Quality Estimation of Alaryngeal Speech

R.Dhivya [#1], Judith Justin [*2], M.Arnika [#3]

#PG Scholars, Department of Biomedical Instrumentation Engineering, Avinashilingam University
Coimbatore, India
[1] dhivyaramasamy21@gmail.com
[3] arni.biomed@gmail.com
[*] Faculty, Department of Biomedical Instrumentation Engineering, Avinashilingam University
Coimbatore, India
[2] judithvjn@yahoo.co.in

***Abstract*— Quality assessment can be done using subjective listening tests or using objective quality measures. Objective measures quantify quality. The sentence material is chosen from IEEE corpus. Real world noise data was taken from the noisy speech corpus NOIZEUS. Alaryngeal speaker‘s voice (alaryngeal speech) is recorded. To enhance the quality of speech produced from the prosthetic device, four classes of enhancement methods encompassing four algorithms mband spectral subtraction algorithm, Karhunen–Loéve transform (KLT) subspace algorithm, MASK statistical-model based algorithm and Wavelet Threshold-Wiener algorithm are used. The enhanced speech signals obtained from the four classes of algorithms are evaluated using Perceptual Evaluation of Speech Quality (PESQ). Spectrograms of these enhanced signals are also plotted.**

Keyword- Tracheoesophageal prosthesis (TEP), alaryngeal speech, Speech Enhancement, Perceptual Evaluation of Speech Quality (PESQ)

I. INTRODUCTION

The larynx is an enlargement in the airway and superior to the trachea. It is a passageway for air moving in and out of the trachea which prevents foreign objects from entering the trachea. The larynx also houses the vocal cords. Cancer can develop in any part of the larynx, but the cure rate is affected by the location of the tumor. Advanced stage of laryngeal carcinoma leads to total laryngectomy. It is the surgical process of removal of larynx. The speaker who has undergone laryngectomy cannot produce voice in a conventional manner because the vocal folds have been removed in the process. Therefore, they require an alternative speaking method to produce voice, using sound sources which generate voice without vibrating the vocal folds. The voice thus produced is called alaryngeal speech.

There are various techniques of producing alaryngeal speech [5]. In this paper, we focus on the voice produced by a laryngectomee speaker implanted with voice prosthesis. The device implanted is Blom-singer duckbill voice prosthesis. The alaryngeal voice of a speaker implanted with prosthetic device is recorded through unidirectional microphone and is stored on a computer. The quality of the voice of the laryngectomee is assessed through objective measure - PESQ which is recommended by International Telecommunication Union.

Speech enhancement algorithms are commonly used to improve the performance of modern communication devices in noisy environments [1]. We evaluate the enhanced speech signals obtained from existing speech enhancement algorithms [7], [10] to assess the performance as per the standardized methodology based on ITU-T P.835 [12]. PESQ is used to evaluate the quality of the alaryngeal speech sentence. Short time Fourier Transform (STFT) analysis is also performed.

This paper is organized as follows: Section II describes the nature of the signal and the method selected for our study, in section III we present the algorithms taken for the enhancement, section IV gives an insight into the evaluation protocol. In section V we present the results of the objective measure PESQ and the spectrograms. The conclusions are given in section VI.

II. MATERIALS AND METHOD

Speech sounds are sensations of air pressure vibrations produced by air exhaled from the lungs and modulated and shaped by the vibrations of the glottal cords and the resonance of the vocal tract as the air is pushed out through the lips and nose as shown in Fig. 1. If there is an originate of larynx cancer (shown in Fig. 2.) the primary treatment option leads to laryngectomy, then speech sounds are produced with the help of an implanted prosthetic device, placed using an tracheoesophageal (TEP) puncture (shown in Fig. 3.), [6] created by the surgeon. The speaker pushes air into the esophagus and then pushes it back up, articulating sounds. A male speaker implanted with the Blom-singer voice prosthesis was considered for the study. The valve used is 18mm long with a 16Fr. puncture and voice restoration practice helped him produce a pseudo voice, which was clear and produced almost perfect pronunciations. The speaker is given a standard text (a sentence to read from the IEEE sentence database). “Sentence: Kick the ball straight and follow through”. Four real world noises at 0, 5, 10 and 15dB levels are added to the data [11]. The corrupted sentence is enhanced by existing enhancement algorithms and PESQ scores are obtained which helps to compute the performance of the algorithm. The method followed is best explained with a flow chart in Fig. 4.

