ALT Speech Recognition System using F0 Improvement and Spectral Tilt Method

Inbanila K, KrishnaKumar E

ABSTRACT--- ***Human Beings use voice as the medium for communication. Human Speech is a very complex signal with multiple frequencies, amplitudes and intensities that mix up to convey specific information. In international terminology, voice disorders are described as dysphonia. Various dysphonia's are clearly organic origin due to nervous, muscular, neuro or cellular degenerative disease affecting the body or it is from local laryngeal changes. Other dysphonia's having no visible laryngeal causes are grouped as non organic involving habitual dysphonia's that arise from faulty speaking habits or the psycho genic dysphonia's that stem from emotional causes. This paper looks at a speech recognition system for disordered speech generated by Physically Disabled people using Artificial Larynx Transducer (ALT) device from the perspective of Speech Signal Processing. From the ALT speech features like formant, pitch and spectral tilt is estimated. For formant frequency estimation RNN technique is used. Before training the system pitch frequency improvement is accomplished. Now the features and homomorphic based coefficients are used for training the system. The same operation is performed during the test phase and compared with the training set. Comparison and decision making is accomplished using distance estimator.***

Index Terms— ALT speech, Formant frequency, Spectral Tilt, Disordered speech, Healthy speech (HE speech), DREL noise.

I. INTRODUCTION

The human beings affected by laryngeal cancer undergo a surgery for removal of entire larynx, thereby loosing the capability to produce speech.[1]. It becomes essential to restore the speech for them. Regular esophageal speech and trachea esophageal speech are the methods used for restoration [4]. In general, the above methods are very uncomfortable for the patients in anatomical and personal sense [6]. So widely Electro larynx (EL) is used for speech production. It is a small handheld battery operated device which has preset pitch band which can be adjusted for male, female and children [1]. Using EL, long sentences can be communicated [7]. There are 2 types of EL. One is the neck type and the other is intra-oral type. Among the two, neck type is most widely used. During the utterance, the device is kept near the glottis [8]. The sound is let in to oral and pharyngeal cavity. The vibration created by this electro mechanical device is transmitted through the tissues of neck. The user should modulate this in to speech by proper movement of the articulators like lips, tongue, teeth, jaw and velum [2]. This is shown in Fig 1. The substitution voice sound produced by this device is monotonous and has a robotic speech nature. So the naturality of the speech will be missing. The other drawback of this method is the directly radiated EL (DREL) noise [5]. The direct noise induces irritation and fatigue for the listener. Removal of DREL noise using spectral subtraction [11] partially helps to overcome the limitations. The usage of the ALT is shown in Fig.2. In the previous our work we have extensively explained the DREL noise removal method and RNN based method for Formant estimation. In this paper an attempt is made on F0 estimation, improvement of F0 and spectral tilt estimation. After the above estimations the system is trained with the features, Linear predictive cepstral coefficients, Homorphic coefficients. During the test phase all the above parameters are estimated and best match is found using the distance estimation algorithm developed. This process completes the speech recognition operation of an ALT speech.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author
Inbanila K*, CHRIST (Deemed to be University), Bangalore, Karnataka, India.
(Email: inbakrish_11@rediffmail.com)
KrishnaKumar E, East Point College of Engineering and Technology, Avalahalli, Bengaluru, Karnataka, India.

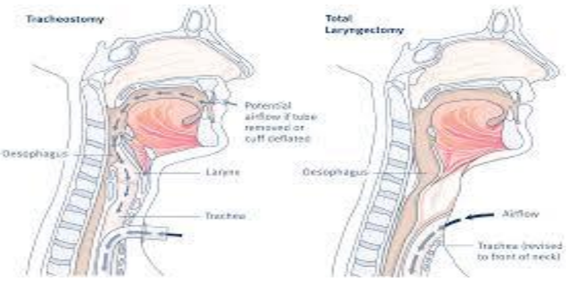
Tracheostomy
Total Laryngectomy
Potential airflow if tube removed or cuff deflated
Oesophagus
Larynx
Trachea
Oesophagus
Airflow
Trachea (revised to front of neck)

Fig.1: The anatomy of the total Laryngetomy [13]

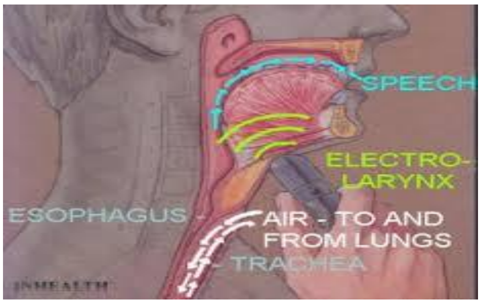

Fig.2: ALT device [14]

The general problem of ALT speech intelligibility is very poor. An attempt is made to improve the pitch contour which will contribute for speech quality enhancement. Usually for applications like voice morphing, pronunciation correction and speech perception a mechanism is adapted for changing the prosody and speech characteristics. The most important part is the change of pitch in an utterance. Pitch extraction is an essential process for speech modification. Similarly formant frequency estimation helps **in** speech characterization for different applications of speech enhancement of ALT speech, speech synthesis from ALT speech and Automatic speech recognizer of ALT speech. The block diagram Fig.3 explains the overall process of Speech recognition system. The speech is subjected to pre-processing such as sampling, windowing, silence and speech discrimination.

