ISSN 2348 – 9928
IJAICT Volume 3, Issue 2, February 2016 Doi:01.0401/ijaict.2014.05.05 Published on 05 (2) 2016

# PERFORMANCE ANALYSIS OF FEATURE EXTRACTION AND ENHANCEMENT OF ESOPHAGEAL SPEECH

***Ms. P. Malathi***
***Research Scholar,***
***Electronics and Communication Engineering,***
***Easwari Engineering College***
***Chennai, Tamilnadu, India***

***Dr. G.R. Suresh***
***Professor,***
***Electronics and Communication Engineering,***
***Easwari Engineering College,***
***Chennai, Tamilnadu, India***

***Abstract— Laryngectomees are patients suffering from Laryngeal Cancer who undergo the removal of the Larynx and obviously lose a part of their vocal folds. The post surgery speech therapy for such patients are of three forms either of them preferred by the patients. They include esophageal speech, tracheoesophageal speech and electrolaryngeal speech or Alarynx speech. The features extracted in the above cited speech and the normal speech differs in their accuracy. This paper discusses the various analyses and enhancement techniques involved by various researchers.***

***Keywords— Esophageal Speech, Tracheoesophageal Speech, Electro Laryngeal Speech, Hidden Markov Model, Gaussian Mixture Model, Eigen Vectors.***

## I. INTRODUCTION

Laryngectomees are trained to produce ALaryngeal (AL) speech which is characterized by low intelligibility and naturalness caused by associated noise and low fundamental frequency. The features such as pitch, shimmer, jitter, HNR etc. Extracted from the ES (Esophageal Speech) speech differs from the features extracted from normal speech. García B., Vicente et al. had proposed the approximation of vocal tract using LPC [1][2] and has made a comparison of formant extraction from esophageal speech using LPC and wavelet transform[3]. Hironori Doi et al. have proposed a statistical approach using GMM [4].

M.Carello et al. have performed a comparative study of acoustic features esophageal and prosthesis speech such as frequency intensity, jitter, and shimmer, noise to harmonic ratio. Shunsuke Ishimitsu et al. has performed recognition of body conducted speech and has compared the recognition parameters of this method with MFCC and Perceptual Linear Prediction parameters. This paper involves the review of various techniques employed in the analysis of the acoustic features of esophageal speech and impaired speech and their performance analysis. The paper is organized as follows, Section 2 describes about esophageal speech, Tracheoesophageal speech and Electrolaryngeal speech, Section 3 elaborates on the techniques employed by various researchers for feature extraction. Section 4 deals with performance analysis of the various enhancement techniques. Section 5 elaborates on the results and discussion about the different analysis.

## II. ALARYNGEAL SPEECH

### *2.1 Esophageal Speech*

Esophageal speech appears to be more natural among all the other artificial laryngeal speech. The only drawback is that the speaker should possess speaking skills for which training will be provided after surgery. The spectral envelope varies randomly compared to normal speech. The excitation produced is found to be unnatural and less periodic which leads to poor pitch extraction. Hironori Doi et al. have observed that this pitch information is found in the spectral envelope.

### *2.2 Tracheoesophageal Speech*

Voice prosthesis is surgically inserted between the trachea and the esophagus. The prostheses inserted are a one way valve which allows flow of air from the lungs to the esophagus but prevents food or saliva from esophagus to the lungs. This has better quality than esophageal speech. It is characterized by longer phonatory duration, better intelligibility and louder voice. The speaking rate is high which improves the fluency.

### *2.3 Electrolaryngeal speech*

An electrolarynx is a handheld device with an electromagnetically vibrating membrane. The vibrations of this membrane are modulated by the articulatory organs into speech. This speech appears to be monotonous but offer rapid acquisition of speech. Sometimes there is an impossibility of transmitting vibrations through skin with scars and damaged tissues. Generally the electrolarynx has more drawbacks than the other two forms of AL speech. The exciting characteristic is that the extracted fundamental frequency exhibits high periodicity and the aperiodic components are extracted easily but with less information since the excitation is artificial.

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.05.05 Published on 05 (2) 2016

## III. FEATURE EXTRACTION IN AL SPEECH

Dubuisson et al. proposed a system to extract features such as spectral decrease and first spectral tristimulus in the Bark Scale and their correlation was found to be 94.7 % for pathological voices and 89.5 % for normal voices.

Ghoraani and Krishnan proposed another methodology for the automatic detection of pathological voices using adaptive time- frequency distribution (TFD) and nonnegative matrix factorization (NMF). The adaptive TFD dynamically tracks the nonstationarity in the speech, and NMF quantifies the constructed TFD.

