A NOVEL SPEECH ENHANCEMENT TECHNIQUE

R.Dhivya[1], Judith Justin[2]

[1]PG Scholar, Department of Biomedical Instrumentation Engineering, Avinashilingam University, TamilNadu, India
[2]Faculty, Department of Biomedical Instrumentation Engineering, Avinashilingam University, TamilNadu, India

Abstract

This enhancement technique is a novel one and is based on the combination of Wavelet thresholding and Spectral Subtraction. Five wavelet filters are compared and the best filter is selected based on their performance of Signal to Noise Ratio. The selected filter is applied to the detail coefficients for thresholding. Approximation coefficient is applied to spectral subtraction filter. The reconstructed signal is evaluated using the metrics such as SNR (Signal to Noise Ratio), Correlation coefficient and PESQ (Perceptual Evaluation of Speech Quality). Real time data is recorded from Alaryngeal speakers and real world noise from Noizeus corpus is used for the study.

Keywords: *Wavelet Thresholding, Signal to Noise Ratio, Correlation coefficient, Perceptual Evaluation of Speech Quality*.

----------***----------

1. INTRODUCTION FOR SPEECH ENHANCEMENT

Numerous speech enhancement algorithms have been proposed to improve the performance of communication devices working in normal environments surrounded in noise. The rapid increase in usage of speech processing algorithms in multi-media and telecommunication applications raises the need for speech quality evaluation [1]. Accurate and reliable assessment of speech quality is thus becoming vital for the satisfaction of the end-user or customer of the deployed speech processing systems (e.g., mobile phone, video conferencing equipment, speech synthesis system, etc.). It is possible for speech to be highly intelligible and still be of poor quality. Also, although two different algorithms may produce equal word intelligibility scores, listeners may perceive the speech of one of the algorithms as being more natural, pleasant and acceptable. Therefore there is a need to measure the attributes of a speech signal. Reliable rating of speech quality is a challenging task because quality assessment is highly subjective and reliability of subjective measurements becomes an issue.

In this research a novel method of speech enhancement technique is proposed and its performance is evaluated using some validated measures.

The data used for the study is unique. A speaker who suffered from laryngeal carcinoma is surgically treated with Total Laryngectomy, which is the removal f the larynx. This is followed by a loss of voice after which voice prosthesis is implanted. The speaker gets trained to speak again with the help of a speech therapist. The voice thus produced is called alaryngeal voice. The quality of the voice of the laryngectomee is assessed through Signal to Noise Ratio (SNR), Correlation coefficient and Perceptual Evaluation of Speech Quality (PESQ) which is recommended by International Telecommunication Union.

This paper is organized as follows: Section II describes the materials and the method selected for the study, Section III describes the proposed technique adopted for the enhancement, and a study of an existing speech enhancement algorithm and Section IV presents the quality metrics adopted for the evaluation. Section V elaborates on the results obtained and a comparison between the two algorithms. In section VI Conclusions are drawn about the novel technique of speech enhancement.

2. MATERIALS AND METHODS

A male speaker implanted with the Blom-singer voice prosthesis was considered for the study. A sentence is presented to the alaryngeal speaker from the IEEE sentence database. The sentences in this database have the features of being phonetically balanced. The sentence, "Kick the ball straight and follow through" is used for the study. The voice generated by the speaker after implantation with Blom-Singer Duckbill Voice Prosthesis is recorded using a unidirectional microphone in an anechoic room and is stored on a computer. In order to study the performance of the algorithm in a real world situations, real world noises like babble, car, street and train at different levels (0 dB, 5 dB,10 dB and 15 dB) are added and the performance of the algorithm are evaluated using validated metrics. The quality metrics used for the evaluation are SNR, Correlation coefficient and PESQ scores. The proposed method is explained with a block diagram shown in Figure 1.

3. PROPOSED TECHNIQUE FOR SPEECH ENHANCEMENT

The recorded speech signal is added to different real world noises - Babble, Car, Street and Train at four different levels 0, 5, 10 and 15 dB respectively. The noisy speech signal generated is decomposed using wavelet filters [2], [3]. Mother wavelet is carefully selected to better approximate and capture the transient spikes of the original signal. "Mother wavelet" is determined how well we estimate the original signal in terms of the shape and at the same time it will affect the frequency spectrum of the denoised signal. The choice of mother wavelet can be based on modifying the detail coefficients using wavelet thresholding techniques and keeps the approximation Coefficients unaltered. The selection of wavelet [4], [5] is based on Signal to Noise Ratio (SNR) among the signal of interest and the wavelet-denoised signal shown in table1.

