ELECTRONICS AND COMMUNICATION ENGINEERING DEPARTMENT

NATIONAL INSTITUTE OF TECHNOLOGY SIKKIM-737139

A MAJOR PROJECT REPORT ON

"Analysis of Variational Mode Function for Speech Enhancement with Application of Vowel detection"

SUBMITTED ON-13/12/2019

SUBMITTED BY:

AKSHEETHA MUTHUNOORU
(B160032EC)

LOHITH SURISETTI
(B160014EC)

KOTA HEMANTH KUMAR
(B160112EC)

UNDER GUIDANCE OF:

Dr. AVINASH KUMAR
ASSISTANT PROFESSOR
ELECTRONICS AND COMMUNICATION ENGINEERING DEAPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY SIKKIM
RAVANGLA BARFUNG BLOCK, SOUTH SIKKIM-737139

CERTIFICATE

This is to certify that the project entitled **"Analysis of Variational Mode Function for Speech Enhancement"** is submitted by Aksheetha Muthunooru (B160032EC), Lohith Surisetti (B160014EC) and Hemanth Kumar Kota (B160112EC), in **"Electronics and Communication Engineering",** during session 2018-2019 at National Institute of Technology Sikkim. It is an authentic record of research work carried in ECED in NIT Sikkim.

The Report is based on the candidate's own work and has not submitted elsewhere.

Dr. Sanjay Kumar Jana
HOD (I/c), ECE Department

Dr. Avinash Kumar
Project Supervisor

ACKNOWLEDGEMENT

“Gratitude is not a thing of expression; it is more a matter of feeling.”

I would like to express my sincere gratitude to our guide **Dr. Avinash Kumar** for his vital support, guidance and encouragement, without which the completion of major project would not have come forth. With his valuable suggestions and guidance, it has been helpful in various phases of the major project.

During the training period many people have best owed and supported me. Special thanks to **Dr. Sanjay Kumar Jana**, HOD(I/c), ECE Department, NIT Sikkim, faculty members & Ph.D. Scholars for their kind gesture and ever helping nature which made me in completion of mini project.

Finally, I would like to thank God Almighty for giving me the strength, knowledge, ability and opportunity to undertake this project study and to preserve and to complete it satisfactorily.

AKSHEETHA MUTHUNOORU **LOHITH SURISETTI** **HEMANTH KUMAR KOTA**

INDEX

ABSTRACT

The aim of the speech enhancement is to improve the intelligibility and quality of the speech. By recording the speech signal in the noisy environment, clean speech signals are degraded. Speech enhancement [2] reduces the noise without distorting the original (clean) signal. In this work, a speech signal is enhanced based on non-local means (NLM) [1,2] estimation and variational mode decomposition (VMD) [1]. The NLM estimation is effective in removing noises whenever non-local similarities are present among the samples of the signal under consideration. However, it suffers from the issue of under-averaging in those regions where amplitude and frequency variations are abrupt. Since speech is a non-stationary signal, the magnitude and frequency vary over the time. Consequently, NLM is not that effective in removing the noise components from the speech signal as observed in the case of image enhancement. To address this issue, the noisy speech signal is first decomposed into variational mode functions (VMFs) [1] using VMD. Each of the VMFs represents a small portion of the overall frequency components of the signal. The VMFs are then combined into different groups depending on their similarities to reduce computational cost. Next, the non-local similarity present in each group of VMFs is exploited for an effective speech enhancement through NLM estimation. The enhancement performance of the proposed method is compared with the existing speech enhancement techniques. The experimental results presented in this study show that, the proposed method provides better performance.

1. INTRODUCTION

Speech enhancement aims to improve speech quality by using various algorithms. The objective of enhancement is improvement in intelligibility and/or overall perceptual quality degraded speech signal using audio signal processing techniques. Enhancing of speech degraded by noise, or noise reduction, is the most important field of speech enhancement, and used for many applications such as mobile phones, teleconferencing systems, speech recognition, and hearing aids.

The suppression of noise components from speech signal to improve the quality and intelligibility is not only essential but also extremely challenging. Over the years, several approaches for speech enhancement have been reported. Most of the classical speech enhancement approaches are subtractive in nature. Then, the estimate of the noise spectrum is subtracted from the noisy speech spectrum to enhance the signal quality. The performance of such approaches is highly dependent on the accuracy with which the non-speech region is detected and robust estimation of instantaneous noise spectrum. Several techniques have been proposed for estimating the noise spectrum from the noisy speech signal. However, such spectral enhancement methods introduce distortion in the enhanced speech signal due to deviations in estimated and actual instantaneous noise spectrum. In the enhancement approaches presented in, the high signal to noise ratio (SNR) [6] regions are identified and relatively more enhanced compared to the low SNR regions.

The non-local means estimation, a well explored method for denoising image and electrocardiography (ECG) signals, is effective in removing the noises whenever non-local similarities are present among the samples of the signal. Since speech is a non-stationary signal, the magnitude and frequency vary over the time. Consequently, NLM is not that effective in removing the noise components from the speech signal as observed in the case of image and ECG enhancement.

This issue can be addressed up to an extent by decomposing the signal into different narrow-band regions. The VMD algorithm decomposes a signal into a predefined number of narrow-band variational mode functions (VMFs). Each of the VMFs represents some smaller portion of the overall frequency band of the signal. Unlike the noisy speech signal, the VMFs do not have abrupt amplitude and frequency variations. Through this motivation, a speech enhancement approach is proposed in this report by utilizing the efficacy of VMD and NLM estimation.

2. LITERATURE SURVEY

In frequency-domain speech enhancement, the magnitude spectrum of the speech signal is estimated and combined with the short-time phase of the degraded speech to produce the enhanced signal. The spectral subtraction algorithms and Wiener filtering are well-known examples. In the spectral subtraction algorithms [1,7], the STSA [1,2] of noisy and noise signals are estimated as the square root of maximum likelihood estimation of each signal spectral variance, and then subtracted from the magnitude spectrum of the noisy signal. In the Wiener filtering algorithm, the estimator is obtained by finding the optimal MMSE estimates [1,16] of the complex Fourier transform coefficients.

Both spectrum subtraction [24] and Wiener filtering algorithms are derived under Gaussian assumption for each spectral component. Spectrum subtraction and Wiener filtering are not optimal spectral amplitude estimators under the assumed Gaussian model; the spectral amplitude estimators are more advantageous than spectral estimators from a perceptual point of view. Ephraim and Malah formulated an optimal spectral amplitude estimator, which, specifically, estimates the modulus (magnitude)of each complex Fourier coefficient of the speech signal in a given analysis frame from the noisy speech in that frame.

In order to derive the MMSE STSA estimator, the a priori probability distribution of the speech and noise Fourier expansion coefficients should be assumed since these are unknown in reality. Ephraim-Malah [22] derived their spectral amplitude estimator based on a Gaussian model. They argued that this assumption utilizes asymptotic statistical properties of the Fourier coefficients.

Specifically, according to, they assumed that the Fourier expansion coefficients of each process can be modelled as statistically independent Gaussian random variables, real- and imaginary parts of each component is independent to each other, and the mean of each coefficient is assumed to be zero and the variance time-varying. The assumption is motivated by the central limit theorem. Central limit theorem/asymptotic statistical properties hold more strongly when the analysis frame size is long, which somewhat conflicts with the "short-time" requirement. Porter and Boll [21,24], Brehm and Stammler [28], and Martin [29] recognized that the DFT coefficients of clean speech derived from

speech frames having a length of about the span of correlation within the signal are not normally distributed, and might be better modelled by a Gamma or a Laplacian distribution.

The DFT frame length is in the range of 10-40 ms, which is a typical frame size in speech application.

At a global SNR of more than 40 dB, only a very low, almost stationary, level of quantization and ambient recording noise is present in these signals. Since speech signals are highly non-stationary, it must be sure that the spectral coefficients represented in the histogram are obtained under quasi-stationary conditions. Martin [26] tried to guarantee the stationarity as follows.

The a priori SNR can be used to select DFT coefficients with a consistent quasi stationary SNR, or, in the case of stationary background noise, with a fixed power.

That is, power is used as a criterion for stationarity. It was found that the Laplacian density provided a reasonable fit to the experimental data. Breithaupt and Martin [26] further verified that the Gamma pdf produced a better fit to experimental data than the Gaussian pdf.

The following will give more details about the methods mentioned above. Estimators based on Gamma and Laplacian models. In their approach, they aimed to compute the squared magnitude of the clean speech spectral coefficients by estimating the squared real and imaginary parts of the DFT respectively. In their method, the clean speech signal and noise were modelled by a combination of Gaussian, Gamma and Laplacian distributions, which will be detailed later. It is known that signal enhanced by the MMSE estimator of the power spectral density [26] suffers from musical noise. Martin proposed a new estimator, in which the real and imaginary parts of the clean signal were estimated in the MMSE sense conditional on the real and imaginary parts of the observed noisy signal.

The basic idea of Laplacian based MMSE-STSA estimator is of in the optimal estimate of the modulus of the speech signal DFT components in the MMSE sense, based on the assumption that the real and imaginary parts of these components are modelled by a Laplacian distribution. The noise signal DFT components are assumed to be a Gaussian distributed. Note that the assumption about noise DFT component is valid regardless of whether the noise is white or coloured.

One of the challenges in deriving such an estimator is to compute the pdf of spectral amplitude.

It is especially complicated when the parts of DFT coefficients modelled by Laplacian Distribution.

The Laplacian assumption makes the derivation in question even more difficult in that independence between amplitude and phase no longer holds with the Laplacian distribution.