Carello and Magnano measured the tracheostoma pressure at the time of phonation and the fundamental frequency, intensity, jitter, shimmer, and noise-to-harmonic ratio of oesophageal voices (EVs) and tracheo-oesophageal voices (TEPs). They found that the two signals resulted in equal fundamental frequency and the same harmonic components for each TEP subject considered.

According to Table 2, the TEP speech shows less standard deviation for frequency, jitter and shimmer. The ES shows less standard deviation for maximum phonation time. The Cross correlation was found between the Fourier Transforms of the ES speech and TEP speech and was observed to possess the same fundamental frequency and harmonic frequencies.

Hironori Doi et al. proposed a feature extraction for esophageal speech as its parameters vary with time unstably. The spectral segment vector i.e. MFCC will be extracted using STRAIGHT analysis from each frame and concatenated with the previous and the next frames.

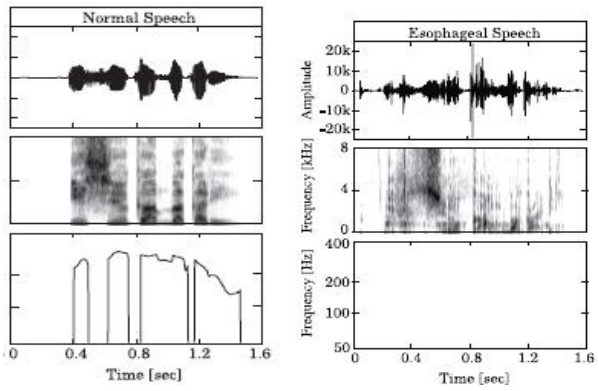


*Fig 1 : a. Spectrograms and $F_0$ contours of Esophageal speech*
*b. Spectrograms and $F_0$ contours of Normal speech*

The spectral segment vector is represented as follows,

$$X_t^{'} = CX_t + d;$$

$$X_t = [x_{t-i}^T, ...., x_t^T, ...., x_{t+i}^T]$$

Where $X_t$ is a joint vector by concatenating the spectrum vector of that frame with the previous and next spectrum vectors. C and d are the transformation matrices (Eigen vector matrix). The redundancy is overcome by reducing the dimension with Principal Component Analysis (PCA). Though the pitch cannot be extracted from the esophageal speech, this information is found from the spectral segment vector. Three different GMMs are used to estimate $F_0$ and the aperiodic components that capture the noise strength of an excitation signal in that frequency band.

## IV. ALARYNGEAL SPEECH ENHANCEMENT

Sheng Li et al. proposed an enhancement algorithm, multiband spectral subtraction method and have compared the performance of the proposed enhancement algorithm with traditional spectral subtraction method, basic Wiener filtering, and a noise-estimation algorithm. Spectrograms have been observed for both the residual noise and speech distortion. In addition, results are also measured objectively by Signal-to Noise ratio (SNR) and subjectively by Mean Opinion Score (MOS) in conditions of different additive white Gaussian noise as well as Babble noise (for MOS) for the algorithm evaluation.

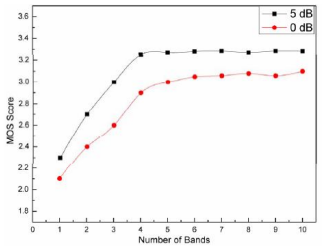


*Fig 2: MOS scores for different bands*

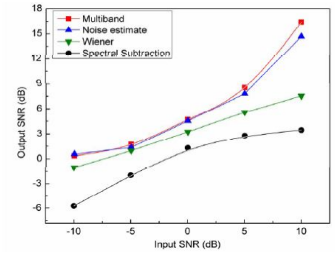


*Fig 3: SNR results at -10, -5, 0, +5,10 dB*

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.05.05 Published on 05 (2) 2016

*Table 1: Average and Standard Deviation for Patient Data, Vocal and Pressure Parameters*

| | Personal data | | | | Vocal parameters | | | | | | Tracheostoma pressure | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Age | Sex | Tracheostoma area | Fundamental frecuancy | Jitter | Jitter perc. | Shimmer | Shimmer perc. | NHR | Maximum phonation time | Tracheostoma pressure | Acoustic pressure/ Tracheostoma pressure |
| | | | [$cm^2$] | [Hz] | [ms] | [%] | [Pa] | [%] | [−] | [s] | [Pa] | [−] $* 10^{(-7)}$ |
| EV average | 64.86 | — | 1.25 | 86.569 | 26.95 | 22.69 | 0.00029 | 0.39 | 1.459 | 0.84 | — | — |
| EV standard deviation | 9.72 | — | 0.52 | 34.063 | 9.96 | 6.24 | 0.00024 | 0.24 | 0.830 | 0.36 | — | — |
| TEP average | 68.57 | — | 1.58 | 91.139 | 8.38 | 7.87 | 0.00016 | 0.31 | 1.322 | 17.30 | 3728 | 2.0053 |
| TEP standard deviation | 8.04 | — | 0.61 | 23.089 | 5.84 | 5.19 | 0.00012 | 0.12 | 1.188 | 15.23 | 1358 | 1.2518 |

