obtained in terms of subjective and objective evaluation scores. The summary and conclusions of the present work, and the scope for future work are mentioned in Section V.

II. Foreground Speech Segmentation

The proposed work mainly consists of foreground speech segmentation and multistage foreground speech enhancement modules. If $s(n)$ is the speech signal recorded in foreground scenario from a natural environment, then there can be other interfering sources along with foreground speech signal. The objective of the current work is to first temporally segment foreground speech regions from rest of the background noise and then enhance foreground speech regions. The overall block diagram of the proposed work is shown in Fig. 1. The details of foreground speech segmentation module is explained in this Section. The acoustic background noise picked up by the recording microphone can be speech like. It is therefore difficult to distinguish the desired foreground speech from rest of the background noise. However, it is important to temporally segment foreground speech regions from rest of the background noise for subsequent processing to enhance foreground regions.

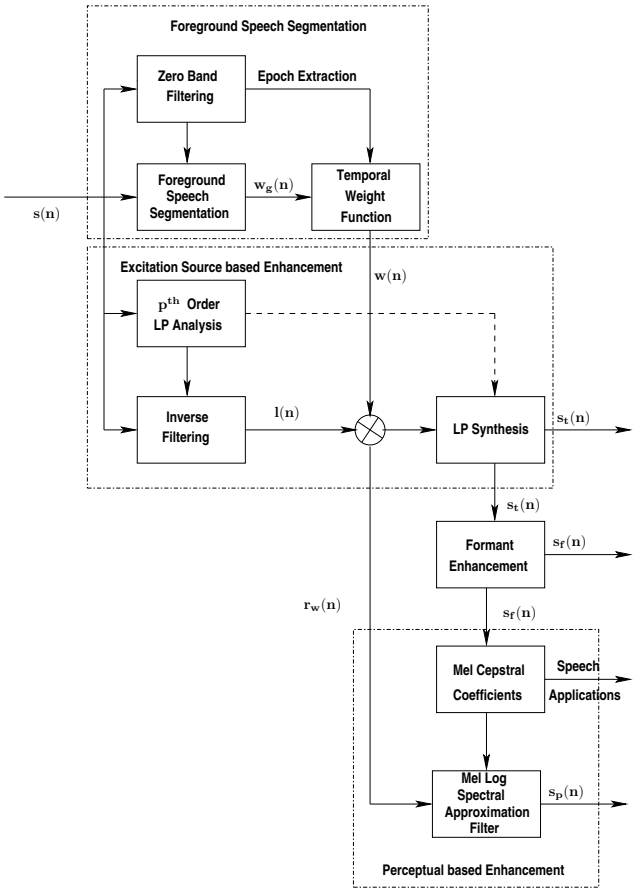

Fig. 1: The overall block diagram of the proposed foreground speech segmentation and enhancement method, where, $s(n)$ is the input speech signal recorded in foreground scenario, $w_g(n)$ is the gross weight function that mainly segments the foreground speech regions from rest of the background noise, $w(n)$ is the final temporal weight function, $l(n)$ is the LP residual signal, $r_w(n)$ is the temporally weighted LP residual signal, $s_t(n)$ is excitation based enhanced output, $s_f(n)$ is formant based enhanced output and $s_p(n)$ is perceptually enhanced output.

A. Excitation Source based Features

Due to the close proximity of foreground speaker to microphone relative to other acoustic sources there are significant differences in the signal characteristics of foreground speech and background noise [15]. Most of the voice activity detection methods developed either in temporal or spectral domains have to deal with all frequency components of speech signal. Alternatively, a method is developed in [15] that analyses foreground speech and background noise at zero frequency using zero frequency filter (ZFF) [16]. The ZFF allows signal components around zero frequency and significantly attenuates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

rest of the frequency components. The output of ZFF is called as zero frequency filtered signal (ZFFS) using which two temporal domain features *viz.*, normalized first order auto-correlation coefficient (NACC) and ZFFS energy are derived to segment foreground speech from rest of the background content. However, ZFF is a marginally stable filter and the output of filter grows or decays exponentially and such a filter requires remove trend procedure to obtain ZFFS. The trend is removed by subtracting the average over 1 to 2 pitch period from each sample of ZFF output. In a method proposed [17] recently, the stable realization of resonator filter called zero band filter (ZBF) is used to obtain zero band filtered signal (ZBFS), similar to ZFFS. The advantage of using ZBF is that, the output of the filter is converging type and does not require F_0 estimation.

Let $s(n)$ be the foreground speech signal recorded in natural environment that can have other interfering sources. The first order difference of this signal is given by

$$d(n) = s(n) - s(n-1) \tag{1}$$

The difference signal $d(n)$ is to deemphasize low frequency fluctuations and emphasize high frequency components present in signal and is passed twice through the transfer functions as given in [17]

$$H_1(z) = \frac{Y_1(z)}{D(z)} = \frac{1}{1 - 2rz^{-1} + r^2z^{-2}} \tag{2}$$

$$H_2(z) = \frac{Y_2(z)}{Y_1(z)} = \frac{1}{1 - 2rz^{-1} + r^2z^{-2}} \tag{3}$$

where, r represents the value of radius on unit circle in z-plane at which the poles are placed, and r value should satisfy $0 < r < 1$ for stability and its value is taken as 0.99 in ZBF. In time domain, the filter outputs can be represented in the form of summation series given by

$$y_1(n) = \sum_{k=0}^{\infty}(k+1)r^k d(n-k) \tag{4}$$

$$y_2(n) = \sum_{k=0}^{\infty}(k+1)r^k y_1(n-k) \tag{5}$$

The output $y_2(n)$ is filtered using a Butterworth 4^{th} order high pass filter having a cutoff frequency 80 Hz in order to attenuate low frequency fluctuations. The nature of the resulting ZBFS is different for background acoustic sources compared to foreground speech. It can be noted from [15] that, the nature of ZBFS is nearly periodic and has higher amplitude levels for foreground speech, while, ZBFS appears to be dispersed in case of background speech and other acoustic sources. Also, the amplitude levels of ZBFS is significantly low in background noise regions compared to foreground speech regions. This can be attributed to the fact that, the proximity of foreground source is closer to microphone than other acoustic sources. This is illustrated in Fig. 2, where, Figs. 2(a) and (b) show 50 ms segments of foreground speech and background speech chosen from same recording, respectively. The signal is recorded at 16 kHz sampling rate using headphone microphone connected to laptop in a living room. The speaker is closer to microphone while talking and a television is playing speech at the background. Since, the signal is recorded in natural environment using single microphone sensor there is no reference clean speech signal available to measure *a priori* signal to noise ratio (SNR). Hence, a similar measurement can be made using foreground to background ratio (FBR) that essentially computes the ratio of normalized foreground speech to background power and it is expressed in decibels (dB). The detailed description of FBR is given in Section IV-B3. The segments shown in Figs. 2(a) and (b) are taken from signal having FBR 6.27 dB. The Figs. 2(e) and (f) illustrate the corresponding ZBFS, while, Figs. 2(i) and (j) show normalized autocorrelation sequence computed from ZBFS, respectively. The slope value measured at positive zero crossings of ZBFS indicates the strength of excitation (SoE) [18]. The Figs. 2(m) and (n) show SoE computed from ZBFS segments shown in Figs. 2(e) and (f), respectively. The Figs. 2(q) and (r) illustrate ZBFS normalized energy computed using the segments shown in Figs. 2(e) and (f), respectively. The ZBFS normalized energy is computed using a frame size of 5 ms with one sample shift. However, for illustration purpose the ZBFS energy is sampled at glottal closure instants to compare them with SoE. It can be noticed that, the signal contours of SoE and ZBFS energy follows similar trend and hence offers the same discriminative abilities to distinguish foreground and background speech regions. The similar trend of these two signals can be attributed to the fact that ZBFS energy depends on SoE and hence they are related. The Fig. 2(c) shows 50 ms segment of background music with vocals chosen from signal having FBR 5.21 dB. The signal is recorded from a foreground speaker when television is playing background music with vocals at the background. The corresponding ZBFS, normalized autocorrelation sequence computed from ZBFS, SoE derived using ZBFS and ZBFS normalized energy plots are shown in Figs. 2(g), (k), (o) and (s), respectively. Further, Fig. 2(d) shows 50 ms segment of background noise chosen from another recording. The speech signal is recorded in foreground scenario from a speaker in office environment when mosaic polishing machine is operating at the background. The segment is chosen from signal having FBR 7.38 dB. The corresponding ZBFS, normalized autocorrelation sequence computed from ZBFS, SoE derived using ZBFS and ZBFS normalized energy plots are shown in Figs. 2(h), (l), (p) and (t), respectively.