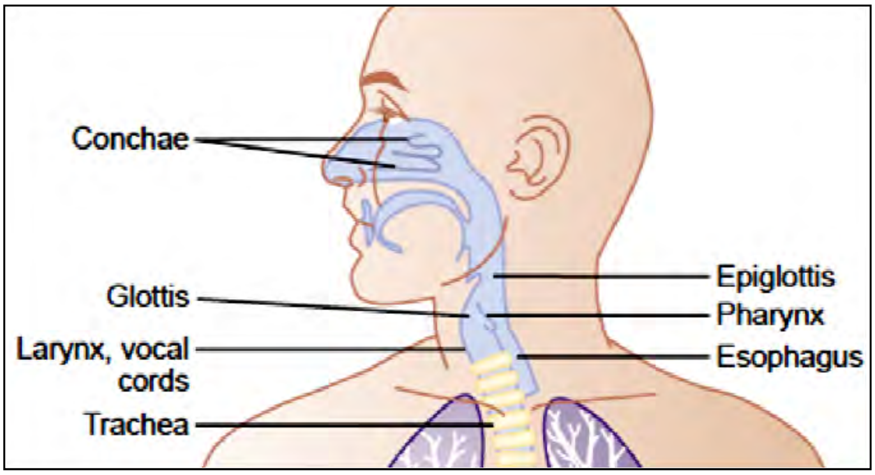
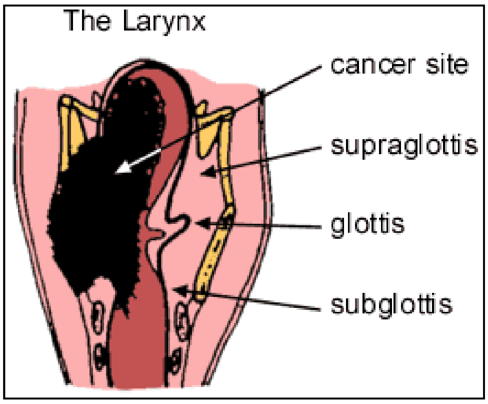


