IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

A Review on Various Speech Enhancement Techniques

Alugonda Rajani [1], Soundarya .S.V.S[2]

Assistant Professor, Department of ECE, University College of Engineering, JNTUK, Kakinada, India[1]

M.Tech Scholar, Department of ECE, University College of Engineering, JNTUK, Kakinada, India[2]

Abstract: Communications in general and telephonic conversations often take place in noisy environments. The most important method for enhancing speech is the removal of background noise and echo suppression. In this paper review various speech enhancement techniques. Various types in single sensor and multi sensor speech enhancement are reviewed. They mainly focus on noise removal in speech signals. Noise cancellation techniques are also reviewed.

Keywords: Speech enhancement, Communication, Noise removal, Echo suppression.

I. INTRODUCTION

Speech enhancement is a large research area in speech signal processing. The goal of many enhancement algorithms is to suppress the noise in a noisy speech signal. In general, noise can be additive, multiplicative, or convolution, narrowband or broadband, and stationary or non stationary [6]. The majority of research in speech enhancement addresses additive, broadband, stationary noise.

Speech enhancement algorithms [3] have many applications in speech signal processing. Signal enhancement can be invaluable to hearing impaired persons because the ability to generate clean signals is critical to their comprehension of speech. Enhancement algorithms are also used in conjunction with speech recognizers and speech coders as front end processing. It has been shown that enhancing the noisy speech signal before running the signal through a recognizer can increase the recognition rate and thus create a more robust recognizer. Similarly, front end enhancing to speech coding has been shown to decrease the number of bits necessary to code the signal.

The paper is organized as follows. In Section II, the speech enhancement methods are reviewed. The various single sensor methods are reviewed in Section III. The multi sensor methods are given in Section IV and. Conclusion is given in Section V.

II. SPEECH ENHANCEMENT METHODS

Speech enhancement improves the quality of speech signal by using various algorithms. The main objective of enhancement is improvement in intelligibility and the overall perceptual quality of degraded speech signal using audio signal processing techniques. Enhancing of speech signal [16] which is degraded by noise, or noise reduction, is the most important field of speech enhancement. The algorithms of speech enhancement for noise reduction can be categorized into three classes: spectral restoration, filtering techniques and model-based methods.

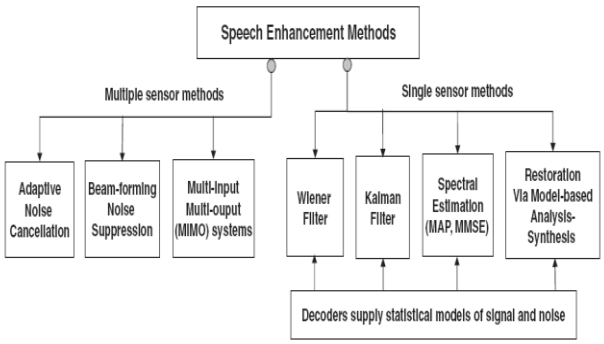

Fig.1. speech enhancement methods

III.SINGLE SENSOR METHODS

Single-input speech enhancement [1] systems where the only available signal is the noise-contaminated speech picked up by a single microphone. Single-input systems do not cancel noise; rather they suppress the noise using estimates of the signal-to-noise ratios [10] of the frequency spectrum of the input signal. Single-input systems rely on the statistical models of speech and noise, which may be estimated on-line from the speech-inactive periods or decoded from a set of pre-trained models of speech and noise. An example of a useful application of a single-microphone enhancement system is a mobile phone system used in noisy environments.