This estimator, however, is not the optimal spectral amplitude estimator, which, we have pointed out is more preferable perceptually. Lotter and Vary addressed the issue of finding the optimal spectral amplitude estimator based on Gamma or Laplacian models. They found a closed form of the spectral amplitude estimator, but they had to employ a parametric function to approximate the pdf of the spectral amplitude since the analytic solution of the pdf was unknown to them.

The parametric function had a simpler form and facilitated their derivation but the authors did not specify how they obtained the parametric function, and the parameters needed to be determined empirically.

The MMSE estimation of speech spectrum has received considerable research attention. Although significant progress has been made in the above-mentioned methods, each of them has its own weaknesses and limitations either on the underlying assumptions or derivation of the estimators. The fact that Laplacian and Gamma are better than the Gaussian PDF for modelling the speech DFT coefficients has been recognized and empirically verified by many researchers. Therefore, many have focused on Laplacian/Gamma based estimators. None of them, however, has succeeded in finding the preferred optimal spectral amplitude estimator in the MMSE sense.

3. SPEECH ENHANCEMENT BY VMD-NLM METHOD

Speech enhancement approach in this report by utilizing the efficacy of VMD [10,11]and NLM [1] estimation. The block diagram summarizing in the method for speech enhancement is shown in the following figure.

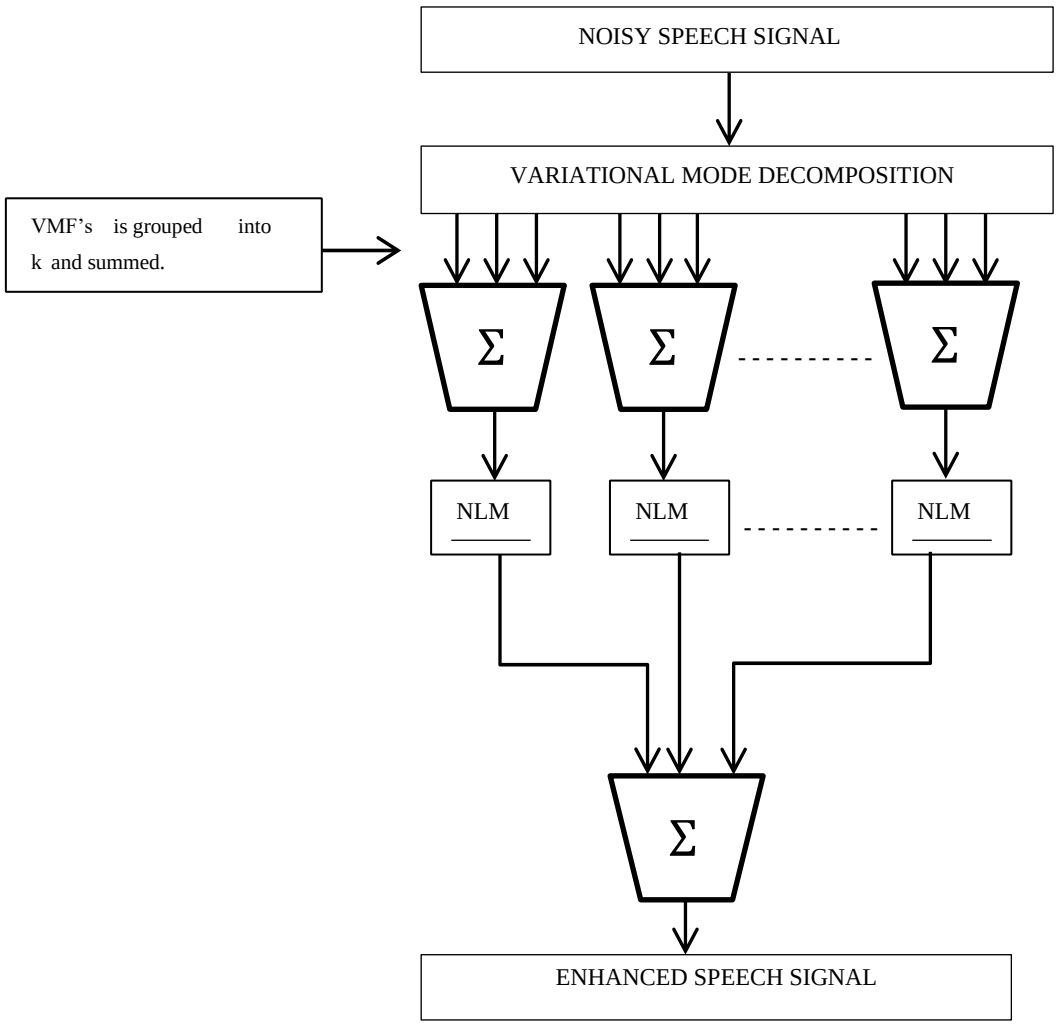

Figure 1: The above block diagram represents the proposed method for enhancing the speech signal.

3.1. VARIATIONAL MODE DECOMPOSITION

In this section we introduce our proposed model for variational mode decomposition, essentially based on the three concepts outlined in the previous section. The goal of VMD [11] is to decompose a real valued input signal into a discrete number of sub-signals (modes) that have specific sparsity properties while reproducing the input. Here, the sparsity prior of each mode is chosen to be its bandwidth in spectral domain.

i. The block diagram shows that the noisy signal is getting decomposed into a set of Variational Mode Function's (namely: - VMF_1, VMF_2, VMF_3 …VMF_K).
ii. The VMF's having lower centre frequency represents the high magnitude regions and the VMF's with higher centre frequency represents the low magnitude regions.
iii. The VMF's are divided into set of j groups based on the similarity in their centre frequencies. Now they are summed and NLM [1] estimation is performed in order to remove the noisiness in the speech signal. The groups are finally combined to obtain the enhanced speech signal.

In other words, we assume each mode k to be mostly compact around a centre pulsation ωk, which is to be determined along with the decomposition. In order to assess the bandwidth of a mode, we propose the following scheme**:**

i. For each mode ωk, compute the associated analytic signal by means of the Hilbert transform in order to obtain a unilateral frequency spectrum.
ii. For each mode, shift the mode's frequency spectrum to "baseband", by mixing with an exponential tuned to the respective estimated centre frequency.
iii. The bandwidth is now estimated through the H1 Gaussian smoothness of the demodulated signal i.e. the squared L2-norm of the gradient.

The resulting constrained variational problem is the following:

$$min_{\{u_k\},\{\omega_k\}}\left(\sum_k \left|\left|\partial_k\left[\left(\delta(t)+\frac{j}{\pi t}\right)* u_k(t)\right] e^{-j\omega_k t}\right|\right|_2^2\right) \qquad \text{———} \qquad (1)$$

Such that Pk vk(t) = s(t). Where, {vk} = {v1, v2 ...vk}, {ωk} = {ω1, ω2 ...ωk}, k, δ (t) and represents the VMFs (modes), the center frequencies for each of the VMFs, total number of modes, Dirac distribution and convolution operator, respectively.

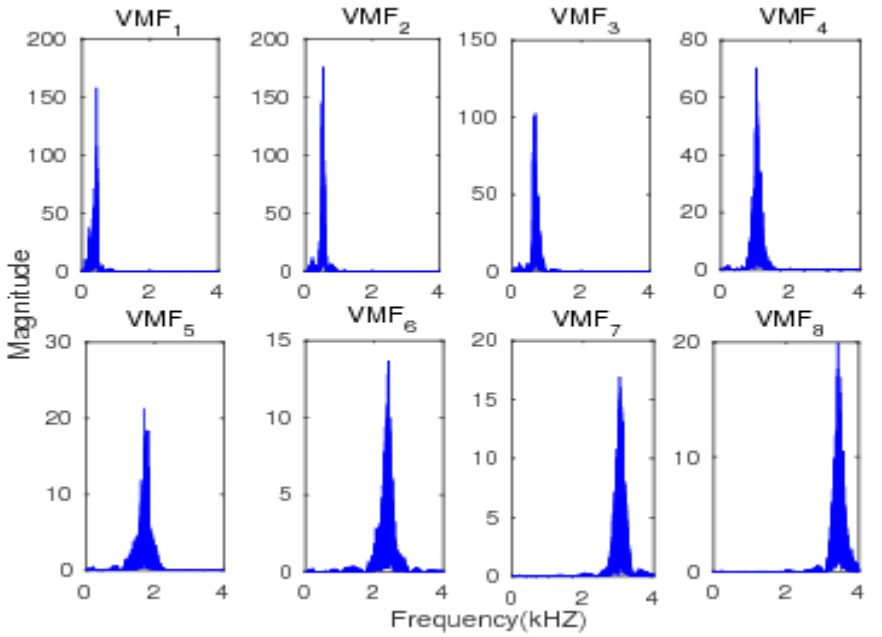

Figure 2: Magnitude spectrum of VMFs for a 0-dB white noise added speech signal. The modes are arranged from low- to high frequency band (left to right).

Variational method for decomposing a signal into an ensemble of band-limited intrinsic mode functions, that we call Variational Mode Decomposition, (VMD). In contrast to existing decomposition models, like the empirical mode decomposition (EMD), we refrain from modelling the individual modes as signals with explicit IMFs. Instead, we replace the most recent definition of IMFs, namely their characteristic description as AM-FM signals, by the corresponding narrow-band property.

If a large number of modes are selected for decomposition, under-binning of modes (loss of information) happens. On the other hand, lower number of modes results in over-binning of modes (mode duplication).

During the preliminary experiments performed on development set, it was observed that for effective decomposition and reconstruction of speech signal, a minimum of k = 12 levels of decomposition is required.

The magnitude spectra for the 12 VMFs derived from a 0dB white noise added speech signal are shown in Figure 2. The magnitude spectra shown from left to right in ascending order of VMFs.