It can be observed that, the ZBFS derived from foreground speech region appears to be nearly periodic, while, it appears distorted for background speech, music, and noise regions. This fact is further illustrated using normalized autocorrelation sequence plots. It can be noticed that the value of first largest peak (excluding center) normalized with respect to the center value in autocorrelation sequence defined as normalized auto-correlation coefficient (NACC) is larger in case of foreground speech regions compared to background regions. The reason for such a change in ZBFS characteristics can be ascribed to the fact that, foreground source is closer to microphone sensor and hence the excitation source is least affected by interfering sources. The amplitude level of ZBFS is directly related to epoch strength of excitation source and hence energy at foreground speech region is relatively larger compared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

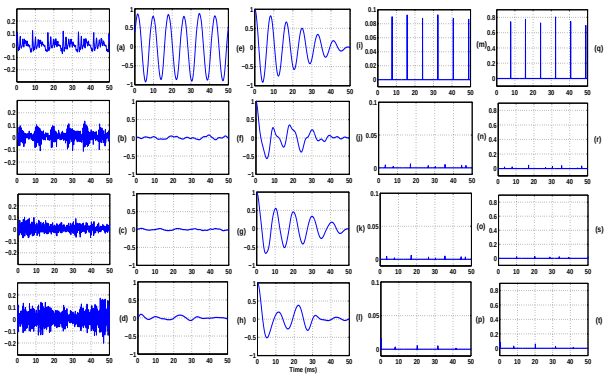

Fig. 2: 50 milliseconds of (a) foreground speech, (b) background speech, (c) background music with vocals, and (d) background noise. Respective, ZBFS ((e) - (h)), normalized autocorrelation sequence using ZBFS ((i) - (l)), strength of excitation computed using ZBFS ((m) - (p)), and ZBFS energy sampled at glottal closure instants ((q) - (t)).

to background regions. Hence, such discriminative features derived from ZBFS can be used to segment foreground speech from rest of the background noise that may include speech like sources. This is further illustrated by choosing a full sentence spoken by foreground speaker in Fig. 3. The Fig. 3(a) shows speech signal recorded from a male speaker in office, while, mosaic polishing machine is operating in background, the FBR value is 7.38 dB. The signal is recorded using a headphone microphone connected to a laptop for recording. It can be observed that there is significant amount of noise present in background. The NACC and ZBFS energy computed from signal is shown in Figs. 3(b) and (c), respectively. The two features derived from ZBFS offers discriminative information between foreground and background regions. The NACC and ZBFS energy are relatively higher in case of foreground speech regions relative to background noise regions. Hence, theses two features can be used to segment foreground speech from recorded signal.

B. Vocal Tract Articulatory Feature

The features discussed so far are derived using excitation source information and does not make use of vocal tract articulatory gestures that are unique to foreground speech. One way of capturing vocal tract information is by measuring the spectral envelope. However, it is suggested that, the temporal dynamics of spectral envelopes as more reliable means for carrying the linguistic context of speech message and can be obtained through modulation spectrum energy of speech signal. The temporal envelope of speech is dominated by low frequency components that are in similar range to the dynamics of speech production, in which, the articulators move at such slow rate [19]. It is studied that, the linguistic information of speech signal lies in the range of 2 to 16 Hz and centered at 4 Hz. Subsequently, filtering slow and fast varying trajectories of spectral envelopes can be useful in alleviating the effects of interfering background sources. Hence, modulation spectrum energy can be used as a feature to segment foreground speech from rest of the background noise. In this work, the modulation spectrum energy is extracted from signal as given in [20]. However, the signal is divided into 18 subbands using compressive gammachirp auditory filter (cGC) [21], [22]. The auditory filter designed is level dependent and nonlinear, this emulates psychophysical data on masking and two tone suppression. The cGC consists of two filters *viz.*, a passive gammachirp filter (pGC) and a dynamic filter which is asymmetric function that shifts in frequency with stimulus level. The cGC filter is realized using the following relationship

$$\begin{aligned} g_c(t) = at^{n_1-1}exp(-2\pi b_1 ERB_N(f_{r1})t) \\ \times exp(j2\pi f_{r1}t + jc_1 \ln t + j\phi_1) \end{aligned} \quad (6)$$

where, a is amplitude; n_1 and b_1 are parameters defining the envelope of gamma distribution; c_1 is chirp factor; f_{r1} is frequency referred to as asymptotic frequency, since the instantaneous frequency of carrier converses to it when t is infinity; $\mathrm{ERB_N}(f_{r1})$ is equivalent rectangular bandwidth of average normal hearing subjects; ϕ_1 is initial phase; and $\ln t$ is natural logarithm of time. The Fourier magnitude spectrum of cGC is given by

$$|G_c(f)| = a_\Gamma . |G_T(f)| . \exp(c_1\theta_1(f)) \quad (7)$$

$$|\theta_1(f)| = \arctan\left(\frac{f - f_{r1}}{b_1 \mathrm{ERB_N} f_{r1}}\right) \quad (8)$$

where, $|G_T(f)|$ is the Fourier magnitude spectrum of gammatone filter; $\exp(c_1\theta_1(f))$ is an asymmetric function because θ_1 is an antisymmetric function centered at asymptotic frequency f_{r1} and a_Γ is constant. Further the asymmetric function $\exp(c_1\theta_1(f))$ is decomposed into lowpass and highpass asymmetric filter functions to represent passive and dynamic components separately. The resulting compressive cGC filter $|G_{cc}(f)|$ is

$$|G_{cc}(f)| = [a_\Gamma . |G_T(f)| . \exp(c_1\theta_1(f))] . \exp(c_2\theta_2(f)) \quad (9)$$

$$|G_{cc}(f)| = |G_{cp}(f)| . \exp(c_2\theta_2(f)) \quad (10)$$

The compressive gammachirp is composed of a level independent passive gammachirp filter (pGC) $G_{cp}(f)$ which represents passive basilar membrane and a level dependent highpass asymmetric function which simulates active component in the cochlea. The Amplitude envelope is computed from each individual subband outputs obtained from filter bank. Each of the filter's output is first halfwave rectified and then subjected through a lowpass filter having a cutoff frequency 28 Hz. The amplitude envelope obtained is downsampled by a factor of 100 and normalized by average value obtained from the respective filter bank output. The modulations of the normalized envelope signals obtained are further analyzed using Discrete Fourier Transform (DFT). The DFT is computed using 250 ms Hamming window with shift of 12.5 ms, this essentially captures dynamic properties of the signal. However, 2-16 Hz components from each such channel are summed together to obtain the modulation spectrum energy. The modulation spectrum energy can be computed using the following relationship

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

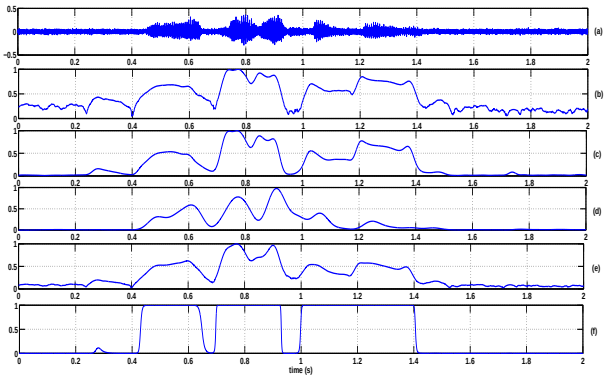

Fig. 3: The figure illustrates the foreground segmentation using the combined evidence from different features (a) speech signal recorded in foreground scenario, (b) normalized first order autocorrelation coefficients derived from ZBFS, (c) short term energy derived from ZBFS, (d) modulation spectrum energy derived from speech signal, (e) combined evidence, and (f) combined evidence passed through sigmoidal function to segment the foreground speech regions from rest of the background noise.

$$m(i) = \sum_{p=1}^{18} \sum_{k=k1}^{k=k2} \mid \hat{S}_p(k,i) \mid^2 \tag{11}$$

where, i is frame index, p represents critical band filter and $k1$, $k2$ represents frequency index of 4 Hz and 16 Hz, respectively. $\hat{S}_p(k,i)$ is obtained using the following relationship

$$\hat{S}_p(k,i) = \sum_{n=0}^{N-1} \hat{s}_p(n + i \times F) w(n) e^{-j2\pi nk/N} \tag{12}$$

where, $\hat{s}_p(n)$ represents normalized envelope of p^{th} filter output, F is frame shift, $w(n)$ is Hamming window, and N is number of points used for computing DFT. The modulation spectrum energy computed for each frame are upsampled to $16000\ samples/s$. The modulation spectrum energy computed from the signal shown in Fig. 3(a) is illustrated in Fig. 3(d). It can be noticed that, the modulation spectrum energy values are significantly higher in foreground speech regions compared to background noise regions. This can be attributed to the fact that vocal tract articulatory gestural movements operate in an exclusive frequency range of 2-16 Hz compared to other acoustic sources. Hence, the modulation spectrum energy can be used to distinguish foreground speech regions from rest of the background content.

C. Combined Evidence

The excitation source based features such as NACC and ZBFS energy are combined with vocal tract based modulation spectrum energy in order to identify the foreground speech regions. In order to illustrate the concepts of NACC, ZBFS energy, and modulation spectrum energy, features are computed with 1 sample shift and each such signal is amplitude normalized. Let us denote such normalized sequence of NACC, ZBFS energy, and modulation spectrum energy as $N_c(n)$, $Z_E(n)$, and $M_E(n)$, respectively, where, n represents the number of samples in a given signal. The features are temporally added and normalized with respect to maximum value of the added sequence to obtain the combined evidence given by $E(n) = (N_c(n) + Z_E(n) + M_E(n))/\max(N_c(n) + Z_E(n) + M_E(n))$. The combined evidence of all three features is shown in Fig. 3(e). However, it is difficult to set the threshold directly on such signal for foreground segmentation. Alternatively, the gross level feature is obtained by passing $E(n)$ through sigmoidal function given by

$$w_g(n) = (1 - w_{gm}) \frac{1}{1 + \exp\left(-\lambda(E(n) - T_h)\right)} + w_{gm} \tag{13}$$

where, $w_g(n)$ is sigmoidal function of $E(n)$, λ is slope parameter set to 20, T_h is threshold derived from mean value of the signal $E(n)$ and w_{gm} is minimum value of sigmoidal function which is set to 0 in this case. The $w_g(n)$ function forms gross level feature that mainly helps to segment foreground speech in the presence of background noise. This is illustrated using Fig. 3(f), where, the foreground regions are further enhanced relative to other background regions. The weight function $w_g(n)$ derived from three features can be used as gross level feature to temporally enhance the foreground speech.