Fig.1. Anatomy of Larynx



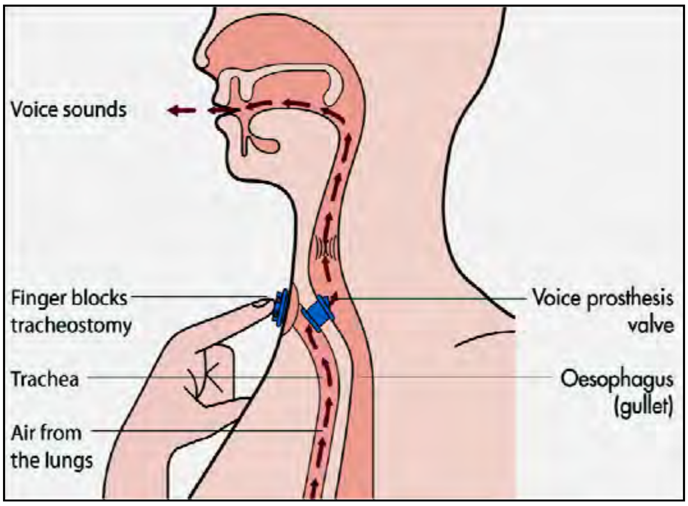


Fig.2. Larynx Cancer

Fig.3. Voice Prosthesis

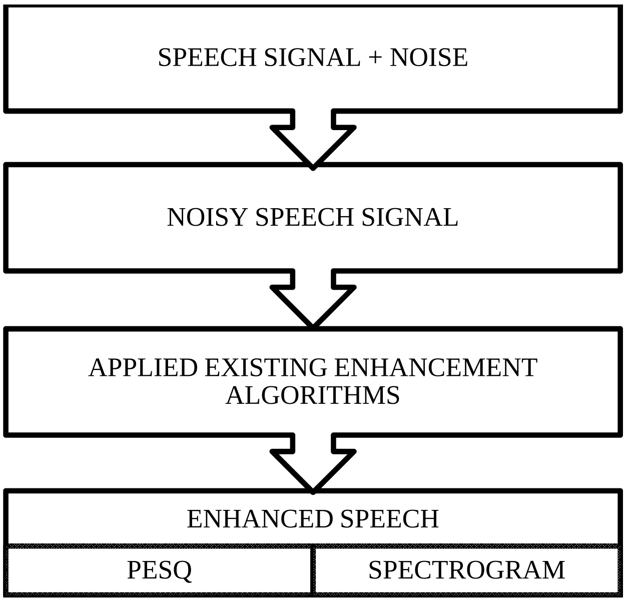


Fig.4. Steps for analysing speech signal

III.ALGORITHMS EVALUATED

In this study four algorithms are considered. One representative class from each of the four algorithms is chosen for our study; mband spectral subtraction algorithm, KLT subspace algorithm, Wavelet thresholding wiener algorithm and MASK statistical-model based algorithm.

A. Spectral Subtraction

The basic principle of the spectral subtraction method is to subtract the magnitude spectrum of noise from that of the noisy speech [9]. The noise is assumed to be uncorrelated and additive to the speech signal.

Let y (n), the noise corrupted input signal, be composed of the clean speech signal x (n) and the additive noise signal d (n). i.e.

$$y(n) = x(n) + d(n) \qquad (1)$$

Taking discrete-time Fourier transform on both sides gives

$$Y(\omega) = X(\omega) + D(\omega) \qquad (2)$$

Y (ω) can be expressed in polar form as

$$\mathrm{Y}(\omega) = \left|Y(\omega)\right| e^{j\phi_y(\omega)} \qquad (3)$$

Where |Y(ω)| is the magnitude spectrum and $\phi_y(\omega)$ is the phase (spectrum) of the corrupted noisy signal. The noise spectrum D(ω) is expressed in terms of magnitude and phase spectra

$$\mathrm{D}(\omega) = \left|D(\omega)\right| e^{j\phi_d(\omega)} \qquad (4)$$

The magnitude noise | D (ω)| is unknown, but can be replaced by its average value computed during non-speech activity (during speech pauses). Similarly, the noise phase $\phi_d(\omega)$ can be replaced by the noisy speech phase $\phi_y(\omega)$. The phase spectra do not affect speech intelligibility, but may affect speech quality to some degree. We can estimate the clean signal spectrum simply by subtracting the noise spectrum from the noisy speech spectrum given as:

$$\widehat{\mathrm{X}}(\omega) = [\left|Y(\omega)\right| - \left|\widehat{\mathrm{D}}(\omega)\right|] \cdot e^{j\phi_y(\omega)} \qquad (5)$$