It can be observed that, in the each of the VMFs, frequency and amplitude variations are very small. It can also be noted that, depending upon the similarities in the location of their center frequency and mean magnitude, some of the VMFs can be combined together.
For example, VMF−1 & VMF−2 can be combined to represent a single group. The VMFs are combined to reduce the computational cost for NLM estimation without loss in denoising capability. In this study, the VMFs are finally clustered into four groups.

3.2 NON-LOCAL MEANS ESTIMATION

The NLM [1] approach estimates the true signal from the noisy signal by exploiting the non-lo similarities among the sample points. In NLM filtering, for each sample point of the signal y(n), estimate $\hat{y}$(n) is computed as a weighted sum of the signal values at another sample point y(m). T final denoised signal is computed with the help of two local patches with starting points being n m, respectively. Both the patches consist of P samples and they lie within the search neighbourhood *M(n)*. The estimated denoised signal is computed.

$$\hat{y} = \frac{1}{r(n)} \sum_{m \in M(n)} r(n,m) y(m) \qquad \text{——————} \quad (2)$$

For each sample point, the mapping is decided by weight values r (n, m) that represent the non-lo similarity present in the neighbourhood with respect to the sample points y(n) and y(m), respective the weight value r (n, m) is computed.

$$r(n,m) = \exp\left(-\frac{\sum_{j=1}^{R}(R(n+j)-R(m+j)^2)}{2PA^2}\right) \qquad (3)$$

where, A represents the bandwidth parameter which controls the amount of smoothing to be apply to the denoised signal. The difference values are summed over P samples (length of the patch) normalized in order to get the weight value. (n) represents the normalized weight value at sample point in which, in turn, is computed as follows:

$$r(n) = \sum_{m \in M(n)} r(n,m) \qquad (4)$$

3.3 SPEECH ENHANCEMENT BY NLM ESTIMTION OF VMF'S

In the case of speech, the amplitude and the frequency change over the frames depending on the sound units. Therefore, the NLM is not effective in enhancing noisy speech signal. those variations are suppressed to a great extent by grouping the VMFs. The NLM estimation is performed on the signal obtained by adding the VMFs belonging to any particular group. The final reconstruction is done by adding each of the NLM estimated outputs as shown in Figure (1).

The effectiveness of the approach for speech enhancement is demonstrated in Figure (3). It is evident that, the fluctuations in each group of VMFs are very less. The NLM effectively removes the noise components from the VMFs. By comparing the original clean and enhanced speech signals, it is evident that the approach is very effective in removing the noise components from the given speech data. The below figure showing the NLM estimation of VMF.

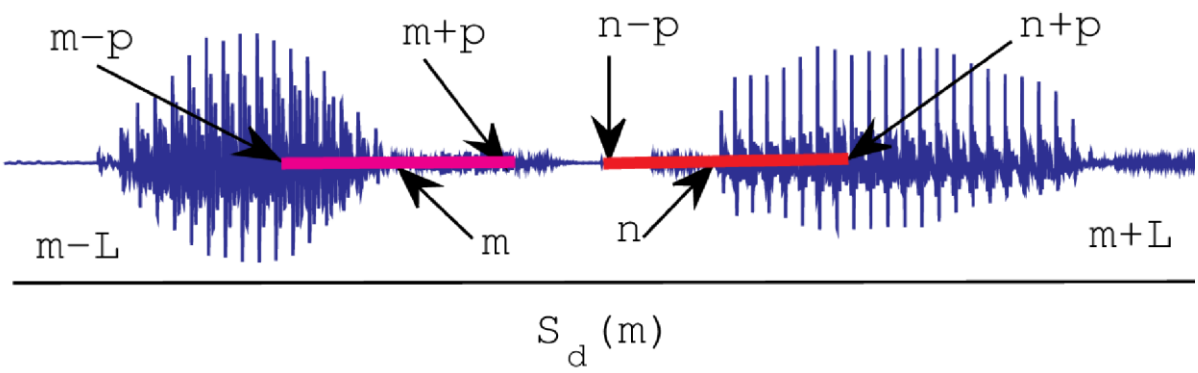

Figure 3: NLM estimation & construction of patches from neighbourhood having non-local similarities.

In this approach, the speech enhancement is performed by processing the noisy speech signal through the following steps:

i. The noisy speech signal is decomposed into k number of VMFs using VMD. The VMFs having lower centre frequency predominantly represents the high magnitude vowel-like regions where as the VMF having higher centre frequency represent the unvoiced sound units.
ii. Then, the VMFs are divided into j groups depending on the similarity in their centre frequencies and magnitude spectrum since those VMFs represents similar sound units.
iii. The VMFs in each group are summed and NLM estimation is performed to remove the noise components. The grouping of VMFs reduces the computational cost.
iv. Finally, the NLM estimated signals obtained from each of the groups are combined to obtain the enhanced signal.

3.4 METRICS FOR PERFORMANCE EVALUTION

These parameters provide an overview of the various methods and techniques used for assessment of speech quality and for vowel detection. Considerations for conducting successful subjective listening tests are given along with cautions that need to be exercised. While the listening tests are considered the gold standard in terms of assessment of speech quality, they can be costly and time consuming. For that reason, much research effort has been placed on devising objective measures that correlate highly

with subjective rating scores. An overview of some of the most commonly used objective measures is provided along with a discussion on how well they correlate with subjective listening tests.

Assessment of speech quality can be done using subjective listening tests or using objective quality measures. Subjective evaluation involves comparisons of original and processed speech signals by a group of listeners who are asked to rate the quality of speech along a pre-determined scale.

Objective evaluation involves a mathematical comparison of the original and processed speech signals. Objective measures quantify quality by measuring the numerical "distance" between the original and processed signals. Clearly, for the objective measure to be valid, it needs to correlate well with

subjective listening tests, and for that reason, much research has been focused on developing objective measures that modelled various aspects of the auditory system.

This section provides an overview of the various subjective and objective measures proposed in the literature for assessing the quality of processed speech. Quality is only one of many attributes of the speech signal. Intelligibility is a different attribute and the two are not equivalent. For that reason, different assessment methods are used to evaluate quality and intelligibility of processed speech. Quality is highly subjective in nature and it is difficult to evaluate reliably.

This is partly because individual listeners have different internal standards of what constitutes "good" or "poor" quality, resulting in large variability in rating scores among listeners. Quality measures assess "how" a speaker produces an utterance, and includes attributes such as "natural", "raspy", "hoarse", "scratchy", and so on. Quality is known to possess many dimensions, encompassing many attributes of the processed signal such as "naturalness", "clarity", "pleasantness", "brightness", etc.

For practical purposes we typically restrict ourselves to only a few dimensions of speech quality depending on the application. Intelligibility measures assess "what" the speaker said, i.e., the meaning or the content of the spoken words. In brief, speech quality and speech intelligibility are not synonymous terms, hence different methods need to be used to assess the quality and intelligibility of processed speech.

The distortions introduced, for instance, by waveform coders (e.g., ADPCM) operating at high bit rates (e.g., 16 kbps) differ from those introduced by linear-predictive based coders (e.g., CELP) operating at relatively lower bit rates (4-8 kbps). The distortions introduced by hearing aids include peak and centre

clipping, Automatic Gain Control (AGC), and output limiting. The AGC circuit itself introduces non-linear distortions dictated primarily by the values of attack and release time constants. Finally, the distortions introduced by the majority of speech enhancement algorithms depend on the background noise and the suppression function used (note that some enhancement algorithms cannot be expressed in terms of a suppression function).

Segmental SNR (segSNR): The segmental signal-to-noise ratio [1] can be evaluated either in the time or frequency domain. The time-domain measure is perhaps one of the simplest objective measures used to evaluate speech enhancement or speech coding algorithms. For this measure to be meaningful it is important that the original and processed signals be aligned in time and that any phase errors present be corrected.

The segmental signal-to-noise (SNRseg) is defined as:

$$SNRseg = \frac{10}{W}\sum_{w=0}^{W-1} log_{10}\frac{\sum_{n=Nw}^{Nw+N-1} x^2(n)}{\sum_{n=Nw}^{Nw+N-1}(x(n)-\hat{x}(n))^2} \qquad (5)$$

where x(n) is the original (clean) signal, $\hat{x}(n)$ is the enhanced signal, N is the frame length (typically chosen to be 15-20 ms), and W is the number of frames in the signal. Note that the SNRseg measure is based on the geometric mean of the SNRs across all frames of the speech signal.

One potential problem with the estimation of SNRseg [1] is that the signal energy during intervals of silence in the speech signal (which are abundant in conversational speech) will be very small resulting in large negative SNRseg values, which will bias the overall measure.

One way to remedy this is to exclude the silent frames from the sum in equation above by comparing short-time energy measurements against a threshold or by flooring the SNRseg values to a small value. In the SNRseg values were limited in the range of [-10 dB, 35 dB] thereby avoiding the need for a speech/silence detector. The segmental SNR can be extended in the frequency domain to produce the frequency-weighted segmental SNR.

Log-likelihood ratio (LLR): Several objective measures were proposed based on the dissimilarity between all pole models of the clean and enhanced speech signals. These measures assume that over short-time intervals speech can be represented by a p^{th} order all-pole model of the form. Perhaps two of the most common all-pole based measures used to evaluate speech-enhancement algorithms are the log

likelihood ratio and Itakura Saito measures. Cepstral distance measures derived from the LPC coefficients were also used.