III. Foreground Speech Enhancement

The foreground speech enhancement is carried using multiple stages. The proposed enhancement scheme is explained using the block diagram shown in Fig. 1, classified into four major modules *viz.*, foreground speech segmentation, excitation source based enhancement, formant based enhancement and perceptual based enhancement. The foreground speech segmentation is discussed in Section II. The enhanced speech signal obtained from different modules are named as excitation source based enhancement ($s_t(n)$), formant based enhancement ($s_f(n)$), and perceptual based enhancement ($s_p(n)$). It can be observed that all three modules are connected sequentially.

A. Excitation Source based Foreground Speech Enhancement

The excitation source information, especially instants of significant excitation are high SNR regions which can be used as anchor points to enhance the foreground speech regions. A robust method is required to locate instants of significant excitation even when speech signal is recorded in noisy environments. It is shown in [17] that, ZBF is robust in identifying the locations of GCIs in foreground scenarios and there is no requirement for estimation of F_0. The positive zero crossings of ZBFS derived from Eqn. (5) precisely match with the locations of GCIs. In order to modify excitation source signal using GCI locations, it is beneficial to resolve speech signal in terms of source and filter components.

In linear prediction (LP) analysis, the vocal tract system can be modeled as a time varying all pole filter using frame based analysis. The LP analysis works on the principle that, the current speech sample can be predicted from past p samples, where, p is called linear prediction order and is chosen as $F_s/1000 + 4$ (F_s is sampling frequency). In order to compute LP coefficients, the frame size is selected to be 20 ms with a frame shift of 10 ms. If $s(n)$ denotes speech signal recorded

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

in foreground scenario, then the predicted sample at the time instant n is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{14}$$

where, a_k is the set of LP coefficients predicted. The residual error is the difference between actual sample sequence $s(n)$ and predicted sample sequence $\hat{s}(n)$ and it is given by the following relationship

$$l(n) = s(n) - \hat{s}(n) \tag{15}$$

From Eqns. (14) and (15), the residual signal $l(n)$ can be written in z-domain transfer function as

$$L(z) = S(z) + \sum_{k=1}^{p} a_k S(z) z^{-k} \tag{16}$$

i.e.,

$$A(z) = \frac{L(z)}{S(z)} = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{17}$$

where, the LP residual signal can be obtained by filtering speech signal $S(z)$ through the filter $A(z)$, which is generally called as inverse filtering. The prediction error is relatively high at GCI locations compared to other regions of speech signal. Hence, the amplitude level of LP residual signal is higher at GCI locations compared to other regions of LP residual signal. The LP residual signal is uncorrelated and therefore any modification of such residual signal introduces least distortion for later synthesis of speech signal. It is shown in [11] that, the enhanced speech can be synthesized by modifying LP residual signal without much audible distortion. The LP residual signal is derived from noisy speech signal and modified by retaining 2 ms regions around GCIs. The method proposed in [12] uses similar approach by utilizing GCIs as anchor points obtained from HELP. The identification rate (IDR) and accuracy (IDA) of locating GCIs using HE of LP residual is inferior in case of noisy speech signal compared to recently proposed methods [23]. Alternatively, in the proposed method, ZBF is used to locate GCIs directly from speech signal recorded in natural environment.

This is illustrated in Fig. 4, where, Fig. 4(a) shows foreground speech signal recorded when music is playing in the background. It can be noticed that there is significant amount of background music present along with foreground speech signal. The foreground speech regions are shown in Fig. 4(a) as dotted lines and the corresponding ZBFS can be obtained by passing speech signal through Eqns. (2) and (3) as shown in Fig. 4(b). The positive zero crossings of ZBFS corresponds to GCI locations of speech signal. The region around GCI locations can be used as anchor points to modify the LP residual signal. The LP residual signal derived from speech signal is shown in Fig. 4(e). In order to emphasize the regions around GCI locations, a fine weight function is obtained similar to [12]. The GCI location is convolved with Hamming window function $h_w(n)$ having a temporal duration of 3 ms that closely corresponds to closed phase interval of a glottal

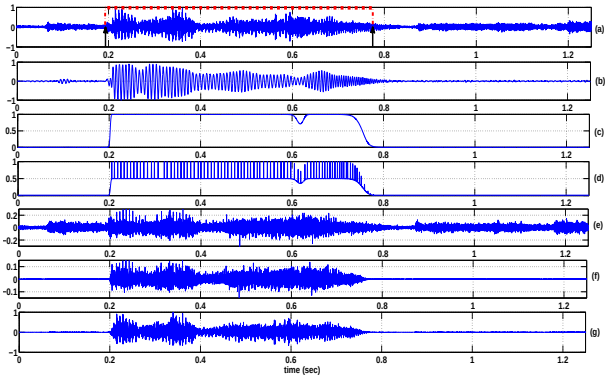

Fig. 4: Illustration of excitation based enhancement. (a) speech signal recorded in foreground scenario (dotted lines and arrows indicate the foreground region), (b) the positive zero crossings of ZBFS indicate the epoch locations, (c) the foreground weight function $w_g(n)$, (d) the temporal weight function $w(n)$ by combining the evidence of epoch locations with foreground segmentation, (e) LP residual signal derived from speech signal, (f) modified LP residual signal $r_w(n)$ (g) excitation based enhanced speech signal synthesized.

cycle. If GCIs are considered as shifted train of impulses, then the fine weight function $w_f(n)$ is given by

$$w_f(n) = \left(\sum_{k=1}^{N_k} \delta(n - i_k) \right) * h_w(n) \tag{18}$$

where, N_k is total number of GCIs located, i_k is estimated location of GCI. The minimum value of $w_f(n)$ is set to a threshold value of T in order to keep the distortion low because of overemphasizing GCI locations in LP residual and the relationship is expressed as

$$w_f(n) = \begin{cases} T, & \text{if } w_f(n) < T \\ w_f(n), & \text{otherwise} \end{cases} \tag{19}$$

where, T is set to 0.5 in this work. It can be noted that, temporal processing is not sensitive to a range of T values [12]. The final weight function $w(n)$ is obtained by multiplying gross weight function $w_g(n)$ as shown in Fig. 4(d) with fine weight function $w_f(n)$ and it is expressed as

$$w(n) = w_g(n) \times w_f(n) \tag{20}$$

The normalized final weight function $w(n)$ is multiplied to residual signal $l(n)$ to obtain the weighted LP residual signal (WLPR) $r_w(n)$ as shown in Fig. 4(f). The temporally enhanced speech signal $s_t(n)$ can be synthesized by the transfer function given in z-domain as

$$S_t(z) = \frac{R_w(z)}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{21}$$

where, $S_t(z)$ is temporally enhanced speech signal and $R_w(z)$ is WLPR in z-domain, while, a_k are the LP filter coefficients. The temporally enhanced speech signal is shown in Fig. 4(g). Temporally there is overall reduction in the background noise after temporal enhancement. This is further illustrated using Fig. 5, where, Fig. 5(a) shows foreground speech signal recorded while music is being played at the background and Fig. 5(e) shows the corresponding narrowband spectrogram. There is significant amount of background noise present throughout the temporal duration of speech recording. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

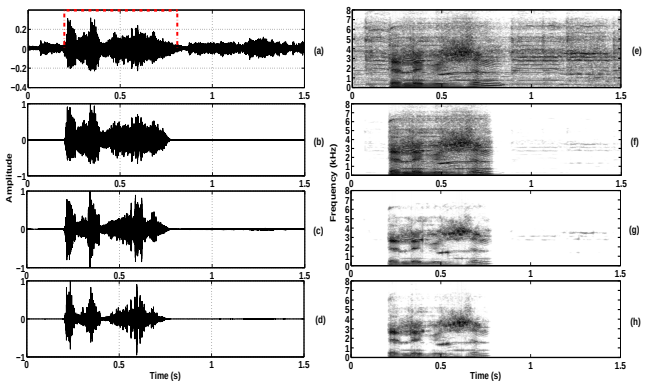

Fig. 5: Illustration of enhancement outputs obtained at different stages and their narrowband spectrogram plots. (a) speech signal recorded in foreground scenario, (b) excitation based enhanced output, (c) formant based enhanced output, (d) perceptual based enhanced output, ((e) - (h)) are corresponding narrowband spectrograms.

narrowband spectrogram shown in Fig. 5(f) obtained from temporally enhanced foreground speech signal and this illustrates that there is significant reduction of background noise. Although, there is reduction of background noise in foreground regions, still there is audible background noise present in the foreground regions and it is evident from the spectrogram plots. Hence, excitation source based enhancement alone may not be sufficient to suppress background noise present in foreground speech regions.