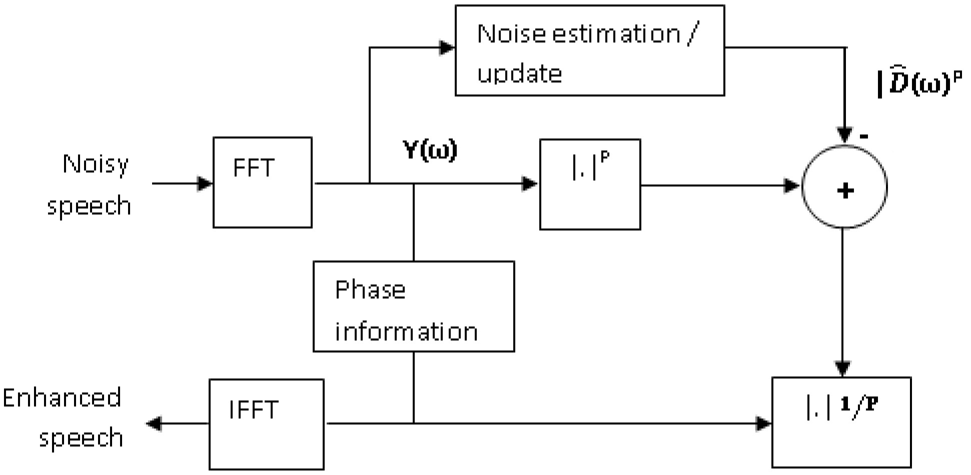


Fig.5. General Form of Spectral Subtraction Algorithm

$\left|\widehat{\mathrm{D}}(\omega)\right|$ is the estimate of the magnitude noise spectrum made during non speech activity. The enhanced speech signal can be obtained by simply taking the inverse Fourier transform of $\widehat{\mathrm{X}}(\omega)$. Equation (5) summarizes the principle of spectral subtraction.

A general expression for the spectral subtraction algorithms is given as:

$$\left|\widehat{\mathrm{X}}(\omega)\right|^p = \left|Y(\omega)\right|^p - \left|\widehat{\mathrm{D}}(\omega)^p\right| \qquad (6)$$

Where p is the power exponent, when p=1 we get the original magnitude spectral subtraction and p=2 yields the power spectral subtraction algorithm. The estimate of the clean speech signal is obtained in the spectral domain with equation (5); the enhanced speech signal is obtained by inverse DFT transformation of $\widehat{X}(\omega)$ .Thus a general form of estimated speech in frequency domain can be expressed as:

$$s(n) = IDFT\left[\left(\left|Y(\omega)\right|^{a} - \widehat{D}(\omega)^{a}\right)^{1/a}\right] \qquad (7)$$

It is reported that a multi-band approach to spectral subtraction removes the colored noise and is suitable for enhancing uncorrelated additive noise. Here the speech spectrum is divided into N non-overlapping bands, and spectral subtraction is performed independently in each band. So, the estimate of the clean speech spectrum in the ith band is obtained by:

$$\left|\widehat{X}_{i}(k)\right|^{2} = \left|\overline{Y}_{i}(k)\right|^{2} - \alpha_{i} \cdot \delta_{i} \cdot \left|\widehat{D}_{1}(k)\right|^{2}, bi \le k \le ei \qquad (8)$$

Where bi and ei are the beginning and ending frequency bins of the ith frequency band, αi is the over-subtraction factor of the ith band and δi is an additional band-subtraction factor that can computed for each frequency band which helps remove the noise. Mband algorithm is proved to be better among 3 classes of spectral subtraction algorithms. Hence, we have chosen it for our study.

B. Statistical Model Based Algorithm

The spectral estimate of speech plays a major role in computing the noise masking threshold when simultaneous masking properties are considered. Spectral-subtraction method roughly estimates the speech spectra and can provide acceptable performance; however, the estimation can be further improved by two-step-decision-directed method [4]. Experimental results show that the background noise can be efficiently suppressed by embedding the two-step-decision-directed algorithm in the perceptual gain factor. A frequency domain optimal linear estimator is proposed which incorporates the masking properties of the human auditory system to make the residual noise distortion inaudible. The human listener will not perceive any noise distortion as long as the power spectrum density of the distortion lies below the masking threshold. If we constrain the k th spectral component of the residual noise to be lower than the masking threshold, denoted asT_k, in frequency bin k we can compute μ_k the value that meet this constraint. Assuming that the constraints α_k are set equal to the masking thresholds T_k and the equality is satisfied, then $\varepsilon^2_{n,k} = \alpha_k$ implies that

$$g^{2}(k) \cdot s_{n}(k) = T_{k}, \qquad \text{For k=1, 2 ...N} \qquad (9)$$

C. Subspace Algorithm

Signal subspace based speech enhancement algorithms are based on the principle that the clean signal might be confined to a subspace of the noisy Euclidean space. Decomposition of the noisy signal into signal and noise subspaces can estimate the clean signal by nulling the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace. The subspace decomposition can be done using the eigen value decomposition (EVD). The idea behind the frequency domain constrained linear optimal estimator is that the signal distortion can be minimized subject to constraints on the spectrum of the residual noise. The decomposition of the vector space of the noisy signal into a signal and noise subspace can be obtained by applying the Karhunen–Loéve transform (KLT) to the noisy signal [2]. The KLT components representing the signal subspace were modified by a gain function determined by the estimator, while the remaining KLT components representing the noise subspace were nulled. The enhanced signal was obtained from the inverse KLT of the modified components.