The log-likelihood ratio (LLR) [1] measure is defined as

$$d_{LLR}(b_x, \bar{b}_{\hat{x}}) = \log \frac{\bar{b}_{\hat{x}}^{\,T} R_x \bar{b}_{\hat{x}}}{b_x^{\,T} R_x b_x} \qquad \text{(6)}$$

where $b_x^T = [1, -b_x(1), -b_x(2), ..., -b_x(P)]$ are the LPC coefficients of the clean signal

$\bar{b}_{\hat{x}}^{\,T} = [1, -b_{\hat{x}}(1), -b_{\hat{x}}(2), ..., -b_{\hat{x}}(p)]$ are the coefficients of the enhanced signal,

and R_x is the $(p + 1) * (p + 1)$ autocorrelation matrix (Toeplitz) of the clean signal.

Perceptual Evaluation of Speech Quality (PESQ) Measure: The original (clean) and degraded signals are first level equalized to a standard listening level, and filtered by a filter with response similar to a standard telephone handset. The signals are aligned in time to correct for time delays, and then processed through an auditory transform, similar to that of BSD [1], to obtain the loudness spectra. The absolute difference between the degraded and original loudness spectra is used as a measure of audible error in the next stage of PESQ [1] computation. Note that unlike most objective measures (e.g., the BSD measure) which treat positive and negative loudness differences the same (by squaring the difference), the PESQ measure treats these differences differently. This is because positive and negative loudness differences affect the perceived quality differently.

A positive difference would indicate that a component, such as noise, has been added to the spectrum, while a negative difference would indicate that a spectral component has been omitted or heavily attenuated. Compared to additive components, the omitted components are not as easily perceived due to masking effects, leading to a less objectionable form of distortion. Consequently, different weights are applied to positive and negative differences. The differences, termed the disturbances, between the loudness spectra is computed and averaged over time and frequency to produce the prediction of subjective MOS score. The final PESQ [1] score is computed as a linear combination of the average disturbance value.

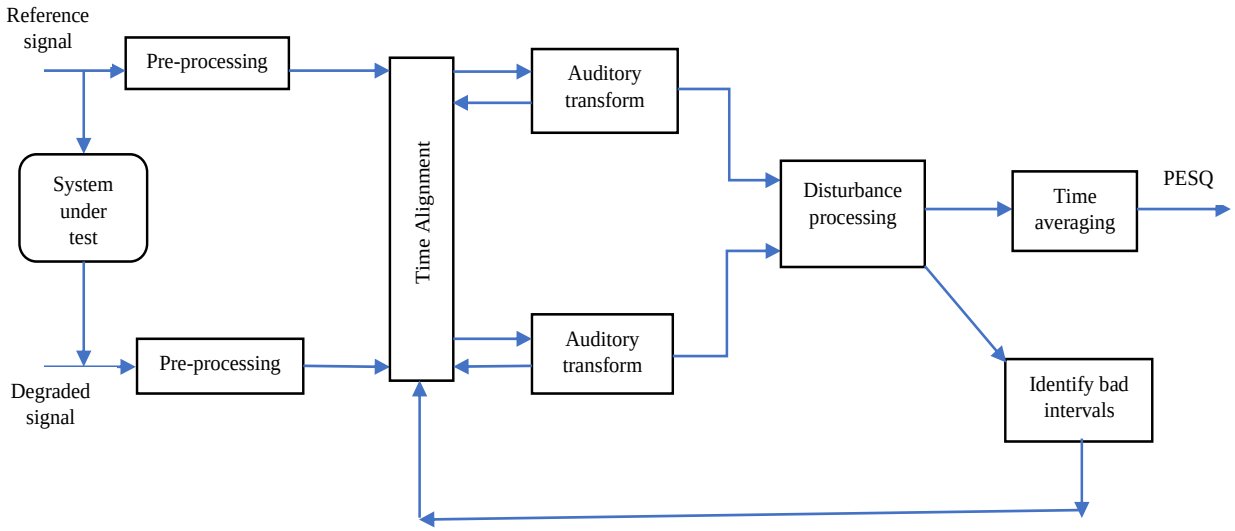

Figure 4: Block diagram of the PESQ measure computation.

*Weighted Spectral Slope (WSS) Distance Measure :*Psychoacoustic studies indicated that when subjects were asked to rate the phonetic distance between synthetic vowels subjected to several spectral manipulations (including low-pass, high-pass filtering, level change, formant frequency differences, etc.), subjects assigned the largest distance to pairs of vowels that differed in formant frequencies. Motivated by this finding, Klatt proposed a measure based on weighted differences between the spectral slopes in each band.

This WSS measure [1] was designed to penalize heavily differences in spectral peak (formants) locations while ignoring other differences between the spectra such as spectral tilt, overall level, etc. Those differences were found to have little effect on ratings of phonetic distance between pairs of synthetic vowels.

This measure is computed by first finding the spectral slope of each band. Let V_x (k) be the original (clean) and $\overline{V_{\hat{x}}}$ (k) the enhanced critical-band spectra, respectively, expressed in dB. A first-order differencing operation is used to compute the spectral slopes as follows:

$$P_x(k) = V_x(k+1) - V_x(k)$$
$$\overline{P_{\hat{x}}}(k) = \overline{V_{\hat{x}}}(k+1) - \overline{V_{\hat{x}}}(k) \qquad (7)$$

where $P_x(k)$ and $\overline{P_{\hat{x}}}(k)$ denote the spectral slopes of the clean and enhanced signals, respectively, of the kth band. The differences in spectral slopes are then weighted according to, first, whether the band is near a spectral peak or valley and, second, according to whether the peak is the largest peak in the spectrum. The weight for bank k, denoted as S(k), is computed as follows:

$$S(k) = \frac{K_{max}}{[K_{max} + V_{max} - V_x(k)]} \frac{K_{loc\ max}}{[K_{loc\ max} + V_{loc\ max} - V_x(k)]} \quad \text{——— (8)}$$

where V_{max} is the largest log-spectral magnitude among all bands, $V_{loc\ max}$ is the value of the peak nearest to band k, and K_{max}, $K_{loc\ max}$ are constants which can be adjusted using regression analysis to maximize the correlation between the subjective listening tests and values of the objective measure. For the experiments in a high correlation was found with K_{max} = 20 and $K_{loc\ max}$= 1.

The WSS measure [1] is finally computed for each frame of speech as:

$$d_{WSS}(V_x, \overline{V_x}) = \sum_{k=1}^{M} S(k)\,(P_x(k) - \overline{P_{\hat{x}}}(k))^2 \quad \text{——— (9)}$$

where M is the number of critical bands used.

The WSS measure [1] is attractive because it does not require explicit formant extraction. Yet, it attends to spectral peak locations and is insensitive to relative peak heights and to spectral details [4] in the valleys. The LPC-based measures [8] (e.g., LLR measure) are sensitive to formant frequency differences, but are also quite sensitive to changes in formant amplitude and spectral tilt changes. It was no surprise that the WSS measure yielded a higher correlation (p = 0.74) than the LPC measures, with subjective quality ratings of speech degraded by speech coding distortions.

This section presented various techniques and procedures that have been used to evaluate the performance of speech enhancement algorithms. Enhancement algorithms can be evaluated in terms of speech intelligibility and speech quality.

Several intelligibility tests were described based on the use of nonsense syllables, monosyllabic words, rhyming words and sentences as speech materials. Several subjective listening tests were also described for evaluating the quality of enhanced speech.

4. VOWEL DETECTION USING THE ENHANCED SIGNAL

Vowels are the most prominent sound units in a speech sequence. Automatic detection of vowels has numerous applications in different domains of the speech signal processing [1]. In earlier reported works, vowels and vowel onset points (VOPs) [36] have been used for the tasks such as speaker recognition [37,38], speech synthesis [1], prosody modification analysis of shouted and laughter speech [30,31] and keyword spotting [45]. The performance of those tasks critically depends on an accurate detection of vowels. In a continuous speech signal, the characteristics of the vowels like duration, average energy, magnitude of transitions at vowel onset point (VOP) and vowel end point (VEP) vary with the context of spoken utterance as well as the environmental conditions [50,55]. Consequently, detection of vowels in noisy speech signal is not only an important but also a very challenging task. In a given speech signal, the vowels are near-periodic, high energy and longer duration sound units [41,45]. Considering these aspects, several front-end speech parameterization techniques capturing the nature of excitation and vocal tract system response have been reported for the detection of vowels and VOPs [36,45**,**47].Further, different acoustic modeling techniques have also been exploited to model the transition at the VOPs as well as the distinct signal characteristics of the vowels [16,31]. The front-end features employed for these tasks include the difference in energy of each of the peaks and their corresponding valleys in the magnitude spectrum [13], zero-crossing rate, energy and pitch information of the speech signal [49], wavelet scaling coefficients of the speech signal[50],Hilbert envelop of the linear prediction(LP)residual[30], spectral peaks, modulation spectrum energies[29],spectral energy present in the glottal closure regions [45], uniformity of the epoch intervals in vowels [47] and the cumulative sum of the short-term discrete Fourier transformed (DFT) magnitude spectrum of non-local means (NLM) estimated speech signal [17]. Several vocal tract and excitation source features have also been combined to represent the complementary information present in the vowels [29,47]. In other group of approaches, statistical modeling methods like hierarchical neural network, multi-layer feed-forward neural network and auto-associative neural

network have been used [31,49]. These models are generally trained on the features estimated from the speech frames around the VOPs.

Recently, hidden Markov model (HMM) was explored for the detection of vowels, VOPs and VEPs [45].

Considering these distinct properties of the vowel regions different method are proposed in the literature for detection of VOPs and VEPs, and used for different speech processing tasks.