A. Weiner filter

Widely utilized algorithm in speech enhancement research is the Wiener filter. If both the signal and the noise estimates are exactly true, this algorithm will yield the optimal estimate of the clean signal. Through minimizing the mean squared error between the estimated and clean speech signals, the Wiener filter is developed and given by

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

$$H(\omega)=\left[\frac{\left|\hat{S}(\omega)\right|^{2}}{\left|\hat{S}(\omega)\right|^{2}+\left|N(\omega)\right|^{2}}\right]^{2} \quad (1)$$

$$\hat{S}(\omega)=H(\omega)S(\omega) \quad (2)$$

Where **H** is the Wiener filter, and **S** and **N** are the noise corrupted speech and noise spectra, respectively. Because the Wiener filter has a zero phase spectrum [4] the phase from the noisy signal is the output phase for the estimation of the PDS of the clean signal. This was similar to the spectral subtraction algorithms. The Wiener filter assumes that the noise and the signal of interest are ergodic and stationary random processes and thus not correlated to each other. To accommodate the non stationary of speech signals, the signals can be broken into frames to assume stationarity, as is commonly done in speech signal processing research. Another generalization to the Wiener filter is found through incorporating a noise correlation constant k and a power constant a to the filter:

$$H(\omega)=\left[\frac{\hat{S}(\omega)^{2}}{\left|\hat{S}(\omega)^{2}+k\left|N(\omega)\right|^{2}\right|}\right]^{a} \quad (3)$$

Again, similar to spectral subtraction, a prioriknowledge of the noise signal is required, but is often difficult to obtain. Incorporating iterative techniques and methods of estimating the noise are therefore important to the Wiener filter algorithm. The iterative techniques re-estimate the Wiener filter with each iteration.

B. Kalman Filter

The Kalman filter [7] was applied for trajectory estimation in the Apollo space programme; it has many applications, for example it is widely used in process control, in radio and other communication devices particularly as phase-lock-loop systems, in GPS position tracking and guidance systems and in signal denoising and system identification problems. The Kalman filter is a Bayesian filter in that it employs the prior probability distributions of the signal and noise processes, the signal is assumed to be a zero-mean Gaussian-Markov process whereas the noise is assumed to be zero-mean independent identically distributed (IID) Gaussian process. The filter also assumes that the parameters of the models of signal and noise generation and channel distortion are known a priori. Kalman filter [17] formulation is based on a state-space approach in which a state Equation models the dynamics of the signal generation process and an observation Equation models the noisy and distorted observation signal. For a signal vector x(m) and noisy observation vector y(m), Equations describing the state process model and the observation model are defined as

$$x(m)=Ax(m-1)+Bu(m)+e(m) \quad (4)$$

$$y(m)=Hx(m)+n(m) \quad (5)$$

Where **x**(m) is the P-dimensional signal, or the state parameter vector at time m; **A** is a P ×P dimensional state transition matrix that relates the states of the process at times m and m−1,the control matrix **B** and the control vector **u**(m) are used in control applications where often an external input may be applied by a controller process to change, adjust or correct the trajectory of the vector process **x**(m).

In communication signal processing applications, such as channel equalisation or speech enhancement, often there is no external control input vector **u**(m) Kalman Equations reduce to

$$x(m)=Ax(m-1)+e(m) \quad (6)$$

$$y(m)=Hx(m)+n(m) \quad (7)$$

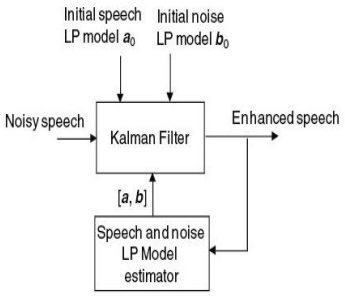

Fig. 2.Kalman filter method

C. MMSE

MMSE estimation of the short time spectral amplitude (STSA); its structure is the same as that of spectral subtraction but in contrast to the Wiener filtering motivation of spectral subtraction, it optimizes the estimate of the real rather than complex spectral amplitudes. Central to their procedures is the estimate of SNR in each frequency bin for which they proposed two algorithms: a maximum likelihood approach and a decision directed approach which they found performed better. The maximum likelihood (ML) [16] approach estimates the SNR by subtracting unity from the low-pass filtered ratio of noisy-signal to noise power (the instantaneous SNR) and half-wave rectifying the result so that it is non-negative. The decision-directed approach forms the SNR estimate by taking a weighted average of this ML estimate and an estimate of the previous frame's SNR determined from the enhanced speech. Both algorithms assume that the mean noise power spectrum is

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

known in advance. Cohen has proposed modifications to the decision-directed approach which are claimed to improve performance further and showed that a delayed response to speech onsets could be avoided by making the estimator non-causal.