B. Formant based Foreground Speech Enhancement

The Sections II and III-A helped to temporally segment foreground speech regions from rest of the background noise and further enhance the foreground speech using excitation source information. However, the foreground speech enhancement $s_t(n)$ using excitation source information is still left with some residual background noise. Considering the fact that, instants of significant excitation are high SNR regions in temporal domain, similarly, the formant peak locations are high SNR regions in spectral domain. The vocal tract information is intact in most of the foreground recording scenarios and hence such information can be exploited to enhance foreground speech regions further. The formant peak enhancement relative to spectral valleys have been exploited in many methods for speech enhancement [24]. In proposed method, the 1^{st} stage of excitation based enhanced output $s_t(n)$ is further subjected through formant based enhancement (FBE). The formant enhancement is carried on LP spectrum which helps to enhance formant locations relative to adjacent valleys. The formant enhancement is shown in the form of block diagram in Fig. 6. $A_t(z)$ is the 1^{st} order LP inverse filter expressed as

$$A_t(z) = 1 + bz^{-1} \tag{22}$$

and $H_{vt}(z)$ is LP filter predicted from p^{th} order LP analysis, where, p is chosen as $F_s/1000 + 4$ [25]

$$H_{vt}(z) = \frac{1}{1 + \sum_{k=1}^{p} c_k z^{-k}} \tag{23}$$

In order to estimate spectral tilt from $s_t(n)$, a 1^{st} order LP analysis is carried using Eqn. (22) as shown in the block

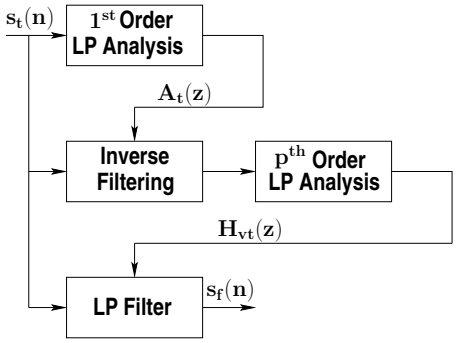

Fig. 6: The formant enhancement block diagram, where, $s_t(n)$ is excitation based enhanced foreground speech signal, $A_t(z)$ is a 1^{st} order LP filter, $H_{vt}(z)$ is the p^{th} order LP filter, and $s_f(n)$ is formant enhanced foreground speech signal.

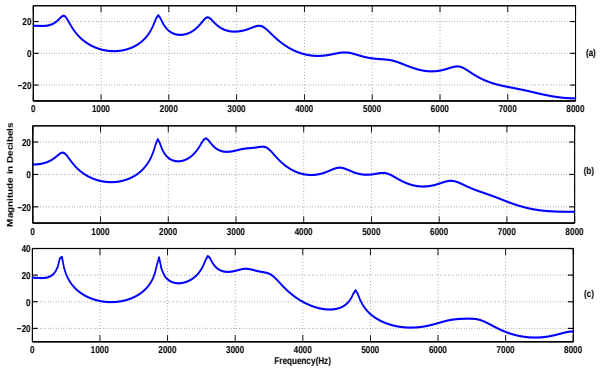

Fig. 7: Illustration of formant enhancement. (a) LP magnitude spectrum obtained from a 20 ms voiced segment before formant enhancement, (b) LP magnitude spectrum of the intermediate stage $H_{vt}(z)$, and (c) LP magnitude spectrum after formant enhancement.

diagram in Fig. 6. The residual signal obtained from 1^{st} order LP analysis is the excitation based enhanced speech signal $s_t(n)$ minus the spectral tilt. Hence, the LP filter estimation using Eqn. (23) would model the vocal tract information without spectral tilt. Therefore, the foreground speech signal $s_t(n)$ passed through Eqn. (23) will further enhance the formant peaks relative to adjacent valleys while maintaining the spectral tilt. The output speech signal $s_f(n)$ is formant enhanced foreground speech signal. The formant enhancement is illustrated using LP magnitude spectrum plots obtained from voiced frame before and after formant enhancement as shown in Fig. 7. The Fig. 7(a) shows the LP magnitude spectrum of coefficients derived from speech signal $s_t(n)$. The Fig. 7(b) shows LP magnitude spectrum derived from $H_{vt}(z)$. It can be noticed that, the LP magnitude spectrum is similar to Fig. 7(b) except that, the magnitude response is relatively flat. It can be noticed that, the LP spectrum derived from speech signal $s_f(n)$ as shown in Fig. 7(c) has sharper formant peaks relative to Fig. 7(a). Also, the peak to adjacent valley ratio is increased.

This is further illustrated using Figs. 5(c) and (g), which shows the foreground speech signal $s_f(n)$ after formant enhancement and its corresponding narrow band spectrogram, respectively. It can be noticed that there is reduction of background noise and formant tracks are enhanced in the foreground regions. The effect of passing foreground speech signal $s_t(n)$ through LP filter makes formant peaks much

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

sharper than original signal by moving poles of the filter closer to unit circle. Hence, Sections III-A and III-B uses the production aspects to enhance foreground speech signal. The enhancement of foreground speech signal is achieved without exclusive modeling of noise and such enhancement does not introduce unwanted musical noise to enhanced speech signal. Since, the poles of all pole LP filter moves close to unit circle in case of such formant enhancement, this makes speech sound unnatural [26]. Consequently, further processing is necessary to perceptually make speech more natural and enhance the foreground speech of any left over residual noise.

C. Perceptual based Foreground Speech Enhancement

The formant enhanced foreground speech signal obtained in Section III-B is further subjected to enhancement using cepstral analysis and synthesis on mel frequency scale. However, the cepstral analysis involves deriving mel cepstral coefficients (MCCs) using which, the enhanced speech signal is synthesized through mel log spectral approximation (MLSA) filter [27]. The advantage of using MCCs is two fold, where, the coefficients can be directly used in speech and speaker recognition applications apart from foreground enhancement. The MCCs $C_\alpha(m)$ are the Fourier cosine coefficients of spectral envelope derived from mel log spectrum. The spectrum represented by MCCs closely resemble human auditory spectral resolution having higher resolution at lower frequencies and lower resolution at higher frequencies [28]. The MLSA filter is applied to get approximate vocal tract response from MCCs using the adaptive algorithm. The true spectrum of MLSA filter for m^{th} order MCCs $c(m)$ is given by

$$H^\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{24}$$

is an all-pass function, which represents the mel-warped frequency characteristics and α is a coefficient corresponding to the mel-scale (for example $\alpha = 0.35$ for 10 kHz sampling rate)

$$\beta_\alpha(\Omega) = tan^{-1}\frac{1 - \alpha^2 sin(\Omega)}{(1 + \alpha^2)cos(\Omega) - 2\alpha} \tag{25}$$

where, α depends on sampling frequency and $\beta_\alpha(\Omega)$ is the phase of all-pass function, the smooth spectral envelope $G_\alpha(\tilde{\Omega})$ of mel log spectrum is expressed as polynomial function of order M given by

$$G_\alpha(\tilde{\Omega}) = \sum_{m=0}^{M} C_\alpha(m)cos(m\tilde{\Omega}) \tag{26}$$

where, $\tilde{\Omega}$ is mel frequency scale given by $\beta_\alpha(\Omega)$ and $C_\alpha(m)$ are the cepstral coefficients of order M.

In order to compute 34 dimensional MCCs, a Hamming windowed frame of size 20 ms with a frame shift of 10 ms is considered. The smoothed spectral envelope is computed from the MCCs using MLSA filter. The MLSA filter provides the best mean square approximation of log spectrum envelope on linear frequency scale and further used to directly synthesize the best quality speech signal. The MLSA filter needs excitation signal along with MCCs in order to synthesize the speech signal. The excitation signal is synthesized using F_0 information along with voiced/unvoiced decision. Alternatively, in the current work, WLPR $r_w(n)$ is used as excitation signal along with MCCs to synthesize perceptually enhanced foreground speech signal $s_p(n)$. This is illustrated in Fig. 8, where, Fig. 8(a) shows log magnitude spectrum of 20 ms voiced frame taken from the original recording. The Fig. 8(b) shows log magnitude spectrum from excitation based enhancement (EBE) method, while Fig. 8(c) illustrates the log magnitude spectrum of FBE method. The Fig. 8(d) shows smoothed log magnitude spectrum derived from MLSA filter. It can be observed that, the smoothed log magnitude spectrum forms the envelope of spectrum shown in Fig. 8(c). Consequently, the log magnitude spectrum of speech signal is synthesized through MLSA filter by using WLPR signal $r_w(n)$ and MCCs and it is shown in Fig. 8(e). It can be noticed that, the sharpness of formant peaks are relatively reduced. This helps to get rid of unnaturalness that is introduced in Section III-B due to formant enhancement. Also, the current block further attenuates remaining background noise. The Figs. 5(d) and (h) show perceptually enhanced speech signal and its narrowband spectrogram, respectively. It can be observed, that the sharpness of formant locations are reduced and further there is attenuation of background noise.

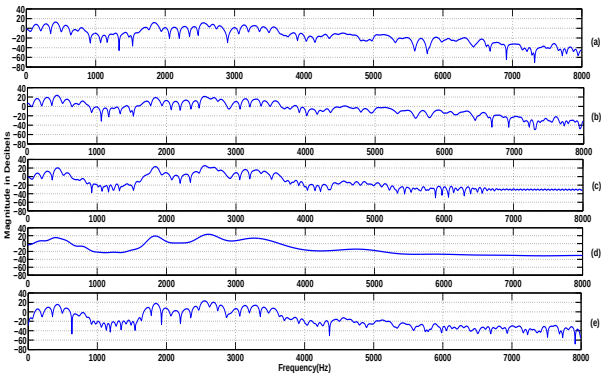

Fig. 8: Illustration of different outputs obtained from 20 ms voiced frame using log magnitude spectrum of (a) original speech recording in foreground scenario, (b) excitation based enhanced output, (c) formant based enhanced foreground speech, (d) smoothed envelope obtained from MCCs using MLSA filter, and (e) perceptually enhanced foreground speech signal using MLSA filter.