Steps for vowel detection after the enhancement of noisy signal:

i. First, apply NLM on noisy speech signal to get an enhanced signal, use patch-half-width and the half neighbourhood-width as 2 ms and 20 ms, respectively.
ii. The enhanced signal is then processed in short-time frames to compute short-term magnitude spectrum.
iii. The cumulative sum of the magnitude spectrum is computed next and smoothed over 50 ms duration with a frame-shift of 1 ms.
iv. To enhance the relatively low energy vowel regions, the smoothed magnitude spectrum is processed through sigmoidal function.
v. The VOP evidence from the feature is obtained by convolving it with a first order Gaussian differentiator (FOGD) window of length 100 ms and standard deviation being one sixth of the window [3].

Analysis of detected vowel regions

- *Identification rate (IR):* The percentage of reference vowel samples that exactly match with the detected vowel samples.
- *Spurious rate (SR):* The percentage of detected vowels that lie outside the reference vowels.

Speech type	Method	IR in %	SR in %		
			Semi-vowel	nasal	Others
Clean Speech	SE-GCI	66.78	10.82	1.48	7.65
	Proposed	85.30	13.55	1.39	3.89

Table 1: Performances of the explored/proposed approach for detecting vowels in a given speech signal under clean test conditions

5. EXPERIMENTAL RESULTS AND DISCUSSION

We have applied 8-level decomposition of noisy speech signal using VMD technique. For VMD, the data fidelity constraint balancing parameter was set 320; time-step was 0 while tolerance of convergence was selected as 10^{-7}. The NLM estimation is dependent on proper selection of some tuneable parameters like patch size (P), search neighbourhood size M(n), and bandwidth parameter (A). In this study, the value of P, M(n) and A are selected as 10, 200 and 0.4σ, respectively on first group VMFs. Similarly, P, N(n) and A are selected as 10, 100 and 0.6σ on second group. For third and fourth groups those parameters are selected as 20, 80 and 0.8σ, respectively.

Where σ represents the standard deviation of the summed signal of respective group of VMFs. All the tuneable parameter values were selected empirically.

In order to evaluate the efficacy of the existing and proposed approaches, speech signals from the TIMIT database were used. A set of 10 speech utterances from speakers was used for experimental evaluations. The clean speech files were corrupted by adding white noise at three different levels of signal to noise ratios (0, 5, and 10dB). These non-stationary background noise sources were obtained from the Noisex-92 database.

The following objective speech quality measures were used for evaluating the performance: perceptual evaluation of speech quality (PESQ), segmental signal to noise ratio (segSNR), log likelihood ratio (LLR), weighted spectral scope (WSS). The results of the experimental evaluations are given in Table 1. Compared to the existing approaches, the proposed speech enhancement technique is expected to result in better segSNR and PESQ values especially for low SNR values (i.e., 0 and 5 dB). Consistent improvements are noted for all the three noise types explored in this study. The best-case performances are presented in boldface to highlight the same. Expect for 10dB white noise case, the proposed approach is expected to be significantly better.

Techniques		SNR (dB)	LLR	SNRseg	WSS	PESQ
		0	0.6762	-5.4671	46.8837	1.0766
mms_mmse_spzc_snru		5	0.5460	-5.4961	36.0635	1.3026
	F	10	0.4346	-5.5020	27.2988	1.4943
	A	15	0.3422	-5.5022	20.8654	1.6695
	C					
	T	0	0.7094	-5.3243	47.0421	1.1023
mms_smpo	O	5	0.5887	-5.3412	36.5441	1.3018
	R	10	0.5029	-5.3822	28.5345	1.4509
	Y	15	0.4467	-5.4030	24.2364	1.5366
		0	0.7042	2.5375	35.0857	1.3180
ics	N	5	0.4757	3.6153	27.1218	1.5097
	O	10	0.3148	4.9376	20.0032	1.7164
	I	15	0.1845	6.4363	14.8890	1.9138
	S					
	E	0	0.8803	-5.5556	48.2617	0.8829
kalman		5	0.8508	-5.5555	38.8897	1.1504
		10	0.7877	-5.5373	29.6399	1.3079
		15	0.6860	-5.2448	22.7929	1.3265
		0	1.2989	-1.1585	83.0606	0.3172
weiner		5	1.3105	-0.9271	77.9919	0.3085
		10	1.3245	-0.7337	73.5915	0.2799
		15	1.3290	-0.4592	70.6374	0.2600
		0	1.1611	0.2558	32.2146	0.7693
nlm		5	1.1014	0.3686	27.8413	1.0075
		10	1.0596	0.4868	24.3258	1.2222
		15	1.1048	0.5789	22.3984	1.3738

Table 2: Values of noise parameters after enhancement using factory noise of SNR (0,5,10,15) dB.

.

Techniques		SNR (dB)	LLR	SNRseg	WSS	PESQ
		0	0.1600	-5.5468	23.4582	1.9229
mms_mmse_spzc_snru		5	0.1512	-5.5526	17.2508	2.0680
		10	0.1499	-5.5519	12.5906	2.1845
	V	15	0.1504	-5.5488	9.4747	2.2633
	O					
	L	0	0.3693	-5.4731	27.9305	1.6231
mms_smpo	V	5	0.3496	-5.4603	23.2281	1.6613
	O	10	0.3307	-5.4317	19.8876	1.6871
		15	0.3163	-5.3917	17.4138	1.6973
	N	0	0.1486	8.3593	9.2850	2.1395
ics	O	5	0.1139	10.0614	6.7089	2.2470
	I	10	0.0876	11.6704	4.8805	2.3330
	S	15	0.0622	12.7644	3.6643	2.3970
	E					
		0	0.7422	-5.5502	35.0734	1.5276
kalman		5	0.7595	-5.4928	31.6560	1.4661
		10	0.7275	-5.1360	28.3979	1.3404
		15	0.7305	-4.1137	24.5483	1.2468
		0	1.2966	-1.0856	80.3836	0.4595
weiner		5	1.2960	-0.8705	75.8040	0.3373
		10	1.3069	-0.6682	73.0403	0.3933
		15	1.3141	-0.4083	71.3118	0.4228
		0	2.6972	0.2683	37.0802	1.2836
nlm		5	2.3411	0.3545	31.2247	1.3851
		10	2.0726	0.4495	28.0676	1.4323
		15	1.8606	0.5372	26.8546	1.4324

Table 3: Values of noise parameters after enhancement using volvo noise of SNR (0,5,10,15) dB.

Techniques		SNR (dB)	LLR	SNRseg	WSS	PESQ
		0	0.7448	-5.5098	29.6239	1.3120
mms_mmse_spzc_snru		5	0.6028	-5.4821	23.3782	1.5265
	W	10	0.4759	-5.4574	18.4506	1.6963
	H	15	0.3624	-5.4588	14.5168	1.8428
	I					
	T	0	0.6931	-5.4623	35.0988	1.3324
mms_smpo	E	5	0.5842	-5.4374	30.1396	1.4659
		10	0.4968	-5.4287	25.6964	1.5502
		15	0.4263	-5.4021	22.1591	1.6005
	N					
	O	0	0.8182	3.1560	29.6253	1.4239
ics	I	5	0.6352	4.2366	24.2246	1.6245
	S	10	0.5027	5.4851	19.1063	1.8230
	E	15	0.3644	6.8315	13.7504	2.0059
		0	0.9928	-5.5556	44.4095	0.9621
kalman		5	0.7068	-5.5556	36.7797	1.2396
		10	0.5770	-5.5556	30.0636	1.5700
		15	0.5462	-5.5556	23.5653	1.7443
		0	1.4333	-1.2368	78.9025	0.3257
weiner		5	1.4338	-0.9789	75.4640	0.3010
		10	1.4346	-0.7944	72.9174	0.2751
		15	1.4148	-0.5433	71.0224	0.3023
		0	0.8742	0.1987	46.5939	1.1572
nlm		5	0.9869	0.2817	42.8506	1.2312
		10	1.0441	0.3664	39.0101	1.2947
		15	1.1130	0.4510	35.1788	1.4357

Table 4: Values of noise parameters after enhancement using white noise of SNR (0,5,10,15) dB.