D. Restoration

Model-based speech enhancement uses prior knowledge in the form of an explicit stochastic model of speech and, in some cases, of the interfering noise. A number of different speech models are available including some combination of autoregressive (AR) models, coefficient models, hidden Markov models and pitch track models. Enhancement methods based on an AR model of speech generally place no constraint other than stability on the estimated set of AR coefficients. In speech coding applications however, strong constraints are invariably placed on the permitted coefficient values by transforming them into the LSP domain before quantization . Gaussian mixture models in the log spectral domain for both speech and noise using a large number (128) of frequency bins for an 8 kHz sample rate. The speech model comprised either 512 or 1024 mixtures while the noise model had only a single component. They reported that their technique gave enhanced speech of exceptional quality with an improvement in segmental SNR of 7 to 4 dB over the SNR range -5 to +15 dB SNR with no noticeable speech distortion at high SNR values.

E. Single Input Speech Enhancement System

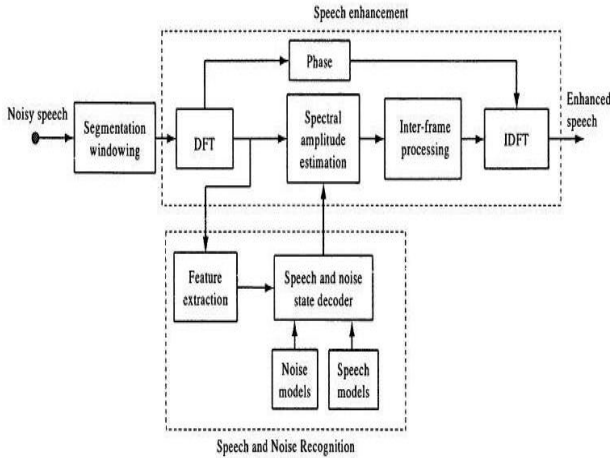

Fig.3. Block diagram of single input speech enhancement system

a. Segmentation and Windowing of Speech Signals

Speech segmentation into a sequence of overlapping frames (20–30 ms) followed by windowing of each segment with a popular window such as the Hamming, Hanning or Hann windows. Speech processing [2] systems divide the sampled speech signal into overlapping frames of about 20–30 ms duration. The N speech samples within each frame are processed and represented by a set of spectral features or by a linear prediction model of speech production. The signal within each frame is assumed to be a stationary process. The choice of the length of speech frames is constrained by the stationarity assumption of linear time-invariant signal processing methods, such as Fourier transform or linear prediction model, and by the maximum allowable delay for real-time communication systems such as voice coders. Note that with a window length of N samples and a sampling rate of Fs Hz the frequency resolution of DFT is Fs/N Hz.

Windowing of a simple waveform, like cos ωt causes its Fourier transform to develop non-zero values known as spectral leakage at frequencies other than ω. The leakage tends to be worst near ω and least at frequencies farthest from ω. The rectangular window has excellent resolution characteristics for signals of comparable strength, but it is a poor choice for signals of disparate amplitudes. This characteristic is sometimes described as low-dynamic-range. At the other extreme of dynamic range are the windows with the poorest resolution. These high-dynamic-range low-resolution windows are also poorest in terms of sensitivity; this is, if the input waveform contains random noise close to the signal frequency, the response to noise, compared to the sinusoid, will be higher than with a higher-resolution window. In other words, the ability to find weak sinusoids amidst the noise is diminished by a high-dynamic-range window. High-dynamic-range windows are probably most often justified in wideband applications, where the spectrum being analyzed is expected to contain many different signals of various amplitudes. In between the extremes are moderate windows, such as Hamming and Hanning which are commonly used in narrowband applications such as the spectrum of a telephone channel.