IV. Experimental Results and Discussions

The proposed scheme of foreground speech enhancement is evaluated with some of the well known speech quality measures along with 3 state-of-the-art methods considered. In order to evaluate the proposed work both subjective and objective measures are used. The mean opinion scores (MoS) and preference test scores are used for subjective analysis which are discussed in Sections IV-B1 and IV-B2, respectively. The clean speech signal is seldom available while recording in noisy environments. Therefore, two objective measures *viz.*, foreground to background ratio (FBR) similar to *a posteriori* SNR and epoch to non-epochal ratio (ENR) measurements are introduced in Sections IV-B3 and IV-B4, respectively. All methods considered for evaluation are compared using FBR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

and ENR measurements. The data set consisting of natural recordings in noisy environments are used to evaluate different methods using MoS, preference test, FBR, and ENR measures. Subjective analysis of enhancement methods are time consuming task and often it is difficult to get right subjects for subjective evaluations. Therefore, perceptual evaluation of speech quality (PESQ) is one of the important objective measures that closely resemble subjective analysis [29]. However, PESQ measurement requires reference clean speech signal for evaluations. Hence, TIMIT speech database is used to benchmark the proposed work [30].

A. Performance Evaluation of Foreground Speech Segmentation

The foreground speech segmentation can be compared to voice activity detection (VAD) as they have similar objective. Hence, the performance is compared with considered two state-of-the-art VADs. One of the VAD considered for performance evaluation is the latest G.729 ITU-T VAD standard developed for fixed telephony and multimedia communications [31]. The G.729 VAD uses multiple boundaries for voice activity decision. More recently a variable frame rate approach is proposed on the basis that, the speech signal is not stationary in short period of fixed frame rate [32]. Therefore, a variable frame rate (VFR) approach is followed based on *a posteriori* SNR weighted energy distance. In order to compare the performance of all three different methods, 10 different TIMIT speech files consisting of 5 female and 5 male speakers are considered [30]. The speech files are modified by appending 50 ms silence on either sides of speech signal to closely simulate the natural recording scenario. Also, it can be noted that the ground truth of VAD is manually marked in TIMIT files and they form the reference boundaries for evaluation. The noisy speech recordings are simulated by additively combining clean speech signal with 4 different types of noise at different levels. The 4 different types of noise considered are mosaic machine noise, hostel mess recording (babble noise), background music with vocals and background speech. The background noise files are recorded in 4 different natural environments as explained in Section II using headphone microphone connected to laptop. However, the background noise sources are far away from microphone. In order to compare the performance of all three methods, the following parameters are used

- Correct VAD decision (CD): Correct decisions made by the VAD.
- Front End Clipping (FEC): Clipping introduced while passing from noise to speech activity.
- Missed Speech Clipping (MSC): Clipping due to speech misclassified as noise.
- Noise Detected as Speech (NDS): Noise interpreted as speech within silence period.
- OVER: Noise interpreted as speech due to VAD flag remaining active in passing from speech activity to noise.

All parameters are expressed in terms of percentage. The Fig. 9 illustrates the performance of three different methods in the form of graph plots. The Figs. 9(a), (e), (i), (m) and (q) show the performance of different methods in terms of CD, FEC, MSC, NDS, and OVER for additive background mosaic machine noise, respectively. It can be noted that, the noise is added at 20, 15, 10, 5, 0, and -5 decibels. It is observed that, the performance of foreground segmentation and VFR VAD are equally robust for background machine noise even at low SNR levels. However, the performance of G.729 VAD degrades as the noise level increases. The reason for degradation in case of G.729 is due to increase in FEC, MSC and OVER. Similarly, Figs. 9(b), (f), (j), (n) and (r) show the performance of three methods for additive background noise recorded in a large hostel mess environment when it is crowded (babble noise). It can be observed from the plots that the performance of foreground segmentation and VFR VAD remains robust upto additive noise level of 5 dB and followed by degradation in performances. However, it can be noticed that the performance of VFR VAD is better compared to foreground segmentation and G.729 VAD at lower SNR levels. The deterioration in the performance of foreground segmentation is mainly due to the increased OVER rate at low SNR levels, particularly at 0 and -5 dB levels. The Figs. 9(c), (g), (k), (o) and (s) show the performance curves for additive background music with vocals at different levels. It can be observed from the plots that the foreground segmentation performance is better than other two methods at low SNR levels. Similarly, Figs. 9(d), (h), (l), (p) and (t) show the performance of 3 different methods for additive background speech at different levels. It can be noticed that there is degradation in performance of all 3 methods at low SNR levels. However, the performance of foreground segmentation is marginally better compared to VFR VAD at low SNR levels. The signal characteristics of background music with vocals and background speech are similar to foreground speech signal. Hence, analyzing the signal at zero frequency rather than entire set of frequencies helps in better discrimination of foreground speech from rest of the background noise as in case of foreground speech segmentation. As explained in Section II, the features derived from ZBFS such as ZBFS energy and NACC are used as features for foreground segmentation. Also, modulation spectrum energy centered at 4 Hz is used in combination with features derived from ZBFS. The performance evaluation shows that, the foreground segmentation is robust to interfering background noise.

B. Performance Evaluation of Foreground Speech Enhancement using Natural Recordings

In order to evaluate proposed method, the speech files recorded in natural environment from ten different speakers are considered, and this include 3 female and 7 male speakers. Typically each speaker has spoken two to three different sentences in each such recordings. The spoken sentences are chosen from English radio broadcast and it has the composition of 74.10% of voiced sounds and 25.90% unvoiced sounds. All speakers were native Indians and were well versed with English as their second language. The speech is recorded in 7 different natural environments that includes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

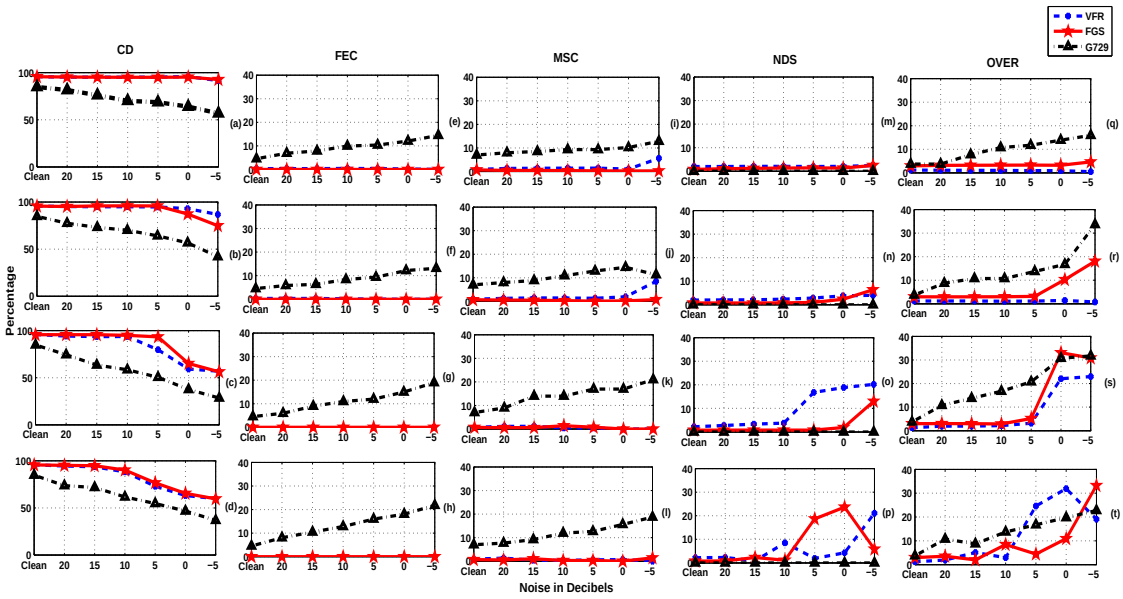

Fig. 9: Illustration of performance evaluation of VAD using foreground segmentation, VFR VAD, and G.729 VAD for 4 different noise types, ((a) - (d)) is CD, ((e) - (h)) is FEC, ((i)-(l)) is MSC, ((m) - (p)) is NDS, ((q) - (t)) is OVER plots with different additive noises of machine noise, babble noise, background music with vocals and background speech, respectively.

busy city road traffic, office room when mosaic machine is operating at the background, air condition machine room, home environment when television is on, room environment when background music is playing along with vocals, crowded hostel mess (babble noise), and building construction site when concrete mixing machine is on. All speech files are recorded in foreground scenario, where, the foreground speaker is closer to microphone sensor relative to other interfering sources. The recording setup includes headphone connected to laptop. The speech files are recorded using WAVESURFER [33] tool at the sampling rate of 16 kHz. Three different headphone sets of different make and prices were used for the recordings in order to maintain the variability of sensors. It can be noted that, the headphones used for recording had no special front end preprocessing circuits to enhance the speech signal. In all such recordings, the speaker is wearing headphone and typically microphone is closer to mouth of the speaker (within 1 to 2 inches). The background acoustic sources are far away from microphone sensor compared to foreground speaker.

The proposed method is compared with three other methods chosen from literature. The spectral domain subtraction based method is one of the earliest methods for enhancing speech signal distorted by additive noise. In spectral subtraction based methods the noise components are modeled by average magnitude spectrum using several frames of noise only regions. The noise spectrum is subtracted from signal spectrum to obtain the enhanced output [3]. The method proposed in [4] assumes that the Fourier coefficients of speech and noise can be independently modeled as zero mean Gaussian random variables. The method aims to minimize the mean square error between the clean speech and enhanced speech signal and hence it is called as minimum mean square error (MMSE). There are several modifications suggested to these two basic methods in literature to improve the quality of enhancement by reducing the musical noise. The matlab implementations available in VOICEBOX [34] are used to compare with the proposed method in this paper. Recently, a method is proposed that makes use of temporal and spectral processing to enhance the degraded speech signal [12]. The temporal processing relies on HELP signal to extract the epoch locations and enhance those locations using LP residual signal, similar to excitation based enhancement module in this work. The signal is enhanced in spectral domain by a comb like function that emphasizes the fundamental component and its harmonics. Both temporal and temporal-spectral enhancement are compared with the proposed method.