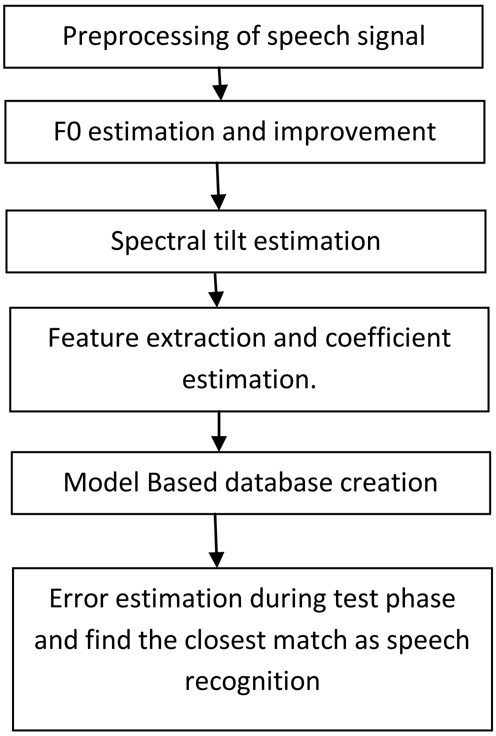
Preprocessing of speech signal
F0 estimation and improvement
Spectral tilt estimation
Feature extraction and coefficient estimation.
Model Based database creation
Error estimation during test phase and find the closest match as speech recognition

Fig.3: Block Diagram of the speech recognizer system

The speech signal having both voiced and unvoiced speech is analysed in time domain and frequency domain for estimation of F0 and formant respectively. The above analysis also yields coefficients which has vital role on recognition system. The formant listing is used as such whereas methods are adapted for F0 improvement. The spectral tilt estimation is also used as feature. During test phase all parameter are estimated and distance estimation is done. The paper is organised by giving an overview of how template based system works considering characteristic value approach and non characteristic value approach followed by methods of F0 estimation and improvement. Significant discussion is done in understanding the purpose of spectral tilt estimation and method to do so. Next the procedure done during the test phase is explained including distance estimation and concluded with results and discussion for the recogniser system.

II. DATABASE

In this paper totally 200 ALT speech and 200 ALT speech samples are taken for training and testing respectively. The utterance made by both categories is the same. The speech duration is 10 sec in average and recorded in noise free environment. The samples have both gender such as male and female with different age group. The database is taken from PTDB-TUG: Pitch Tracking Database from Graz University of Technology.

III. SYSTEM MODELLING

Characteristic value approach

In this method the modelling of speech utterance is done considering the linear and non linear nature of speech. Speech signal in general is a highly non stationary signal having both linear and non linear dynamics with piecewise constant approximations and assumptions. Together in this method the speech is broken and stored in the form of words which in general is achieved from language models, then word is broken into phoneme got from the lexicon model and further divided in to states got from HMM model or similar methods. This nature of consideration makes it perform well in simple recognition systems. When this method is tried to be adapted for ALT speech segregation as word to create language model, as phonemes to get lexicon model and get states from HMM model is not easy as the articulation is hugely different. So when the test sample has difference in articulation the method fails. In this paper, ALT speech is considered during the test sample which has huge variation in regard of utterance when compared to HE speech. Together in practise there are situation when hypo and hyper utterance is dealt with. During training operation the utterance is made very clear and intelligible as most time it is done as a recording. Whereas during test sequence the utterance need not be so well articulated because it is a natural conversation. During this process, there is situation where the syllable is not well clearly uttered and distinction of the syllable will not be sufficient. This is hypo utterance. On the other hand when the environment is not calm much time the speaker speaks loudly with higher stress for each utterance making it hyper utterance. In the case of ALT sample also the same problem exists. Hence characteristic value approach may not give success for all test utterance.

Non Characteristic value approach

From the previous paragraph it is clear that characteristic value approach is not a suitable method for ALT speech recognition. So an attempt was made for Non characteristic value method. In this process the speech is broken and stored as a sequence of features rather than states.

During the test phase the features are compared with the stored features and not with the states as done in the previous method. As the comparison happens with the stored features, the hypo and hyper utterance issue dealt in the previous methods can be overcome if a corrective warping technique is used.

In this method also Dynamic Time warping(DTW) is used in order to have the utterance duration same and it was discussed in the previous work of us.