b. DFT

Speech is segmented into overlapping frames of N samples and transformed to the frequency domain via discrete Fourier transform (DFT). In the frequency domain the noisy speech samples can be represented as

$$Y(k) = X(k) + N(k) \quad k = 0,....N-1 \tag{8}$$

The discrete Fourier transform (DFT) is a specific kind of discrete transform, used in Fourier analysis. It transforms one function into another, which is called the frequency domain representation, or simply the DFT [14], of the original function which is in the time domain. The DFT requires an input function that is discrete. Such inputs are often created by sampling a continuous function, such as a person's voice. The DFT bins can then be processed individually or in groups of frequencies, taking into account the psychoacoustics of hearing in critical bands of the auditory spectral analysis systems.

c. Spectral analysis

When the DFT is used for spectral analysis, the $\{x_n\}$ sequence usually represents a finite set of uniformly-spaced time-samples of some signal $\{x_t\}$,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

where t represents time. The conversion from continuous time to samples discrete-time changes the underlying Fourier transform of x(t) into a discrete-time Fourier transform (DTFT), which generally entails a type of distortion called aliasing. Choice of an appropriate sample-rate is the key to minimize that distortion. Similarly, the conversion from a very long (or infinite) sequence to a manageable size entails a type of distortion called leakage, which is manifested as a loss of detail aka resolution in the DTFT [2]. Choice of an appropriate sub-sequence length in Coherent sampling is the primary key to minimizing that effect. When the available data and time to process it is more than the amount needed to attain the desired frequency resolution, a standard technique is to perform multiple DFTs, for example to create a spectrogram. If the desired result is a power spectrum and noise or randomness is present in the data, averaging the magnitude components of the multiple DFTs is a useful procedure to reduce the variance of the spectrum also called a periodogram .Two examples of such techniques are the Welch method and the Bartlett method; the general subject of estimating the power spectrum of a noisy signal is called spectral estimation.

d. Inter-Frame and Intra-Frame Correlations

Two important issues in modelling noisy speech are Modelling and utilisation of the probability distributions and the intra-frame correlations of speech and noise samples within each noisy speech frame of N samples.

Modelling and utilisation of the probability distributions and the inter-frame correlations of speech and noise features across successive frames of noisy speech.

Most speech enhancement systems are based on estimates of the short-time amplitude spectrum or the linear prediction model of speech. The phase distortion of speech is ignored. In the case of DFT-based features, each spectral sample X(k) at a discrete frequency k is the correlation of the speech samples x(m) with a sinusoidal basis function $e^{-\frac{j2\Pi km}{N}}$. The intra-frame spectral correlation, that is the correlation of spectral samples within a frame of speech, is often ignored, as is the inter-frame temporal correlation of spectral samples across successive speech frames. In the case of speech enhancement methods based on linear prediction models (LP) of speech, the LP model's poles model the spectral correlations within each frame. The de-noising of linear prediction model is achieved through de-noising the discrete samples of the frequency response of noisy speech and that process ignores the correlation of spectral samples.