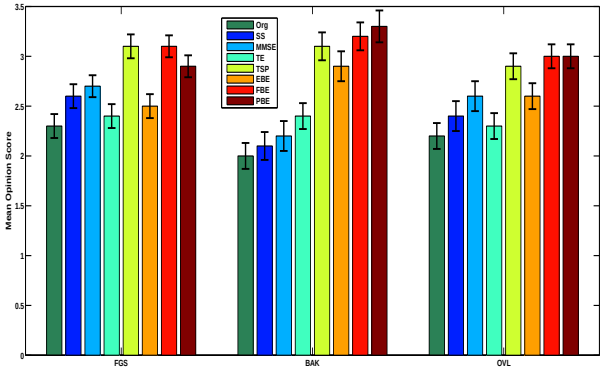

Fig. 10: Bar graph representing the mean opinion scores obtained from different subjects, where, FGS is foreground speech, BAK is background noise and OVL is overall ratings. The graph also depicts the error computed using 95% confidence interval from the subjective scores.

1) Subjective Evaluation using MoS Score: The subjective evaluation is carried using three parameters as foreground speech (FGS), background noise (BKG) and the overall quality (OVL) of enhanced speech signal. The original and enhanced speech files were provided to 24 different subjects in random

order for subjective evaluation scores similar to [35]. The subjects include 16 male and 8 female adult listeners having an average age of 26 ± 5 years. All listeners are well versed with English language and English was their medium of instruction in their academic studies. The listeners are mainly working in the area of speech processing, speech synthesis, speaker verification and speech recognition topics and they typically have 1 to 3 years of experience in this domain. All listeners had normal hearing abilities without any kind of hearing impairment. The files were listened to in a typical lab environment when air condition is switched on. The files were listened without visualizing the waveform using similar types of headphones connected to personal computer. The subjects were provided with the following instructions for each of such parameters.

- Focus listening on FGS regions alone in terms of reduced background noise, intelligibility and lesser distortion with scales suggesting [1 - Very Unnatural, 2 - Fairly Unnatural, 3 - Somewhat Natural, 4 - Fairly Natural, 5 - Very Natural].
- Focus listening on BAK regions alone in terms of reduced background noise, and lesser distortion with scales suggesting [1 - Very Intrusive, 2 - Somewhat Intrusive, 3 - Noticeable but not Intrusive, 4 - Somewhat Noticeable, 5 - Not Noticeable].
- Focus listening on the OVL in both foreground and background regions in terms of reduced background noise, intelligibility and lesser distortion with scales suggesting [1 - Bad, 2 - Poor, 3 - Fair, 4 - Good, 5 - Excellent].

The subjects were provided with three sets of same files arranged in random order for evaluation. The Fig. 10 shows the results of subjective evaluation in terms of MoS using the original and enhanced speech files generated using different methods. The Fig. 10 shows the bar graph plots along with margin of error with 95% confidence interval. It can be observed from Fig. 10 that all three scores i.e., FGS, BAK and OVL is lower in case of original signal as expected and this forms the reference score to evaluate other methods. The subjective scores obtained for MMSE is better compared to SS in all three parameters. Also, MMSE based method is better than temporal enhancement in terms of FGS and OVL, while, TE remains better in terms of BAK compared to SS and MMSE. Though, conceptually TE and EBE are similar in approach, it can be noticed that EBE performs better than TE in all three parameters. The reason of EBE performing better than TE is because of better identification rate (IDR), identification accuracy (IDA), and lower miss rate (MR) to extract epoch locations using ZBF compared to HELP [17]. Also, ZBF remains robust to noisy conditions and hence leads to lower distortion while temporally processing the LP residual signal. This helps to set reasonably lower threshold value to attenuate the background noise in case of EBE. Consequently, further processing modules can benefit from such higher attenuation of background noise in the proposed work.

It can be noticed that TSP, FBE, and PBE methods perform better than SS, MMSE, TE, and EBE methods. The TSP and FBE based methods outperform other methods in terms of FGS, while, FBE and PBE based methods are better in terms of BAK and OVL. It can be noted that spectral processing in case of TSP is useful in attenuating the valleys of spectral magnitude and formant peaks remain unmodified. In case of FBE, the formant peaks are boosted further relative to spectral valleys. The formant locations are high SNR regions in spectral domain and hence relative enhancement of formant peak locations help to reduce the background noise and enhance the foreground speech regions. This is evident from subjective evaluation using FGS and OVL parameters. Any residual background noise left over can be reduced further by PBE method and this can be observed from Fig. 10 by the increased BAK and OVL. Also, it can be observed that MoS of TSP, FBE and PBE methods in terms of FGS, BAK, and OVL are better than TE and EBE methods.

TABLE I: Subjective evaluation of different methods using preference test score in terms of percentage, where, SS - spectral subtraction, MMSE - minimum mean square error approximation, TE - temporal enhancement, TSP - temporal and spectral processing, EBE - excitation based enhancement, FBE - formant based enhancement, and PBE - perceptual based enhancement.

Method	SS	MMSE	TE	TSP
EBE	51%	44%	53%	39%
FBE	58%	54%	61%	49%
PBE	61%	58%	63%	54%

2) Subjective Evaluation using Preference Test: Preference test is one of the simpler tests to asses the speech quality. A subset of speech files that were used for MoS evaluations in Section IV-B1 are used in preference test. The speech files recorded from 5 different speakers in 5 different environments are used. The speakers include 2 female and 3 male speakers and each has spoken a English sentence chosen from English Broadcast. The speech files are assessed by 10 different subjects that include 5 female and 5 male listeners. All listeners were adults and not having any hearing impairments and the average age of listeners is 28 ± 5 years. The listening environment was in a typical computer laboratory with relatively less background noise. The listeners used high quality headphones of similar make to listen speech files. The files are enhanced using SS, MMSE, TE, TSP, EBE, FBE and PBE methods. A pairwise listening test was conducted, where, one of the speech files were either EBE, FBE, or PBE, while, the other is SS, MMSE, TE, or TSP. The listeners were asked to listen to reference file before listening to pair. Here, reference file is the original recording in foreground scenario. The 5 different noisy environments include mosaic machine noise, building construction noise, traffic noise, background music with vocals, and background speech. The subjects were given with following instructions to asses files

- Listen to reference file before listening to individual pair of files.
- Asses the speech files in terms of reduced background noise, lesser distortion, and speech intelligibility by giving equal weightage to all 3 features.

Table I shows the preference test scores averaged for 5 different types of noisy files assessed by 10 different subjects.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

The percentage score refers to which subjects have preferred EBE, FBE, and PBE over other methods. It can be observed that, the subjects have preferred EBE over SS and TE, while, MMSE and TSP scores better compared to EBE. The EBE does not cause distortion in the form of musical noise. Also, due to better epoch extraction method, the background noise is suppressed relatively more compared to TE method. However, the temporal enhancement alone is not sufficient to reduce the background noise and hence MMSE and TSP scores better than EBE. The FBE is preferred relatively higher compared to SS, MMSE, and TE. The formant enhancement helps to increase the spectral peaks which are essentially high SNR regions in spectral domain compared to spectral valleys. Hence, there is reduction of background noise in spectral domain and this may be the reason for preference over SS, MMSE, and TE. The sharpening of formant peaks leads to audible distortion and this can be the reason for low preference score compared to TSP. However, the spectral envelope is smoothened by PBE module to reduce the distortion and further elimination of any residual noise left in previous stages. Hence, overall PBE performs better in terms of preference score. In case of preference test it is difficult to asses the specific reason for the choice made by subject as the preference depends on all 3 factors. However, comparing preference test score with MoS test reveals that, the overall trend of preference test score is similar to OVL of MoS from Section IV-B1.

3) Foreground to Background Ratio (FBR): When speech signal is recorded in natural environment, seldom speech and noise signals are available separately to *a priori* estimation of speech and noise signal powers, respectively. Since, the speech signal is recorded in natural environments and there can be many types of interfering noises, and that may include background speakers. As explained in Section II foreground speech segmentation is one of the reliable ways to temporally segment foreground speech from rest of the background regions. Hence, gross weight function $w_g(n)$ obtained from Eqn. (13) is used to segment the foreground and background regions using the following relation

$$f(n) = \begin{cases} 1, & \text{if } w_g(n) > \mu_{w_g(n)} \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

where, $f(n)$ is the binary signal indicating foreground and background regions and $\mu_{w_g(n)}$ is the mean value derived from $w_g(n)$. Hence, the SNR computed using foreground segmentation is called foreground to background ratio (FBR). The FBR can be computed using the following relationships

$$\hat{\sigma}_f^2 = \frac{1}{L_f} \sum_{n=0}^{l-1} s^2(n).f(n) \tag{28}$$

$$\hat{\sigma}_b^2 = \frac{1}{L_b} \sum_{n=0}^{l-1} s^2(n).(1 - f(n)) \tag{29}$$

$$FBR = 10\ log_{10} \frac{\hat{\sigma}_f^2}{\hat{\sigma}_b^2} \tag{30}$$

where, $\hat{\sigma}_f^2$ is foreground speech power estimation, L_f is the number of samples in foreground speech region, $s(n)$ is the naturally recorded speech signal, $f(n)$ is foreground background binary signal, $\hat{\sigma}_b^2$ is the background power estimation, and L_b is the number of samples in background region. In order to study the characteristics of FBR, 10 different speech files from TIMIT database consisting 5 female and 5 male speakers are considered. The clean speech files are added with 4 different types of noise at different levels as shown in Fig. 11. The x-axis represents the added noise levels to speech files at 20, 15, 10, 5 and 0 dB, whereas, y-axis represents the FBR calculated using the relationship given in Eqn. (30) expressed in decibels. It can be observed that there is linear relationship between the noise added to clean speech files and the foreground to background ratio estimated in all 4 noisy cases *viz.*, machine noise, babble noise, background music with vocals and background speech. However, it can be noticed that there is degradation in estimating FBR in case of background speech case at 0 dB noise level. Overall the FBR estimate is reliable and robust to different types of additive noise at different levels. Hence, FBR can be used to measure *a posteriori* SNR that can help indicate the suppression of background noise by different methods.