Speech enhancement problem will be described as a speech signal x being transmitted through a distortion less channel that is corrupted by additive noise w. The resulting noisy speech signal y can be expressed as

$$y = x + w \qquad (10)$$

where, x=[x1,x2,….,xM]T, w=[w1,w2,…..,wM]T and y=[y1,y2,….,yM]T . The observation period has been denoted as M. Henceforth, the vectors w, x, y will be considered as part of CM. The speech enhancement system will attempt to estimate the original signal using a single channel of received speech.

D. Wiener Algorithm

Signal processing for non stationary by Fourier method is not well suited for detection and classification, so wavelets can able to approximate time varying non stationary signals in a better way. Wiener type algorithm estimates the complex spectrum, to obtain a clean signal from the noisy signal corrupted by additive noise.

Wavelet are local in both frequency and time scale, hence localization is advantage for removing noise. Wavelet denoising works only when the signal characteristics are known in advance but will distort some of the desired signal when thresholding is applied. The drawback of the Wiener filter is the fixed frequency response at all frequencies and the requirement to estimate the power spectral density of the clean signal and noise prior to filtering. The noise term η(ω) is stationary and colored ,the variance of the noise wavelet coefficients, $n_{j,k}$, j-denotes the scale or level of wavelet decomposition and k denotes the kth coefficient, will be different for each scale in the wavelet decomposition [3]. Scale-dependent thresholding can be used to account for the different variances of the noise wavelet coefficients in each scale. The noise wavelet coefficients, $n_{j,k}$, where

$$Var(n_{j,k}) = \sigma_j{}^2 = \frac{1}{N}\sum_{k=0}^{N-1} S(k)\left|H_j(k)\right|^2 \qquad (11)$$

$H_j(k)$ is the frequency response of the length N period wavelet filter of level j, and S(k) is the Fourier transform of the noise. The low variance spectral estimators based on wavelet thresholding the multitaper spectra does not encounter the musical noise.

IV. Perceptual Evaluation of Speech quality

Perceptual domain measures are based on models of the human auditory system, compared to time and spectral domain measures and they have a higher chance of predicting the subjective quality of speech. Perceptually relevant information is both sufficient and necessary for precise assessment of perceived speech quality. One of the commonly used perceptual quality measure is Perceptual Evaluation of Speech Quality (PESQ). PESQ is a validated metric recommended by International Telecommunication Union (ITU) [8], [12] for assessing speech quality. PESQ predicts the subjective opinion score of an enhanced speech. In PESQ algorithm, a reference signal and enhanced signal are first aligned in both time and level. The final PESQ score is computed as a linear combination of the average disturbance value dsym and the average asymmetrical disturbance value dasym as follows:

$$PESQ = a0 + a1\cdot dsym + a2\cdot dasym \qquad (12)$$

Where a0=4.5, a1= - 0.1 and a2= - 0.0309

The range of the PESQ score will be a MOS-like score, i.e., a score of rating of 1-2-3-4-5 is given for unsatisfactory-poor-fair-good-excellent on a listening quality scale.