e. Speech model

Speech enhancement usually the spectral amplitude, or a linear prediction model, of speech is estimated and this estimate is subsequently used to reconstruct speech samples. A variety of methods have been proposed for estimation of clean speech including Wiener filter, spectral subtraction, Kalman filter, the minimum mean squared error (MMSE) and the maximum a posterior (MAP) methods. For the proper functioning of the speech estimation module knowledge of the statistics of speech and noise is required and this can be estimated from the noisy speech or it can be obtained from pre-trained models of speech and noise [4]. The implementation of a noise reduction method, such as the Wiener filter, Kalman filter, spectral subtraction or a Bayesian estimation method, requires estimates of the time-varying statistical parameters and in particular the power spectra or equivalently the correlation matrices of the speech and noise processes. An estimate of the noise statistics can be obtained from the speech-inactive periods, however for the best results the speech and noise statistical parameters are obtained from a network of probability models of speech and noise and this essentially implies that in an optimal speech processing system speech recognition and speech enhancement would need to be integrated. The most commonly used probability models for speech are hidden Markov models (HMMs). Hidden Markov models, or alternatively Gaussian mixture models (GMMs), can also be used for modelling non-stationary noise. To model different types of noise a number of HMMs need to be trained for each type of noise. Alternatively, one can use a GMM of noise with a large number of components, with each component effectively modelling a different type of noise.

IV. MULTI CHANNEL METHODS

Multiple-input speech enhancement systems where a number of signals containing speech and noise are picked up by several microphones. Eg: adaptive noise cancellation, adaptive beam-forming microphone arrays and multiple-input multiple-output (MIMO) acoustic echo cancellation systems. In multiple-input systems the microphones can be spatially configured and adapted for optimum performance. Multiple-input noise reduction systems are useful for teleconference systems and for in-car cabin communication systems. In multiple input noise reduction systems several noisy input signals, picked up by an array of microphones, are filtered, time aligned and combined to reinforce the desired signals and reduce distortions due to noise, interfering speech, echo and room reverberations.

A. multi-input multi-output (MIMO)

Multi-input speech enhancement systems include adaptive beam forming, adaptive noise cancellation, multi-input multi-output (MIMO) teleconferencing systems, stereophonic echo cancellation and in-car MIMO communication systems.

In a typical multi-input speech enhancement system, there are several microphones. The output of each microphone is a mixture of the speech signal, feedback from loudspeakers, speech reflections from walls and noise. Assuming that there are M microphones and N sets of signal and noise sources, there are N×M different acoustic

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

channels between the sources of signals and the microphones. We can write a system of linear equations to describe the relationship between the signals emitted from different sources xi(m), and the signals picked up by the microphones yj(m) as

$$y_j(m) = \sum_{i=1}^{N} \sum_{k=0}^{P} h_{ij}(k) x_i(m-k) \quad \text{j=1,……M} \quad (9)$$

Where $h_{ij}(k)$ denotes the response of the channel from source i to microphone j modelled by a finite impulse response (FIR) linear filter. Note that for simplicity each source of signal, noise or interference is denoted with the same letter x and different index as $x_i(m)$; m is the discrete-time index

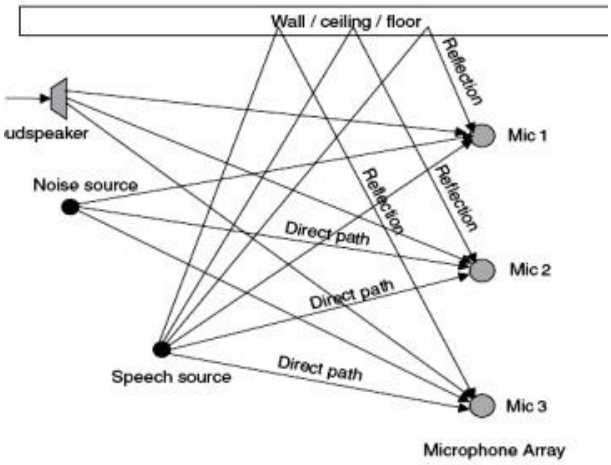

Fig.4 llustration of different sounds and noise arriving at microphones

In the simplest MIMO model, the response of an acoustic channel from sound source i to microphone j via a direct or reflected path can be represented by two parameters, an attenuation factor $\alpha_{ij}(m)$ and a propagation time delay $\tau_{ij}(m)$, as

$$h_{ij}(m) = \alpha_{ij}(m)\delta(m - \tau_{ij}(m)) \quad (10)$$

each source of sound may reach a microphone via a direct path and via a number of indirect paths after reflections in which case the response from source i to microphone j needs to be expressed as

$$h_{ij}(m) = \sum_{k=1}^{L} \alpha_{ijk}(m)\delta(m - \tau_{ijk}(m)) \quad (11)$$

Where $\alpha_{ijk}(m)$ and $\tau_{ijk}(m)$ are the attenuation factor and the propagation time delay along the k th path from source i to microphone j.