Both the objective and subjective test results suggest that a better noise reduction effect was obtained and the perceptually annoying musical noise was efficiently reduced (especially in the high-frequency regions), with little distortion to speech information and has a strong flexibility to adapt itself to rigorous speech environment as compared to the other standard speech enhancement algorithm.

Martin Hagmuller has proposed the enhancement of Alaryngeal speech based on Time-domain pitch-synchronous overlap-and-add (TD-PSOLA) to lower the pitch and period enhancement to reduce breathiness.

Rym Haj Ali, Sofia Ben Jebara have enhanced the Elaryngeal speech by enhancing the excitation source and formant bandwidth without increasing background noise.

C.Ganesh Babu et al. has used an Ephraim Malah filter to enhance the Alaryngeal speech quality and statistical modeling for speech recognition.

*Table 2: Recognition Accuracy Compared Using Kalman Filter And Ephraim Malah Filter.*

| Noise | KF | EM | %Improvement |
|---|---|---|---|
| Airport | 1 | 21 | 95.23 |
| Babble | 28 | 37 | 24.3 |
| Exhibition | 6 | 17 | 64.7 |
| Street | 9 | 23 | 60.86 |
| Restaurant | 23 | 42 | 95.23 |
| Station | 5 | 9 | 44.4 |
| Car | 3 | 24 | 87.5 |

Table 3 depicts the improvement in recognition of Alaryngeal speech using EM filter in various noisy environments.

## V. RESULTS AND DISCUSSION

From the research carried out by many authors it has been observed that the features extracted from Alaryngeal speech are fundamental frequency, jitter, shimmer, formants and MFCC. Among the methods discussed, the spectral segment vector obtained from STRAIGHT analysis method is found to outperform the other methods with better quality and accuracy since it takes into consideration the time varying characteristic of the esophageal speech and also calculates the pitch from the spectral component vector.

The historical methods to enhance the A laryngeal speech include the PSOLA method and to enhance the extracted excitation and the formant frequencies. This method undergoes a drawback of the excitation lacking periodicity and the formant frequencies lacking accuracy. Hence Multiband Spectral Subtraction Method seems to outperform the others overcoming the drawbacks encountered in the previous methods which is evident from the MOS and SNR

## References

[1] García, B., Vicente, J. & Aramendi, E. “Time-Spectral Technique for Esophageal Speech Regeneration” Biosignal ’02, 2002.

[2] García, B., Vicente, J., Ruiz, I., Alonso, A. & Loyo, E. “Esophageal Voices: Glottal Flow Restoration”, ICASSP 2005.

[3] Begona Garcia Zapirain, Ibon Ruiz, Amaia Mendez, “Oesophageal Speech’s Formants Measurement Using Wavelet Transform”,

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.05.05 Published on 05 (2) 2016

Advances In Wavelet Theory And Their Applications In Engineering, Physics And Technology, Intech Open Science.

[4] Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, And Kiyohiro Shikano, "Statistical Approach to enhancing Esophageal Speech based on Gaussian Mixture Models", *ICASSP2010*

[5] Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, And Kiyohiro Shikano, "Esophageal Speech Enhancement Based on Statistical Voice Conversion with Gaussian Mixture Models", IEICE Trans. Inf.&Syst, Vol E93-D, No.9 September 2010

[6] Juan Ignacio Godino-Llorente, Pedro Gomez- Vilda And Tan Lee, "Analysis And Signal Processing Of Oesophageal And Pathological Voices", Eurasip Journal On Advances In Signal Processing,2009

[7] Massimiliana Carello And Mauro Magnano, "A First Comparative Study Of Oesophageal And Voice Prosthesis Speech Production", EURASIP Journal on Advances in Signal Processing, 2009.

[8] Shunsuke Ishimitsu, Kouhei Oda And Masashi Nakayama, "Body-Conducted Speech Recognition in Speech Support System for Disorders", International Journal Of Innovative computing, Information And Control, August 2011

[9] Sheng Li, MingXi Wan, and SuPin Wang, "Multi-Band Spectral Subtraction Method for Electrolarynx Speech Enhancement", Algorithms 2009, 2, 550-564, ISSN 1999-4893