V. Results and Discussion

PESQ values were calculated for the alaryngeal voice enhanced with the four different classes of speech enhancement algorithms - mband Spectral subtraction algorithm, KLT subspace algorithm, Wavelet thresholding wiener algorithm and MASK statistical-model based algorithm for the cafeteria babble noise (shown in Fig.6.), Car noise (shown in Fig.7.), Street noise (shown in Fig. 8.), Train noise (shown in Fig. 9.) at 0, 5, 10, 15 dB levels. For mband spectral subtraction algorithm, the mean PESQ value for 0 dB lies in the range of 2.4291 which is between poor and fair; 5 dB lies in the range of 2.8260, 10 dB falls in the range of 3.0611 lies in fair quality scale; 15 dB is 3.2903 which is in the range between fair and good. For wt-weiner algorithm, the mean PESQ value for 0 dB falls in the range of 1.6651 lies between unsatisfactory and poor; 5 dB falls in the range of 2.2229, which is closer to poor quality scale; 10 dB falls in the range of 2.4812 lies between poor and fair and 15 dB falls in the range of 2.8107 which is closer to fair quality scale. For mask algorithm, the mean PESQ value for 0 dB lies in the range of 1.6419 lies close to poor quality; 5 dB lies in the range of 2.0812, which is closer to poor quality scale; 10 dB falls in the range of 2.4124 between poor and fair and 15 dB falls in the range of 2.7768 which is closer to fair quality. For KLT algorithm, the mean PESQ value for 0 dB falls in the range of 2.0402 lies closer to poor quality scale; 5 dB falls in the range of 2.5293 between poor and fair; 10 dB falls in the range of 2.8013- closer to fair quality scale;15 dB falls in the range of 2.9698 closer to fair quality scale. The result of the mband spectral subtraction algorithm seems to be better than all the other algorithms since it has given good results for the higher noise level (15dB).

The spectrogram is a visual representation of the spectrum of frequencies in a sound. It is graph of the energy content of a signal expressed as function of frequency and time. A spectrogram represents the frequency change in a signal with time During regions of silence, and at frequency regions where there is little energy, the spectrogram appears blue; red regions indicate areas of energy - caused for example by vocal fold closures, harmonics or formant vibration in a speech signal. The spectrogram of a signal s (t) can be estimated by computing the squared magnitude of the STFT of the signal S (t), as follows:

$$Spectrogram(t,\omega) = \left|STFT(t,\omega)\right|^2 \qquad (13)$$

From the spectrogram results, we can find out the difference between the frequency information of before (shown in Fig.10.) and after (shown in Fig. 11.) the speech enhancement. It also deduces the energy patterns, formant transition and frequency values of the formants.