B. Beam-forming with Microphone Arrays

Microphone array beam-forming, is a noise reduction method in which an array of microphones and adaptive filters provide steerable directional reception of sound waves. The effect is that sounds arriving at microphones along the main beam of the array are constructively combined and reinforced whereas sounds including disturbances such as noise, reverberation and echo arriving from other directions are relatively attenuated. However, unwanted signals propagating together with the desired signal along the direction of the beam are not suppressed. the microphone is in the far field of the sound source and hence the sound waves reaching the microphone array can be considered as planar (as opposed to spherical) waves. Beam-forming has applications in hands-free communication such as in-car communication, personal computer voice communication, teleconferencing and robust speech recognition. Beam-forming can also be combined with acoustic feedback cancellation.

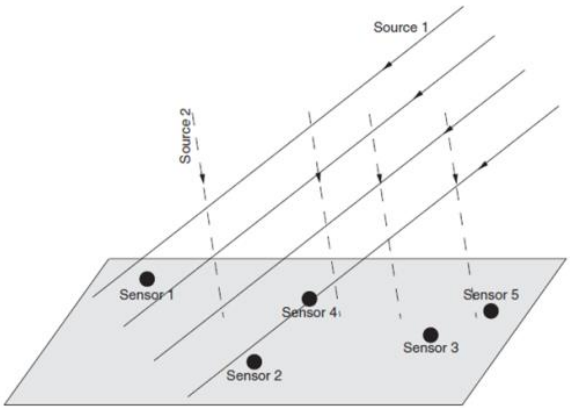

Fig.5. An array of sensors

Beam-forming employs temporal-spatial filtering to steer the microphone array towards a desired direction. In its simplest form a beam-forming algorithm is a delay-sum device; the signal received at the kth microphone is delayed by an amount τk such that all the signals coming from the desired direction are time-aligned and in-phase. The summation stage results in constructive reinforcement of the signals from the desired direction whereas the signals arriving from other directions are not time aligned, are out of phase and relatively attenuated after summation. To adaptively adjust the filters to selectively steer the array and pick up a sound wave from the direction of the source and or where the sound energy is strongest and screen out noise, feedback and reflections of sounds from other directions.

C. Adaptive Noise Cancellation

Adaptive noise Cancellation is an alternative technique of estimating signals corrupted by additive noise or interference. Its advantage lies in that, with no priori estimates of signal or noise, levels of noise rejection are attainable that would be difficult or impossible to achieve by other signal processing methods of removing noise

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

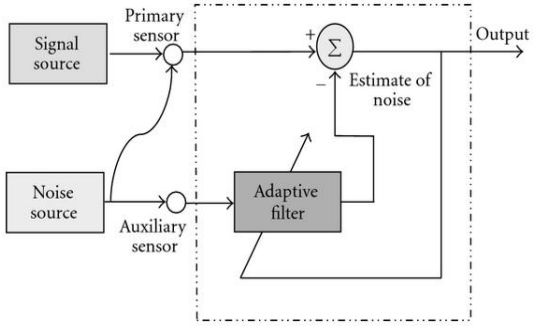

Fig.6. adaptive noise cancellation

V. CONCLUSION

In this paper, we have reviewed various speech enhancement techniques. The speech signal is degraded due to various types of noise. We have presented types of single and multi sensor speech enhancement. Avoiding the noise completely is not possible we hence focus on reduction based on various criterias noise cancellation echo suppression are also important parts of speech enhancement. Kalman filter is one of the efficient forms of enhancement as it is recursive we can improve the quality f the signal,

REFERENCES

[1] Navneet Upadhyay, Abhijit Karmakar, "An Improved Multi-Band Spectral Subtraction Algorithm for Enhancing Speech in Various Noise Environments", International Conference On Design and Manufacturing, IConDM 2013, Elsevier, Vol. 64, Pages 312-321, 13 November 2013.