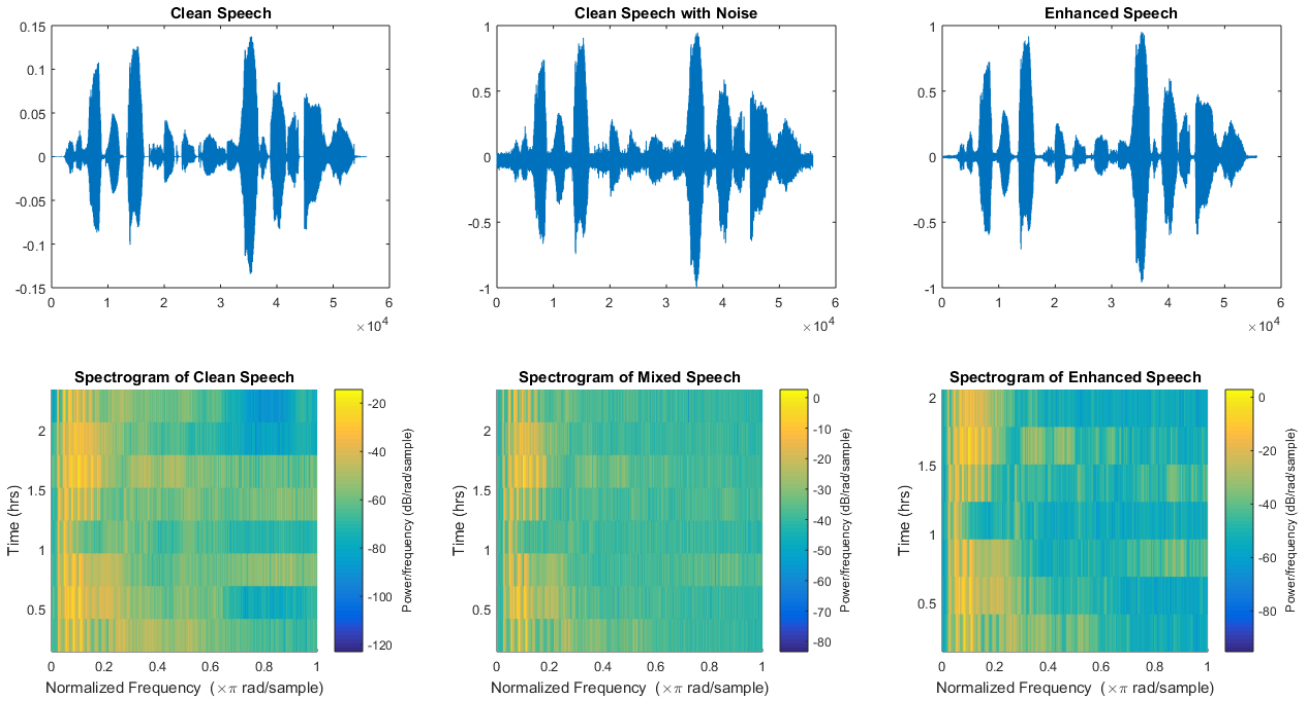

Fig 4: Plot of Clean Speech, Clean Speech with Noise, Enhanced Speech & their Spectrograms of MMS-MMSE-SPZC technique with White noise of SNR (10dB).

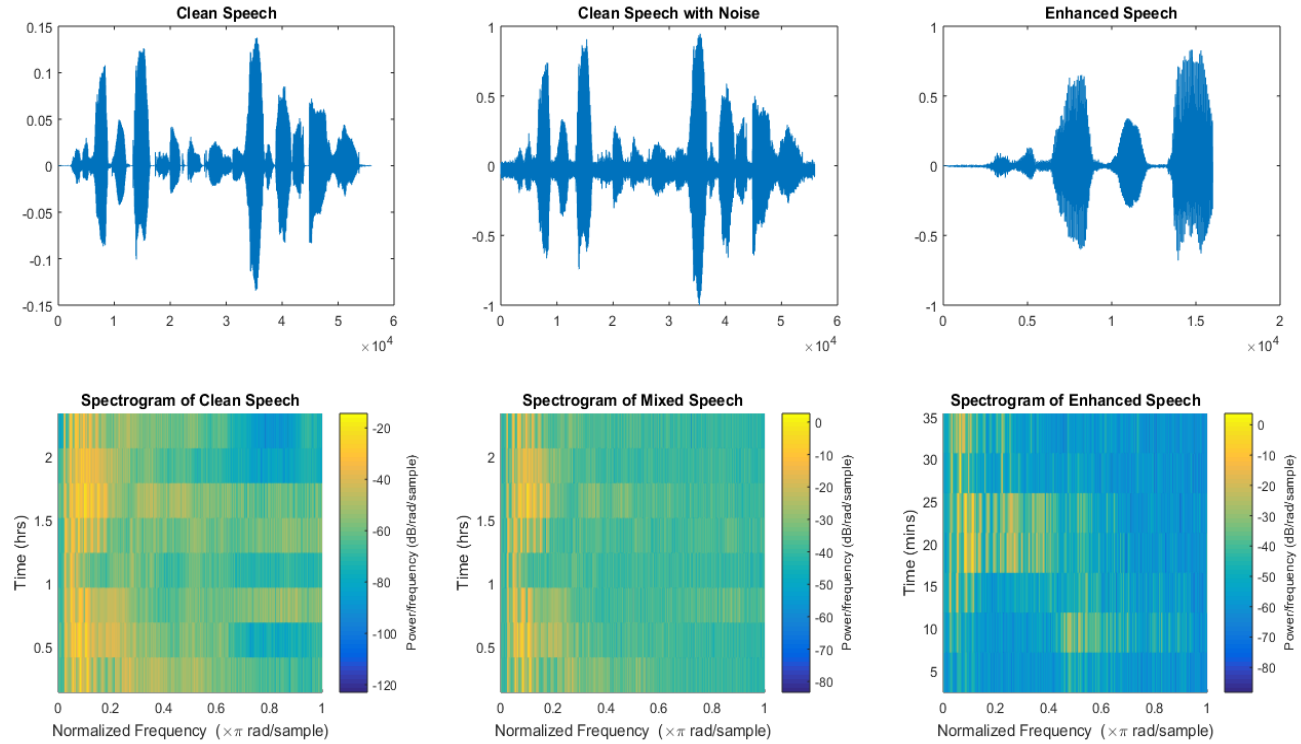

Fig 5: Plot of Clean Speech, Clean Speech with Noise, Enhanced Speech & their Spectrograms of MMS-SMPO technique with White noise of SNR (10dB).

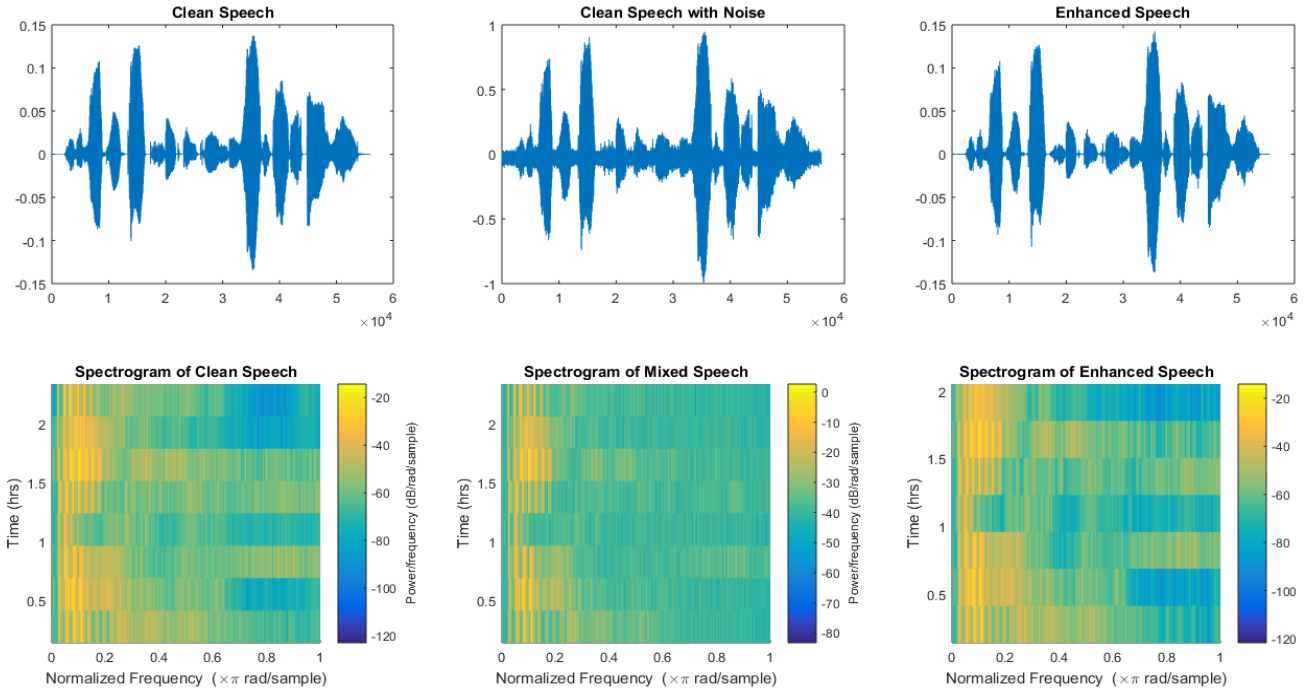

Fig 6: Plot of Clean Speech, Clean Speech with Noise, Enhanced Speech & their Spectrograms of ICS technique with White noise of SNR (10dB).