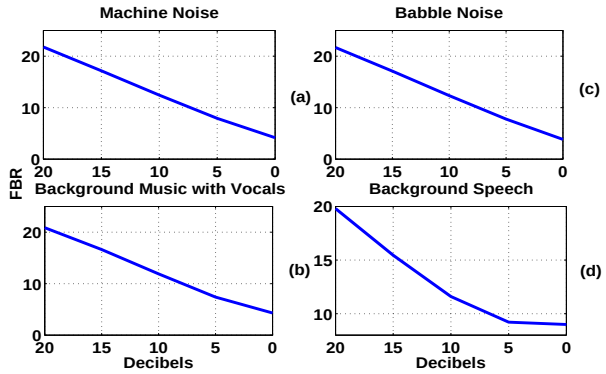

Fig. 11: Average FBR obtained from 10 different TIMIT speech files by adding noise at different levels of 20, 15, 10, 5 and 0 dB, with different types of noise like (a) Mosaic Machine Noise, (b) Hostel Mess Noise (Babble Noise), (c) Background Music with Vocals, and (d) Background Speech.

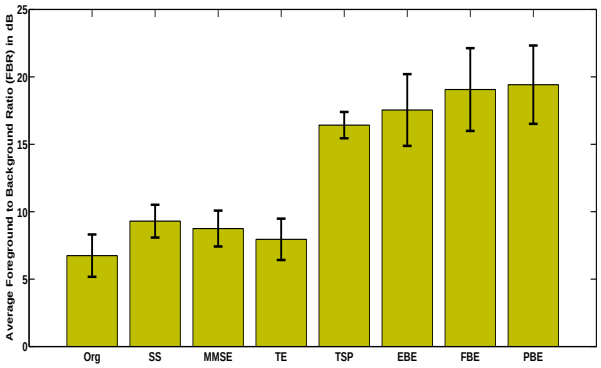

Fig. 12: Bar graph representing objective scores in terms of average FBR obtained from original and processed outputs. The graph also depicts the error computed using 95% confidence interval from the objective scores.

The same set of files used to evaluate subjective measures as described in Secion IV are used to evaluate the performance of different methods using FBR. Totally 27 different speech

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

files collected naturally in different noise environments are considered for FBR measurement. The FBR is computed for all the enhanced outputs using different methods. The objective scores are illustrated using bar graph in Fig. 12 in terms of average FBR expressed in decibels (dB). Also, the plot indicates the margin of error with 95% confidence interval. It can be noticed that the performance of TSP, EBE, FBE, and PBE methods are superior compared to SS, MMSE and TE methods. The objective scores obtained further corroborates the subjective analysis as discussed above. However, the performance of FBE, and PBE remains best in terms of FBR. The difference between the average scores of original signal and the enhanced outputs from FBE and PBE shows that there is an improvement of 12 dB. This shows that, the proposed method performs best in attenuating the background noise compared to other methods and still maintains the overall quality of enhanced foreground speech. The similar trend can be observed from subjective analysis using BAK parameter as shown in Section IV-B1.

4) Epoch to Non-Epochal Ratio (ENR): The effect of interfering sources are not uniform throughout the foreground regions. There are certain regions of foreground speech that are relatively more robust to interfering sources, especially the regions around instants of significant excitation are high SNR regions compared to other regions within a glottal cycle. Hence, the ratio between energy around epochal region to non-epochal region within a glottal cycle of foreground speech can be an important objective measure to evaluate different methods. In order to compute such a ratio, Hilbert Envelope of LP residual (HELP) signal is considered similar to [36], where, Hilbert Envelope of LP residual $e(n)$ is given by the following relationship

$$h(n) = \sqrt{e^2(n) + {e_h}^2(n)} \tag{31}$$

where, $h(n)$ is Hilbert envelope and $e_h(n)$ is the Hilbert transform of $e(n)$. This measurement, when compared between the original recording and enhanced output will help to assess the performance of different methods in terms of enhancing high SNR regions further relative to other regions. The quantity essentially measures the energy between epochal and non-epochal regions and hence termed as *Epoch to Non-Epochal Ratio* (ENR).

The epochal energy is calculated by considering the summation of normalized energy around 3 ms of epoch locations, where, 3 ms closely corresponds to glottal closure interval. The non-epochal energy is computed by the summation of normalized energy excluding 3 ms region around the epoch locations with reference to each glottal cycle. The ENR can be computed using the following relationships

$$\hat{E} = \sum_{k=1}^{N_k} \frac{\frac{1}{2M+1}\sum_{p=i_k-M}^{i_k+M} h_e^2(p)}{\frac{1}{L_1}\sum_{q=i_{k-1}+M+1}^{i_k-M-1} h_e^2(q) + \frac{1}{L_2}\sum_{s=i_{k+1}-M-1}^{i_k+M+1} h_e^2(s)} \tag{32}$$

and

$$\hat{O} = \sum_{k=1}^{N_k} \frac{\frac{1}{2M+1}\sum_{p=i_k-M}^{i_k+M} h_o^2(p)}{\frac{1}{L_1}\sum_{q=i_{k-1}+M+1}^{i_k-M-1} h_o^2(q) + \frac{1}{L_2}\sum_{s=i_{k+1}-M-1}^{i_k+M+1} h_o^2(s)} \tag{33}$$

where, $h_e(p)$ is HELP derived from enhanced foreground speech, $\hat{E}$ is the estimation of ENR using enhanced foreground speech, N_k is total number of epochs in foreground speech regions, M corresponds to samples of 1.5 ms, i_k is the epoch location at k^{th} epoch, i_{k-1} is the previous epoch location to i_k, i_{k+1} is the epoch location after i_k, $L_1 = i_k - i_{k-1} - 2M - 2$, $L_2 = i_{k+1} - i_k + 2M + 2$, and $\hat{O}$ is the estimation of ENR using HELP derived from original recording $h_o(p)$. The ratio between $10\log 10\frac{\hat{E}}{\hat{O}}$ represents the improvement achieved in terms of enhancing high SNR regions further relative to low SNR regions. A similar approach is followed to characterize ENR as in case of FBR explained in Section IV-B3. The 10 speech files are from different speakers chosen from TIMIT database is additively corrupted using 4 types of noise at different levels. The Fig. 13 shows the ENR evaluation for clean and additive noise cases. It can be observed that there is a linear relationship between additive noise and ENR for all 4 different noise types. It can be noticed that ENR increases linearly as additive noise increases. The ENR is computed

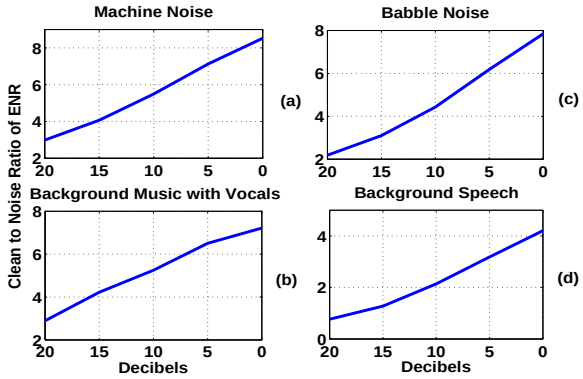

Fig. 13: Average ENR obtained from 10 different TIMIT speech files by adding noise at different levels of 20, 15, 10, 5, and 0 dB, with different types of noise like (a) Mosaic Machine Noise, (b) Hostel Mess Noise (Babble Noise), (c) Background Music with Vocals, and (d) Background Speech.

for all the enhanced outputs using different methods. The objective scores are as shown in Table II that is computed by averaging the scores from all files. It can be noted that the ratio is computed between enhanced output to original recording while computing ENR. It can be observed that EBE and PBE methods are best amongst all other methods in terms of ENR. The improvement achieved in case of EBE is due to better identification of epochs using ZBF, as a result of which the noise components are suppressed in LP residual to obtain WLPR. The speech signal synthesized using WLPR results in enhanced output in which high SNR regions are enhanced further compared to low SNR regions. However, in case of PBE the spectral envelope is smoothened and this leads to reduction of background noise, mainly in low SNR regions. Hence, EBE and PBE methods have high ENR compared to other methods. In case of FBE, due to LP filter enhancement, the formant peaks are sharpened and thereby increasing the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2549699, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE II: Objective evaluation of different methods using Epoch to Non-Epochal Ratio (ENR) that is computed as a ratio between ENR of enhanced foreground speech to original recordings. The table represents the average ratio expressed in decibels (dB) computed across all the enhanced speech files obtained from different methods.

SS	MMSE	TE	TSP	EBE	FBE	PBE
1.28	0.83	3.68	2.31	4.23	1.26	5.46

TABLE III: Perceptual Evaluation of Speech Quality Scores, where, MMN - Mosaic Machine Noise, MN - Hostel Mess Noise, TN - Traffic Noise, BM - Background Music with Vocals, BS - Background Speech SS - spectral subtraction, MMSE - minimum mean square error approximation, TE - temporal enhancement, TSP - temporal and spectral processing, EBE - excitation based enhancement, FBE - formant based enhancement, and PBE - perceptual based enhancement.