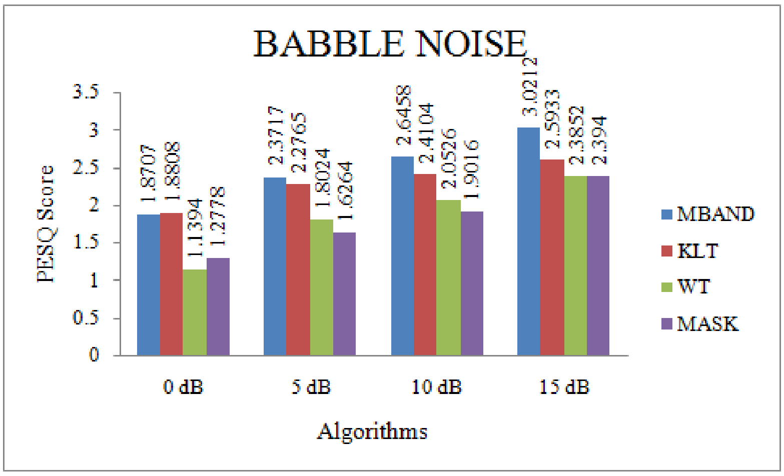


Fig.6. PESQ scores for babble noise

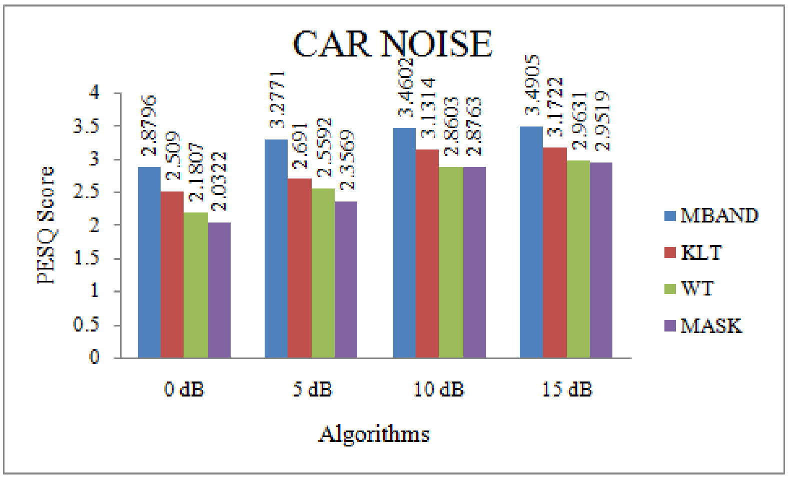


Fig.7. PESQ scores for car noise

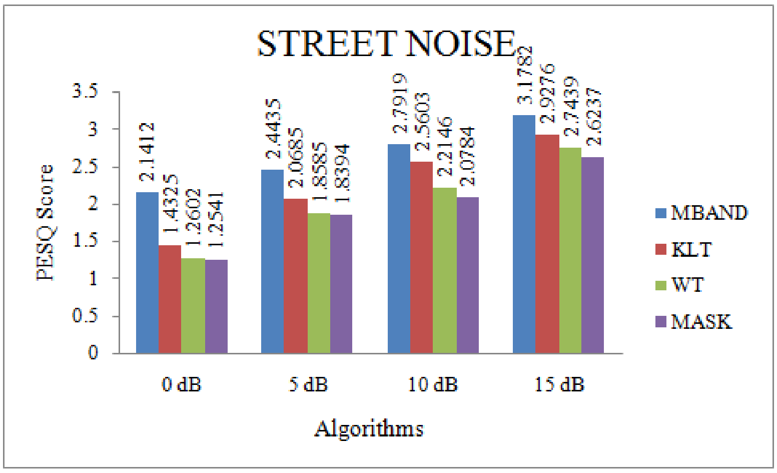


Fig.8. PESQ scores for street noise

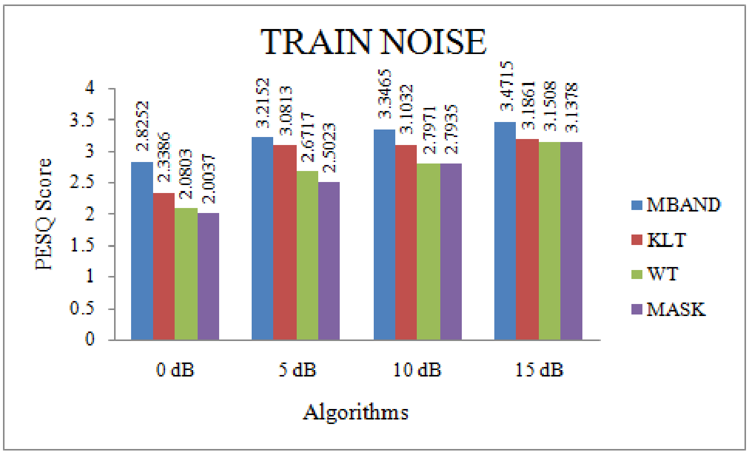


Fig.9. PESQ scores for train noise

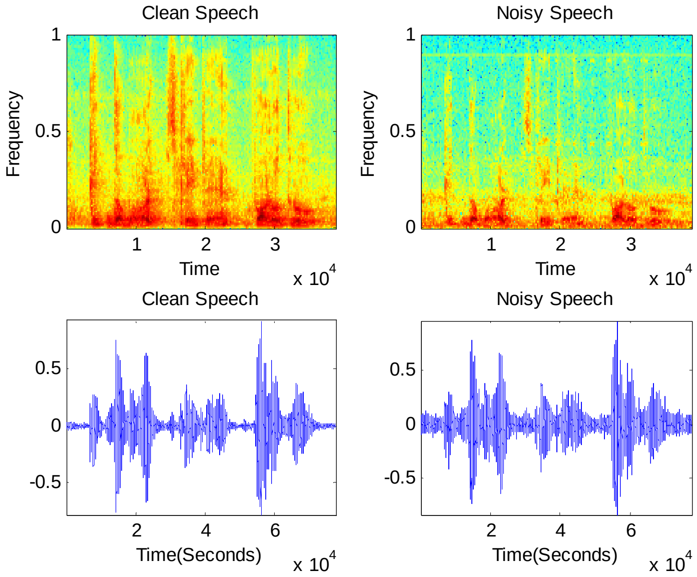


Fig.10. Spectrogram and time plot of Noisy signal with 10 dB

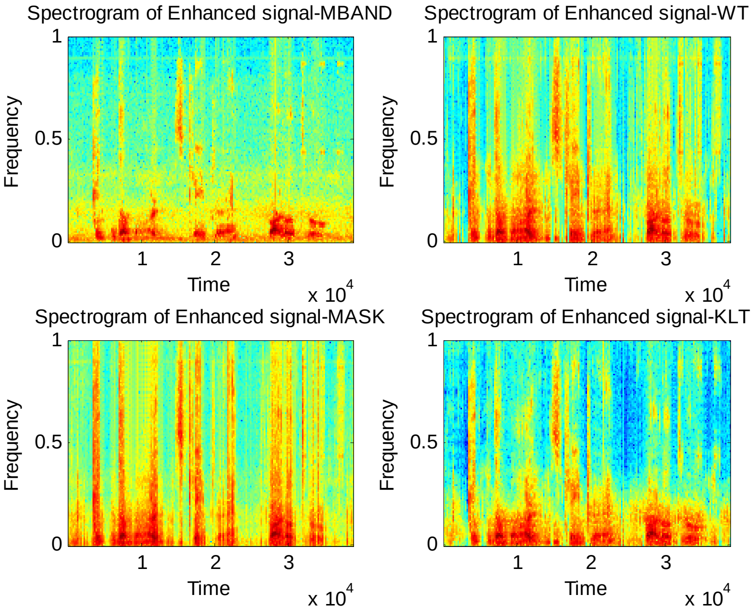


Fig.11. Spectrogram of enhanced signal with 5 dB

VI. CONCLUSION

The PESQ value is closer to fair quality scale as observed from the results of all the four algorithms. Thus we conclude that of the four algorithms chosen, the mband spectral subtraction algorithm performed better enhancement compared to all the other algorithms. The alaryngeal speech produced by voice prosthesis tracheoesophageal puncture (TEP) produces a pseudo speech which is of good quality in the presence of real world noises. Acoustic analysis can be done further to estimate how close it is to the natural voice produced by normal persons.

REFERENCES

[1] P.Loizou, *Speech enhancement: Theory and Practice,* CRC Press, LLC, Boca Raton, 2007.

[2] Yi Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp.334-341, 2003.

[3] Y.Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. on Speech and Audio Processing,* vol.12(1), pp.59-67, 2004.