[2] Yi Zhang, Yunxin Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement", Journal on Speech Communication, Elsevier, Vol. 55, Pages 509-522, 6 November 2012.

[3] Dr. Shaila D. Apte, "Speech Processing Applications", in Speech and Audio Processing, Wiley India Edition.

[4] Sonia Sunny, David Peter S, K Poulose Jacob, "A New Algorithm for Adaptive Smoothing of Signals in Speech Enhancement", International Conference on Electronic Engineering and Computer Science, IERI Procedia, Elsevier, Vol. 4, Pages 337–343, 12 December 2013.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[5] Erik Visser, Manabu Otsuka, Te-Won Lee, "A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments", Journal on Speech Communication, Elsevier, Vol. 41,

[6] Kotta Manohar, Preeti Rao, "Speech enhancement in nonstationary noise environments using noise properties", Journal on Speech Communication, Elsevier, Vol. 48, Pages 96-109, 15 September 2005.

[7] Stephen So, Kuldip K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement", Journal on Speech Communication, Elsevier, Vol. 53, Pages 818-829, 16 February 2011.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Rongshan Yu, "Speech enhancement based on soft audible noise masking and noise power estimation", Journal on Speech Communication, Elsevier, Vol. 55, Pages 964-974, 25 June 2013.

[9] Nima Yousefian, Philipos C. Loizou, John H.L. Hansen, "A coherence-based noise reduction algorithm for binaural hearing aids", Journal on Speech Communication, Elsevier, Vol. 58, Pages 101-110, 23 November, 2013.

[10] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Detroit, MI, USA, May 1995, vol. 1, pp. 153–156.

[11] Yi Zhang, Yunxin Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement", Journal on Speech Communication, Elsevier, Vol. 55, Pages 509-522, 6 November 2012.

[12] Hwai-Tsu Hu, Fang-Jang Kuo, Hsin-Jen Wang, "Supplementary schemes to spectral subtraction for speech enhancement", Journal on Speech Communication, Elsevier, Vol. 36, Pages 205-218, 7 January 2002.

[13]Radu Mihnea Udrea, Nicolae D. Vizireanu, Silviu Ciochina, "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter", Journal on Digital Signal Processing, Elsevier, Vol. 18, Pages 581-587, 15 August 2007.

[14] Kuldip Paliwal, Kamil Wo´jcicki, Belinda Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain", Journal on Speech Communication, Elsevier, Vol. 52, Pages 450-475, 19 February 2010.

[15] Durgesh, Anil Garg , Pankaj Bactor " Speech Enhancement Algorithms: A Brief Review", International Journal for Advance Research In Engineering And Technology, Volume 1, Issue V, June (2013).

[16] D. Pastor and F. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," IEEE Trans. Signal Process., vol. 60, no. 4, pp. 1545–1555,Apr. 2012.

[17] Hiroshi Ijima & Akira Ohsumi, "Detection of Signals in Nonstationary Noise via Kalman Filter-Based Stationarization Approach" in Intechopen, Japan ,2010.

BIOGRAPHIES

Alugonda Rajani is currently working as assistant professor in Jawharlal Nehru Technological University, Kakinada. In the department of Electronics and Communication Engineering. Her areas of interest are Signal Processing, Embedded Systems & Control Systems.

Soundarya S.V.S received a B.Tech degree in the department of Electronics and Communication from Kakinada institute of engineering and technology for women. Currently pursuing M.Tech in Jawharlal Nehru Technological University, Kakinada. In the department of electronics and communication engineering.