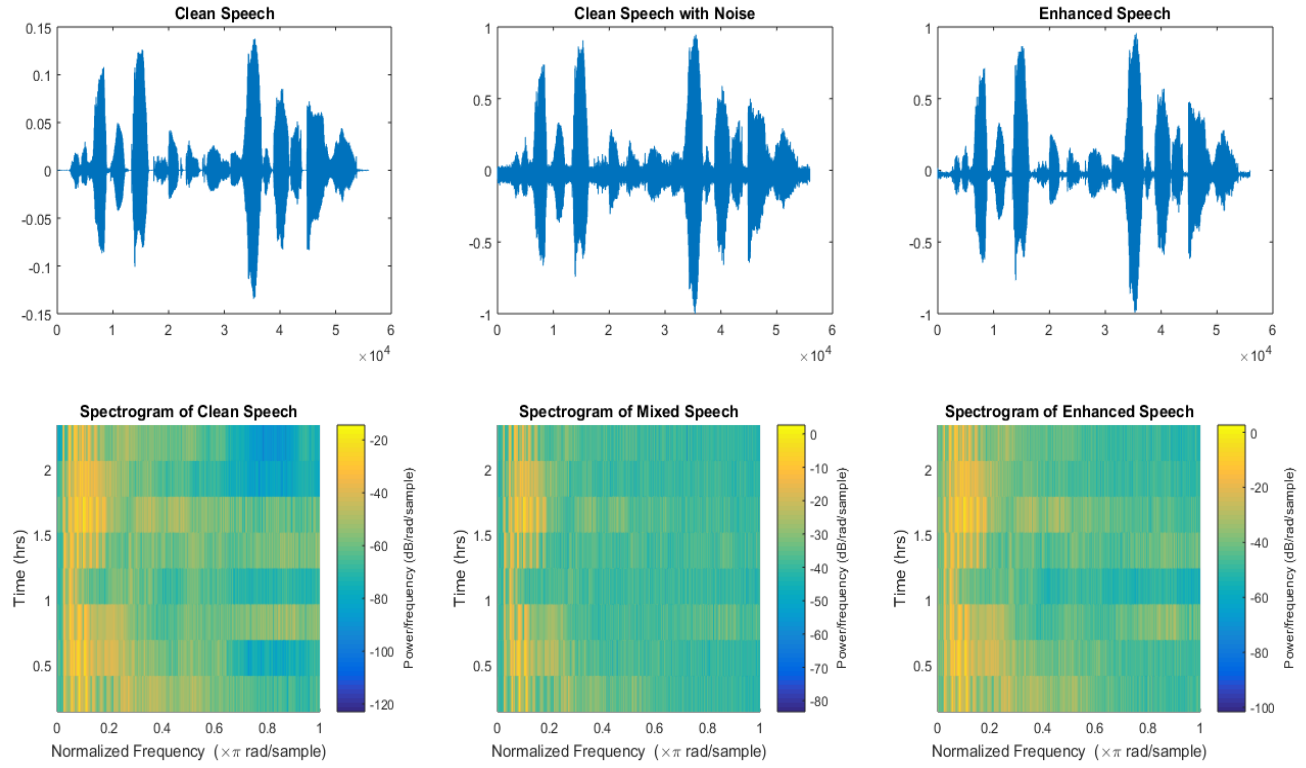

Fig 7: Plot of Clean Speech, Clean Speech with Noise, Enhanced Speech & their Spectrograms of Kalman filter technique with White noise of SNR (10dB).

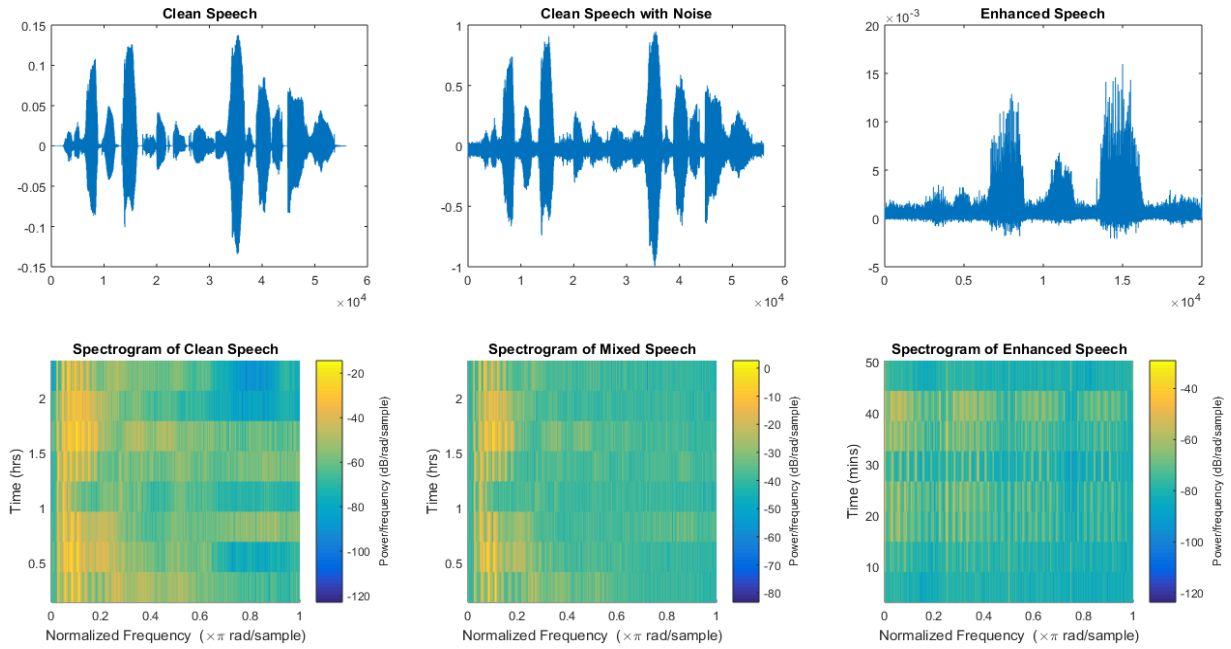

Fig 8: Plot of Clean Speech, Clean Speech with Noise, Enhanced Speech & their Spectrograms of Weiner filter technique with White noise of SNR (10dB).

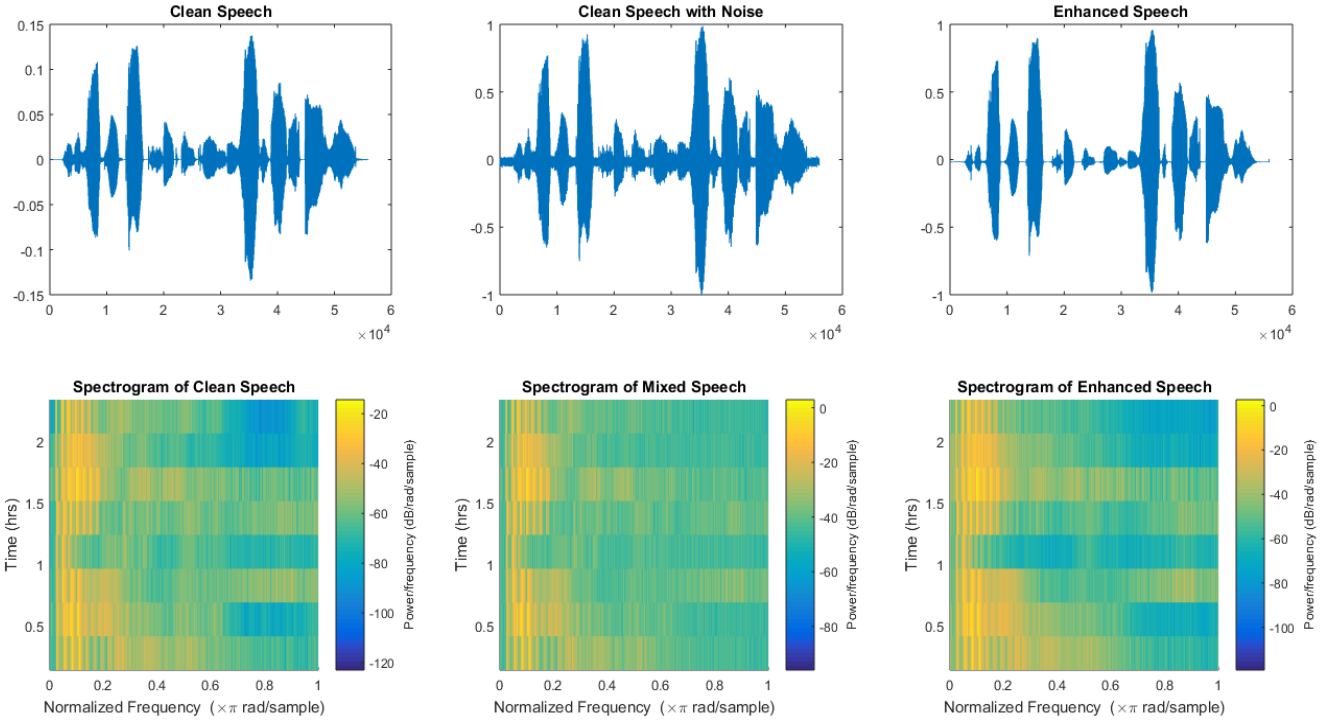

Fig 9: Plot of Clean Speech, Clean Speech with Noise, Enhanced Speech & their Spectrograms of NLM technique with White noise of SNR (10dB)

6. CONCLUSION AND FUTURE SCOPE

In this report, a two-stage VMD-NLM based speech enhancement technique has been explored. The noisy speech signal is first decomposed into 8 VMFs using the VMD algorithm. Next, based on the similarities in the location of centre frequencies and the mean amplitudes, the VMFs are clustered and summed to yield a set of four VMFs.

This step reduces the overall computational cost. The so obtained VMFs are then processed through NLM estimation in order to effectively reduce the ill-effects of interfering noises. This approach is compared with two of the recently developed speech enhancement techniques in terms of objective speech quality measures like segSNR and PESQ. The noise applied at different SNR levels are used for experimental evaluation. The proposed speech enhancement approach is expected to be better than the explored methods.

Adaptive algorithms found effective for such cases and are utilized in this problem. The work is based on decomposition method using variational mode decomposition (VMD) technique, where the decomposed components signify the frequency characteristics of the signal. Since Wiener filtering is used in VMD inherently, it can be modified with the least mean squares (LMS) adaptive algorithm for good accuracy and adaptability in this work.

Different noises like Babble noise, Street noise, and Exhibition noise can be considered and the corresponding signals can be decomposed into set of intrinsic mode functions (IMFs). Basically, the lower modes are of high frequency and noisy; whereas the higher mode IMFs contain the low and medium frequency components and are considered as the enhanced signal. The results of the proposed algorithm are found excellent as compared to earlier techniques. The resultant wave forms are visually observed and the sound is verified for audible range. Also, different measuring parameters are considered for its performance measure.

It is measured in terms of signal-to-noise ratio (SNR), segmental signal to noise ratio (SegSNR), perceptual evaluation of speech quality (PESQ) and log spectral distance (LSD). The technique is verified with standard database NOIZEUS for 0, 5, 10,15 dB respectively and also in real world case.