The signal is decomposed at one level, since the signal is down-sampled by 2, approximate and detail coefficients. To threshold the detail coefficients the threshold value is selected based on soft threshold evaluator of unbiased risk, 'Rigrsure'. Soft thresholding shrinks the coefficients that are larger than the threshold. The selected threshold value is shown in equation 1.

$$l = s\sqrt{w_b} \qquad (1)$$

σis the standard deviation of the noisy signal, ω is the squared wavelet coefficient.

The detail coefficient is denoised and the approximation coefficients are enhanced using the multiband spectral subtraction method [6].

3.1 Wavelet Thresholding

Wavelet Transform is applied to the noisy signal to decompose into approximation and Detail coefficients. Soft thresholding is applied to the detail coefficients to denoise the noisy signal. Assume that y(n) is a noisy signal and is given as in equation 2.

$$y(n)=x(n)+d(n) \qquad (2)$$

Where x(n) is original signal and d(n) is noise signal. High frequency features may present in original signal, which is well-preserved by wavelet transform. The detail coefficient of the noisy signal is shrinked by soft thresholding given in equation (3),

$$\widehat{X} = \begin{cases} y - \mathrm{sgn}(y)T & if\ |y| \geq T \\ 0 & if\ |y| < T \end{cases} \qquad (3)$$

There are different types of wavelet filters such as [6]:

- Daubechies filter is the extension of haar wavelets, that produce smoother scaling and wavelet functions.
- Coiflets filter allow approximately equal number of zero scaling function and wavelet moments.
- Symlet filter is a modified version of Daubechies wavelets with increased symmetry.
- Biorthogonal filter describes a pair of topological vector spaces that are in duality with a pair of indexed subsets in a specific way.
- Reverse biorthogonal

These wavelets are assessed through Signal to Noise Ratio to assess the best performing denoising wavelet as shown in Table 1.

Wavelet Reconstruction: After the detail coefficients are shrinked and approximation coefficients enhanced, these two coefficients are reconstructed using inverse discrete wavelet transform with the help of the mother wavelet. The reconstructed signal is evaluated using the quality metrics such as SNR, Correlation Coefficient and PESQ.

3.2 Spectral Subtraction

Generally noise will not affect the speech signal uniformly over the spectrum; some frequencies will be exaggerated based on the spectral characteristics of the noise. In mband approach, the spectrum is allocated into N overlapping bands, and spectral subtraction is achieved independently in each band [7].

The estimate of clean speech spectrum in ith band is obtained by the equation 4,

$$\left|\widehat{X}_i(w_k)\right|^2 = \left|\overline{Y_i}(w_k)\right|^2 - a_i.b_i.\left|\widehat{D}_i(w_k)\right|^2 \quad b_i \leq w_k \leq e_i \qquad (4)$$

Where $w_k = 2\pi k / N$, are discrete frequencies, $\left|\widehat{D}_i(w_k)\right|^2$ is the estimated noise power spectrum, b_i and e_i are the commencement and termination frequency bins of the ith frequency band, a_i is the over-subtraction factor, and b_i is the additional band-subtraction factor that can be set individually for each frequency band to adapt the noise removal process.

3.3 Quality Metrics

Speech quality measure has three types of objectives [9] such as perceptual domain, spectral domain, and time domain measures. In analog or waveform coding systems, time domain measures are used, which has the goal to reproduce the waveform itself. Among them the most known methods are SNR and Segmental SNR (SNRseg). Spectral domain

measures are the second type that ranges between 15 and 30 ms long, calculated using speech segments. Spectral domain measures are less sensitive to misalignments between the original and distorted signal, so they are much more reliable than time domain measures. Perceptual domain measures are based on models of human auditory acuity that is in contrast to spectral domain measures. It incorporates human auditory models and transform speech signal into a perceptually pertinent domain e.g., Bark spectrum or loudness domain.