Noise	Decibels	Original	SS	MMSE	TE	TSP	EBE	FBE	PBE
	15	2.5	**3.2**	3.1	2.9	3.0	2.9	3.0	3.1
MMN	**10**	2.3	**3.1**	3.0	2.8	2.9	2.8	2.9	3.0
	5	1.8	**2.7**	2.6	2.4	2.5	2.5	2.5	2.6
	15	2.6	3.0	2.9	2.8	**3.1**	2.8	2.9	3.0
MN	**10**	2.2	2.7	2.6	2.7	**3.0**	2.7	2.8	2.9
	5	1.9	2.3	2.2	2.5	**2.6**	2.4	2.5	**2.6**
	15	2.4	2.6	2.5	2.8	**2.9**	2.7	2.8	**2.9**
TN	**10**	2.2	2.3	2.2	2.7	**2.8**	2.6	2.7	**2.8**
	5	1.8	2.0	2.0	2.2	**2.3**	2.0	2.1	2.2
	15	2.4	2.8	2.7	2.7	**3.0**	2.8	2.9	**3.0**
BM	**10**	2.2	2.4	2.3	2.5	2.7	2.6	2.7	**2.8**
	5	1.8	2.1	2.0	2.2	**2.4**	2.2	2.3	**2.4**
	15	2.8	2.8	2.7	2.8	3.0	2.8	3.0	**3.1**
BS	**10**	2.7	2.7	2.6	2.7	2.9	2.8	2.9	**3.0**
	5	1.9	1.8	1.8	2.2	2.3	2.3	2.3	**2.4**

gain of the filter transfer function at formant locations. The enhanced LP filter convolves with excitation signal resulting in relatively higher amplitude at non-epochal regions and hence the ENR is lower in case of FBE. The enhancement of foreground speech signal using FBE is achieved mainly because of enhancement of high SNR regions in spectral domain.

C. Perceptual Evaluation of Speech Quality (PESQ)

The PESQ is carried using 10 TIMIT speech files taken from 5 female and 5 male speakers. The clean speech files are corrupted by adding 5 different types of noise at 3 different levels. The Table III shows the average PESQ scores obtained from 10 different speech files. It can be observed that in case of mosaic machine noise (MMN) the spectral subtraction method performs well. Since, MMN is nearly stationary noise and hence SS is able to better model the background noise in such cases. However, in case of hostel mess noise (babble noise) and traffic noise which are relatively non stationary in nature, TSP and PBE methods performs better than other enhancements. The background music with vocals and background speech cases are the most challenging, as the background noise characteristics is similar to foreground speech regions. Due to robustness of foreground speech segmentation and better estimation of GCI locations using ZBF, the performance of the proposed work is relatively better compared to other methods. Overall, the performance of TSP and PBE are better and comparable. Though, it is difficult to correlate all the parameters of subjective MoS with PESQ score, the trend remains consistent with subjective analysis.

V. Summary and Conclusions

In this paper, a new way to approach the problem of speech enhancement is suggested, where, the distance between the foreground speaker to microphone and rest of the background sources is utilized. The proposed work relies on known production and perceptual features to enhance the foreground speech. The advantage of proposed method is that the distortion is significantly lower and does not introduce musical noise unlike other methods as spectral subtraction and MMSE. The method exploits reliable ZBF for foreground segmentation and extraction of glottal closure instants, due to which, higher attenuation of background noise is possible with least distortion. The advantage of using ZBF is that, there is no necessity of finding F_0 of foreground speaker. The performance of proposed work is compared with 3 other existing state-of-the-art methods in terms of subjective and objective evaluations. It is found that, the proposed method can significantly attenuate the background noise and still maintain the better quality of enhanced foreground speech signal. Future work should focus on using proposed framework of production and perception features for enhancement in other types of degradation like reverberation and multi-speaker environment.

Acknowledgment

This work is part of the projects on the development of ”Prosodically Guided Phonetic Engine for Assamese Language” funded by the Technology Development for Indian Languages (TDIL) Programme initiated by the Department of Electronics & Information Technology (DeitY), Govt. of India and the development of Speech Based Multi-level Person Authentication System funded by the e-security division of Department of Electronics & Information Technology (DeitY), Govt. of India.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Taylor and F. Group, Eds. CRC Press, 2013.

[2] K. Parlak and O. G. Moreno, *Applied Speech Enhancement in Mobile Communication Acoustics: Background Noise Elimination with Filtering Algorithms*. LAP LAMBERT Academic Publishing, 2012.

[3] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[4] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[5] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 9, pp. 504–512, 2001.

[6] T. Gerkmann and R. C. Hendriks, “Unbiased mmse-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 20, pp. 1383–1393, 2011.

[7] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 7, pp. 126–137, 1999.
[8] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 11, pp. 457–465, 2003.
[9] F. S. Cooper, "Acoustics in human communication: Evolving ideas about the nature of speech," *Journal of the Acoustical Society of America*, vol. 68, pp. 18–21, 1980.
[10] J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *Journal of the Acoustical Society of America*, vol. 93, pp. 510–524, 1993.
[11] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25–42, May 1999.
[12] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication* vol. 53, pp. 154–174, February 2011.
[13] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A Stokes, "Effects of noise on speech production: Acoustic and perceptua analyses," *Journal of the Acoustical Society of America*, vol. 84, pp 917–928, 1988.
[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.
[15] K. T. Deepak, B. D. Sarma, and S. R. M. Prasanna, "Foreground speech segmentation using zero frequency filtered signal," in *Interspeech* September 2012.
[16] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 16, pp. 1602–1613, November 2008.
[17] K. T. Deepak and S. R. M. Prasanna, "Epoch extraction using zero band filtering from speech signal," *Circuits, Systems, and Signal Processing*, p. [Online], December 2014.
[18] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Lett.*, vol. 16, no. 6, pp. 469–472, June 2009.
[19] C. L. Smith, C. P. Browman, R. S. McGowan, and B. Kaytt, "Extracting dynamic parameters from speech movement data," *Haskins Laboratories Status Report on Speech Research*, vol. SR-105/106, pp. 107–140, 1991.
[20] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *ICASSP*, vol. 3, April 1997, pp. 1647–1650.
[21] T. Irino and R. D. Patterson, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust. Soc. Am*, vol. 109 (5), pp. 2008–2022, May 2001.
[22] M. Unoki, T. Irino, B. Glasberg, B. C. J. Moore, and R. D. Patterson, "Comparison of the roex and gammachirp filters as representations of the auditory filter," *J. Acout. Soc. Am.*, vol. 120(3), pp. 1474–1492, 2006.
[23] 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 and Lanuage Processing*, vol. 20, no. 3, pp. 994–1006, March 2012.
[24] J. Yang, F. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Commun.*, vol. 39, pp. 33–46, 2002.
[25] L. R. Rabiner and R. W. Scafer, *An Introduction to Digital Speech Processing (Foundations and Trends in Signal Processing)*. Now Publishers Inc, 2007.
[26] V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, "Enhancement of adpcm speech coding with backward-adaptive algorithms for postfiltering and noise feedback," *IEEE Journal On Selected Areas in Communication*, vol. 6, pp. 364–382, 1988.
[27] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," in *ICSLP*, September 1994, pp. 1043–1046.
[28] G. Fant, *Speech sound and features*. MIT Press, Cambridge, 1973.
[29] A. W. Rix, J. G. Beerends, D. S. Kim, P. Kroon, and O. Ghitza, "Objective assesment of speech and audio quality - technology and applications," *IEEE Tran. Audio, Speech and Signal Processing*, vol. 14, pp. 1890–1901, 2006.
[30] TIMIT, *"Timit Acoustic-Phonetic Continuous Speech Corpus"*, NIST Order PB91-505065, National Institute of Standards and Technology, Gaithersburg, MD, USA, 1990, Speech Disc 1-1.1., 1990.
[31] A. Benyassine, E. Shlomot, and H. Y. Su, "ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application," *IEEE Commun. Mag*, vol. 35, pp. 64–73, 1997.
[32] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 798–807, 2010.
[33] K. Sjlander and J. Beskow, "Wavesurfer an open source speech tool," in *Sixth International Conference on Spoken Language Processing*, 2000.
[34] *VOICEBOX: Speech Processing Toolbox for MATLAB*.
[35] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *ICASSP*, May 2006.
[36] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Tran. Audio, Speech and Signal Processing*, vol. 13, pp. 751–761, 2005.

K. T. Deepak was born in India on 1976. He received the B.E. degree in Instrumentation Technology from Malnad College of Engineering, University of Mysore, India, in the year 1999 and M.Tech degree in Bio Medical Instrumentation from Sri Jayachamarajendra College of Engineering, Visvesvaraya Technological University, Mysore, India in the year 2002. He worked as Senior Project Engineer in ST Microelectronics, Bangalore from the year 2005 to 2010, mainly focusing on audio post processing algorithms for the Digital Television group. He is currently working as Technical Leader at NXP Semiconductors India Pvt. Ltd. and pursuing Ph.D. degree in Electronics and Electrical Engineering at the Indian Institute of Technology Guwahati, India. His research interests include speech/audio processing, analysis and recognition.

S. R. M. Prasanna was born in India on 1971. He received the B.E. degree in Electronics Engineering from Sri Siddartha Institute of Technology (then with Bangalore University), India, in 1994. He received the M.Tech. degree in Industrial Electronics from the National Institute of Technology, Surathkal (then KREC Surathkal), India, in 1997, and the Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004. He is currently the Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati. His research interests are in speech and signal processing and published research articles in both national and international journals and conferences.