Also, we are looking forward to work with other techniques such as adaptive variational mode decomposition which is a bit different from the decomposition method used in this work, and EMD

Wavelet method of denoising signals which includes the basic idea that the energy of a signal will often be concentrated in a few coefficients in wavelet domain while the energy of noise is spread among all coefficients in wavelet domain as the hard and soft thresholding methods for denoising, where the former leaves the magnitudes of coefficients unchanged if they are larger than a given threshold, while the latter just shrinks them to zero by the threshold value. We will propose a noise robust feature in coming days.

7. REFERENCES

[1] P. C. Loizou, Speech enhancement: theory and practice. CRC press, 2013.

[2] J. Li, L. Deng,Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, 2014.

[3] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 5, pp. 1711–1723, 2007.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on acoustics, speech, and signal processing, vol. 27, no. 2, pp. 113–120, 1979.

[5] Y. Lu and P. C. Loizou, "Estimators of the magnitude-squared spectrum and methods for incorporating snr uncertainty," IEEE transactions on audio, speech, and language processing, vol. 19, no. 5, pp. 1123–1137, 2011.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Transactions on speech and audio processing, vol. 9, no. 5, pp. 504– 512, 2001.

[7] R. Tavres and R. Coelho, "Speech enhancement with non-stationary acoustic noise detection in time domain," IEEE Signal Processing Letters, vol. 23, no. 1, pp. 6–10, 2016.

[8] B. Yegnanarayana, C. Avendano, H. Hermansky, and P.S. Murthy, "Speech enhancement using linear prediction residual," Speech Communication, vol. 28, no. 1, pp. 25–42, may 1999.

[9] K. Deepak and S. M. Prasanna, "Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1205–1219, 2016.

[10] N. Chatlani and J. J. Soraghan, "EMD-based filtering (EMDF)lio of low-frequency noise for speech enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1158–1166, 2012.

[11] K. Khaldi, A.-O. Boudraa, andA.Komaty, “Speech enhancement using empirical mode decomposition and the teager–kaiserenergy operator,” The Journal of the Acoustical Society of America, vol. 135, no. 1, pp. 451–459, 2014.

[12] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” IEEE Trans. Acoustics, Speech and Signal Processing. 28, pp. 137—145, Apr. 1980.

[13] M. S. Ahmed, “Comparison of noisy speech enhancement algorithms in terms of lpc perturbation,” IEEE Trans. Acoustics, Speech and Signal Processing, vol. 37, pp. 121—125, 1989.

[14] H. R. Abutalebi and H. Sheikhzadeh, “A hybrid sub band adaptive system for speech enhancement in diffuse noise fields,” IEEE Signal Processing Letters, vol. 17, pp. 270—273, 2004.

[15] P. Wolfe and S. Godsill, “Simple alternatives to the ephraim and malah suppression rule for speech enhancement,” Proc. IEEE Workshop on Statistical Signal Processing, vol. 11th, pp. 496—499, Aug. 2001.

[16] R. Martin, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimation,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 504—512, 2002.

[17] D. R. Brillinger, Time Series: Data Analysis and Theory. San Francisco: Holden Day, Inc., 1981.

[18] J. Porterand S. Boll, “Optimal estimators for spectral restoration of noisy speech, “in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 18A.2.1—18A.2.4, 1984.

[19] C. Breithaupt and R. Martin, “Mmse estimation of magnitude-squared dft coefficients with super gaussian priors,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 848—851, 2003.

[20] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 208—211, 1979.

[21] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-27, pp. 113—120, 1979.

[22] J. M. Tribolet and R. E. Crochiere, “Frequency domain coding of speech,” IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-27, p. 522, Oct. 1979.

[23] R. Zelinski and P. Noll, “Adaptive transform coding of speech signals,” IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-25, p. 306, Aug. 1977.

[24] W. B. Davenport and W. L. Root, An Introduction to the theory of random signals and noise. New York: McGraw-Hill, 1958.

[25] H. Brehm and W. Stammler, “Description and generation of spherically invariant speech-model signals,” Signal Processing, vol. 12, pp. 119—141, 1987.

[26] R. Martin, “Speech enhancement based on minimum mean-square estimation and super gaussian priors,” IEEE Trans. Acoustics, Speech and Signal Processing. 13, pp. 845—856, Sept. 2005.

[27] S. Kullback, Information theory and statistics. Dover Publication, 1958.

[28] R. Martin and C. Breithaupt, “Speech enhancement in the dft domain using Laplacian speech priors,” in International Workshop on Acoustic Echo and Noise Control (IWAENC), pp. 87–90, Sept. 2003.

[29] T. Lotter and P. Vary, “Noise reduction by maximum a posteriori spectral amplitude estimation with super gaussian speech modelling,” in Proc. Int. Workshop Acoustic Echo and Noise Control, pp. 83—86, 2003.

[30] K. E. Mueller, “Computing the confluent hypergeometric function, m (a, b, x),” Numer. Math., vol. 90, pp. 179—196, 2001.

[31] S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series and Products. Academic Press, 2000.

[32] G. Pradhan, A. Kumar, S. Shahnawazuddin, Excitation source features for improving the detection of vowel onset and offset points in a speech sequence, in Proceedings of the Interspeech 2017, pp. 1884–1888,2017.

[33] G. Pradhan, S.M. Prasanna, Speaker verification by vowel and non-vowel like segmentation. IEEE Trans. Audio Speech Lang. Process 21(4), 854–867,2013.

[34] S.M. Prasanna, G. Pradhan, Significance of vowel-like regions for speaker verification under degraded conditions. IEEE Trans. Audio Speech Lang. Process 19(8), 2552–2565,2011.

[35] S.R.M. Prasanna, B.V.S. Reddy, P. Krishnamoorthy, Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Trans. Audio Speech Lang. Process 17(4), 556–565,2009.

[36] S.R.M. Prasanna, B. Yegnanarayana, Detection of vowel onset point events using excitation source information, in Proceedings of the Interspeech, pp. 1133–1136 ,2005.

[37] Jao, C.C. Sekhar, B. Yegnanarayana, Neural network-based approach for detection of vowel onset points, in Proceedings of the International Conference on Advanced Pattern Recognition Digital Technologies, vol. 1, pp. 316–320,1999.

[38] K.S. Rao, A.K. Vuppala, Speaker Identification and Time Scale Modification Using VOPs (Springer, Berlin), 2014.

[39] K.S. Rao, B. Yegnanarayana, Duration modification using glottal closure instants and vowel onset points. Speech Commun. 51(12), 1263–1269,2009.

[40] S.P. Rath, D. Povey, K. Veselý, J. Cernocký, Improved feature process for deep neural networks, in Proceedings of the Interspeech,2013.

[41] B.S. Reddy, K.V. Rao, S.M. Prasanna, Keyword spotting using vowel onset point, vector quantization and hidden Markov modeling based techniques, in Proceedings of the TENCON, pp. 1–4,2008.

[42] B. Sarma, S. Prajwal, S.M. Prasanna, Improved vowel onset and offset points detection using bessel features, in Proceedings of the International Conference on Signal Processing and Communication, pp. 1–6,2014.

[43] P. Singh, G. Pradhan, Exploring the non-local similarity present in variational mode functions for effective ecg denoising, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 861–865.2011).

[44] P. Singh, G. Pradhan, Variational mode decomposition based ecg denoising using non-local means and wavelet domain filtering. Australas. Phys. Eng. Sci. Med. 41, 1–14 ,2018.

[45] P. Singh, S. Shahnawazuddin, G. Pradhan, an efficient ecg denoising technique based on non-local means estimation and modified empirical mode decomposition. Circuits Syst. Signal Process. 37(10), 4527–4547,2018.

[46] N. Srinivas, G. Pradhan, Shahnawazuddin, Enhancement of noisy speech signal by non-local means estimation of variational mode functions. Proc. Interspeech 2018, 1156–1160,2018.

[47] K.N. Stevens, Acoustic Phonetics (The MIT Press Cambridge, London), 2000.

[48] B.H. Tracey, E.L. Miller, Nonlocal means denoising of ecg signals. IEEE Trans. Biomed. Eng. 59(9), 2383–2386,2012.

[49] D. VanDeVille, M. Kocher, Sure-based non-local means. IEEE Signal Process.Lett.16(11),973–976,2009.

[50] Alvera, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12(3), 247–251,1993.

[51] A.K. Vuppala, K.S. Rao, Vowel on set point detection for noisy speech using spectral energy at formant frequencies. Int. J. Speech Technol. 16(2), 229–235,2013.

[52] A.K. Vuppala, K.S. Rao, S. Chakrabarti, Improved vowel onset point detection using epoch intervals. AEU Int. J. Electron. Commun. 66(8), 697–700,2012.

[53] H.K. Vydana, S.R. Kadiri, A.K. Vuppala, Vowel-based non-uniform prosody modification for emotion conversion. Circuits Syst. Signal Process. 35(5), 1643–1663,2016.

[54] J. Wang, C. Hu, S. Hung, J. Lee, A hierarchical neural network-based C/V segmentation algorithm for Mandarin speech recognition. IEEE Trans. Signal Process. 39(9), 2141–2146 .1991.

[55] J.H. Wang, S.H. Chen, A C/V segmentation algorithm for Mandarin speech using wavelet transforms, in Proceedings of the International Conference on Acoustics, Speech, Signal Process., vol. 1, pp. 417–420,1999.

[56] P.J. Wolfe, S.J. Godsill, Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement, in Signal Processing Workshop on Statistical Signal Processing, pp. 496–499 (2001) 52. J. Yadav, K.S. Rao, Detection of vowel offset point from speech signal. IEEE Signal Process. Lett. 20(4), 299–302,2013.