3.3.1 SNR

Synchronization of original and distorted signals is necessary, because the waveform is directly compared in time domain or else the performance will be poor. Synchronization is difficult; the simplest possible method is to calculate Signal to Noise Ratio (SNR). It measures the distortion of waveform coders that reproduce the input waveform, calculated as based on equation 5:

$$SNR = 10\log_{10}\frac{\sum_{i=1}^{N} x^2(\mathrm{i})}{\sum_{i=1}^{N}(x(i) - \hat{x}(i))} \quad (5)$$

Where x(i) and $\hat{x}$ (i) are the original and processed speech samples indexed by i and N is the total number of samples.

3.3.2 Correlation Coefficient

Correlation expresses the relationship between two signals numerically. Correlation is expressed in terms of scale from 0 to1. If the correlation value is closer to 0 indicates weaker, and if it is closer to 1 indicates stronger correlation.

3.3.3 PESQ

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 [8]. The commonly used perceptual quality measure is Perceptual Evaluation of Speech Quality (PESQ). PESQ is an authenticated metric suggested by International Telecommunication Union (ITU) 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 allied 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 given in equation 6:

$$PESQ = a0 + a1 \times dsym + a2 \times dasym \quad (6)$$

The assortment 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 inadequate-deprived-reasonable-worthy-outstanding on a listening quality scale.

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

4. RESULTS AND DISCUSSION

Different wavelet filters considered include Daubechies, coif let, Symlet, demy, biorthogonal and reverse biorthogonal. These wavelet filters are compared based on the Signal to Noise Ratio and the results are shown in Table 1. From the comparison, we find Symlet 7 wavelet filter to perform well in denoising the signal.

Table -1: Selection of Wavelet Filter

Wavelet Filter	SNR
db9	9.9256
coif4	9.92
sym7	9.9258
Demy	9.9183
bior3.7	9.922

4.1 SNR

The Signal to Noise Ratio is one among the metric considered for evaluation of the proposed algorithm. The values are increasing for higher noise levels. Maximum value for babble noise at 15dB is 9.1733, for car noise at 15dB is 9.5701, for street noise at 15dB is 9.3026 and for train noise at 15dB is 9.1018, shown in Figure 2. Proposed algorithm give best under high noise levels, which makes it suitable for real life situations.

4.2 Correlation Coefficient

Correlation values are increasing for higher dB noise level. Maximum value for babble noise at 15dB is 0.9763, for car noise at 15dB is 0.9813, for street noise at 15dB is 0.9768 and for train noise at 15dB is 0.9763, shown in Figure 2. Proposed algorithm is found to yield good correlation for higher noise levels.

4.3 PESQ

PESQ values are increasing for higher dB noise level. Maximum value for babble noise at 15dB is 3.1705, for car noise at 15dB is 3.5105, for street noise at 15dB is 3.3086 and for train noise at 15dB is 3.3086, shown in Figure 2. The PESQ values indicate reasonable quality in the listening quality scale.

5. CONCLUSIONS

From assessment of quality metrics, results show that speech enhancement using Wavelet Thresholding integrated with Spectral Subtraction yields better performance. Comparing with all noise levels, proposed algorithm works better in higher noise dB level and also by comparing enhancement of different noises, denoising of Car noise works better. The alaryngeal speech produced by voice prosthesis tracheoesophageal puncture (TEP) produces a pseudo speech which is of reasonable quality in the presence of real world noises. Further the proposed algorithm can be compared with other Speech Enhancement algorithms to prove the efficacy of enhancement.