[4] Yi Hu and P. Loizou, "Incorporating a psycho-acoustical model in frequency domain speech enhancement," *IEEE Signal Processing Letters*, vol.11(2), pp.270-273, 2004.

[5] D.Globlek, S. Stajner- Katusic, M. Musura, D. Horga and M. Liker, "Comparison of Alaryngeal voice and speech," *Logoped Phoniatr Vocol,* vol.29, pp.87-91, 2004.

[6] D.Globek, Boris Simunjak, Mirko Ivkic, Mladen Hedjever, "Speech and voice analysis after near-total laryngectomy and tracheoesophageal puncture with implantation of Provox 2 prosthesis," *Logoped Phoniatr Vocol*, vol 29,pp.84-86, 2004.

[7] Yi Hu and P. C. Loizou, “Subjective evaluations and comparisons of speech enhancement methods,” *Speech Communication* , vol.49, pp. 588-601, 2007.

[8] Yi Hu, Philips C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech & Language Processing, vol.*16(1), pp. 229-238, 2008.

[9] S. Kamath, and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* 2002

[10] Yi Hu and P. Loizou, “Subjective comparison of speech enhancement algorithms,” *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006, vol. I, pp. 153-156.

[11] Loizou. (2007) NOIZEUS website.[Online].Available: http://ecs.utdallas.edu/loizou/speech/noizeus/

[12] *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation P.862, 2000.