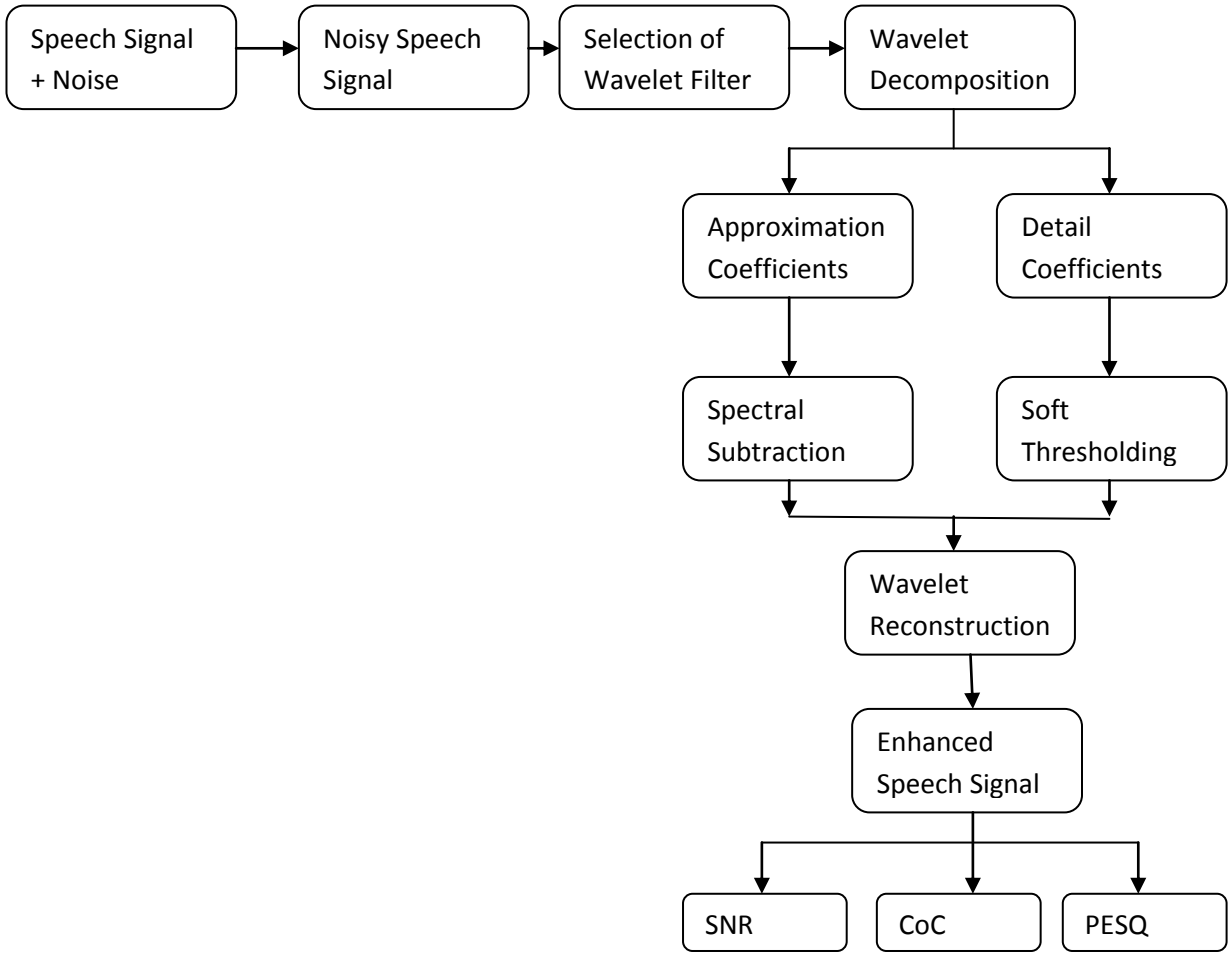

Fig -1: Steps for the Speech Enhancement

REFERENCES

[1] Boutaleb, R. , Meraoubi, H. ; Ykhlef, F. ; Benzaba, W. ; Boucetta, Y. ; Bendaouia, L., " Comparitive Performance Study between Spectral Subtraction and Discrete Wavelet Transform for Speech Enhancement", Computer Systems and Applications (AICCSA), 2013 ACS International Conference on 27-30 May 2013

[2] Dr. Mahesh S. Chavan, Mrs Manjusha N.Chavan & Dr. M.S.Gaikwad, "Studies on Implementation of Wavelet for Denoising Speech Signal", International Journal of Computer Applications (0975 – 8887)Volume 3 – No.2, June 2010.

[3] V.S.R Kumari, Dileep Kumar Devarakonda," A Wavelet Based Denoising of Speech Signal" International Journal of Engineering Trends and Technology (IJETT) – Volume 5 number 2 - Nov 2013.

[4] Roopali Goel, Ritesh Jain," Speech Signal Noise Reduction by Wavelets", International Journal of Innovative Technology and Exploring Engineering (IJITEE)

[5] Ritesh Jain, Suraiya Parveen , "Analysis of Different Wavelets by Correlation ",International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-2, Issue-4, April 2013.

[6] Guang-Yan Wang ," Speech enhancement based on the Combination of spectral subtraction and wavelet thresholding", http://ieeexplore.ieee.org/xpl/23-25 Oct. 2009

[7] 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

[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. Mohamed, F. Cervantes and H. Afifi. ``Real-Time Audio Quality Assessment in Packet Networks", In Network and Information Systems Journal, vol. 3, no. 3-4, 2000, pp. 595-609.

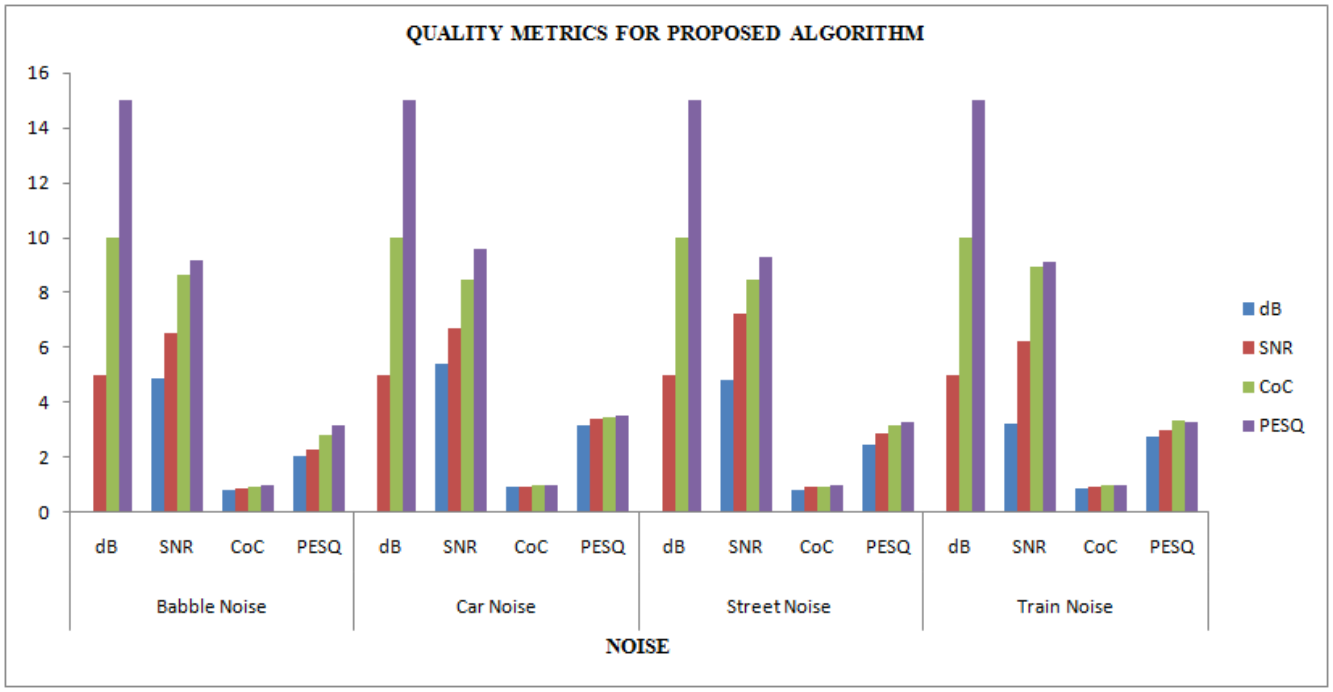

Chart -1: Quality metrics for Proposed Algorithm

BIOGRAPHIES

Dhivya.R completed B.E in Biomedical Engineering from Vellalar Engineering College, Erode and Currently pursuing M.E in the department of Biomedical Instrumentation Engineering in Avinashilingam University, Coimbatore.

Judith Justin graduated from Government College of Technology, Coimbatore and completed M.Tech in Biomedical Engineering from Indian Institute of Technology – Madras. Her research interests are in the field of Biosignal and Image Processing and Medical Instrumentation. She has published around 15 papers in National and International Conferences and has more than 12 papers published in Peer reviewed International Journals to her credit. She is a life member of ISTE and BMESI.