IV. F0 ESTIMATION AND IMPROVEMENT

F0 Estimation

F0 is otherwise called as fundamental frequency or pitch frequency. This feature helps in speaker identification. When a word or sentence is uttered and if it is stored as a signal, it can be observed that the speech signal is a convoluted signal created by many signals with variable frequency component. But when the same signal is observed for large duration, a periodicity is present in the signal. This periodicity envelope is the parameter that contributes for the pitch frequency. This parameter can be estimated easily using any time domain analysis of the speech signal. In general auto correlation method is used for this purpose. When short time speech signal is subjected to autocorrelation, peaks appear at regular interval. The inverse of the duration between the peaks contribute for the pitch frequency. But if the speech signal has many formant frequencies, the autocorrelation creates confusing peaks leading to wrong estimation of pitch. In this regard centre clipping method is used. In this method a voltage threshold is fixed for every short time speech. The portion of the signal which is above the upper threshold and below the lower threshold is segregated. The auto correlation operation happens to this smoothened speech. This operation helps to remove the confusing peaks available in the previous method. This method is a time domain analysis method. For the application discussed in this paper this method of pitch frequency estimation is sufficient because this method always preserves all the information about the harmonics and also the formant amplitudes. Though the phase parameter is ignored, the pitch estimation is not going to be affected much. Similarly F0 contour estimation also is a must for the proposed speech recognition system. In the autocorrelation method or average magnitude difference function method the speech signal is sampled at 16 kHz. The frame length is maintained at 256 points [3].

F0 Improvement:

Pitch frequency estimation discussed in the previous paragraph is adapted for the ALT speech. When the pitch contour shown in fig 4 is observed closely with regard to different speech samples taken from female and male samples with different utterance, it is evident that the pitch frequency contour is not overlapping with the HE speech as well as there are many unusual pitch shoots which disturbs the listener. The same is shown in the graph that visualizes the pitch of same utterance by ALT and HE speech as fig 4 and fig 5 respectively.

The usual pitch frequency range of 50 Hz to 150 Hz for adult male and 150Hz to 250 Hz adult female is deviated in some instant of time. The input to this block is the speech signal for which gender categorization is already adapted. Together the voiced component of the speech signal is only subjected to pitch modification. Identification of voiced speech is achieved by estimating the average energy of the short time speech signal. If the average energy is more than the threshold, it is considered to be speech signal and not a silence. The identified speech signal is let to the block for identification of periodicity. If periodicity exists it is considered as voiced speech sample. If periodicity does not exist it is considered as unvoiced speech sample. So the sample of the speech that is above average energy threshold and has periodicity is sent for pitch period modification. Here two methods are proposed to remove these deviations and bring the speech F0 contour as smoothened translation.

Method 1:

This method adapts the scaling off method at the positions of deviation. Initially the gender identified speech pitch parameter value is estimated against every instant of time. As the gender is known the threshold of pitch frequency is also known. Now the pitch value is compared with the adjacent pitch value. If the transformation of the value is optimum and within threshold it is retained with the same value. If the adjacent value deviation is more or if the value is above threshold then pitch modification is used.

The modification is made based on the formulas mentioned below.

$P(n)= P(n)/i$ if $P(n) – P(n-1) >=$ threshold

$P(n)=P(n)$ if $P(n) – P(n-1) =$ threshold

Where P(n) represents the current pitch value. i represents the scaling factor and it is a constant ranging from 0.5 to 1.The usage of this formula proportionally decreases the pitch deviation to appreciable amount.

Here 3 different conditions occur. They are:

1) Adjacent values are within the deviation limit and within threshold. In this case the values are retained as such.

2) The second sample is far above the proper previous sample value. In this case current sample is subjected to scaling based on the formula and new estimated value is assigned for that instant.

3) Current sample is proper when compared to past sample. In this case the past sample estimated value only to be considered for comparison.

Method 2:

Here Pitch synchronous method is adapted. This aims in increase or decrease of periodicity duration as per the requirement of the sample. Increase in duration is done by padding extra values within the periodicity duration and decrease is by removal of samples in the periodicity duration. The increase or decrease of duration is based on the formula mentioned below.

The compression or expansion takes place based on the window length.

For small window length

$Y(nL,w)= X(f\, nL,w)$

where Y is the pitch scaled output of the input X and f is the scaling factor.

For long window length

$X(nL,w) = U(nL,w)H(nL,w)$

Where U is the short time transform of the source and H is the short time transform of vocal tract spectrum. After the modification new sequence is synthesized by overlap method.

The pitch contour modified by the above two methods are compared and average of that is considered as the final pitch frequency only in the positions of original deviation. The below figures compares the pitch frequency for HE and Alt speech. Fig 4 represents the signal representation of a female ALT speech. The graph shows the speech signal, voice activity and pitch contour. In the contour there are many instants at which the pitch frequency shoots above threshold range. Fig 5 is the representation of all the parameter mentioned above but for a HE speech. So the algorithm compares the pitch frequency of current instant and performs smoothening operation. Fig 6 represents the smoothened contour achieved using method one and two and it is obvious that the peak shoots are smoothened.

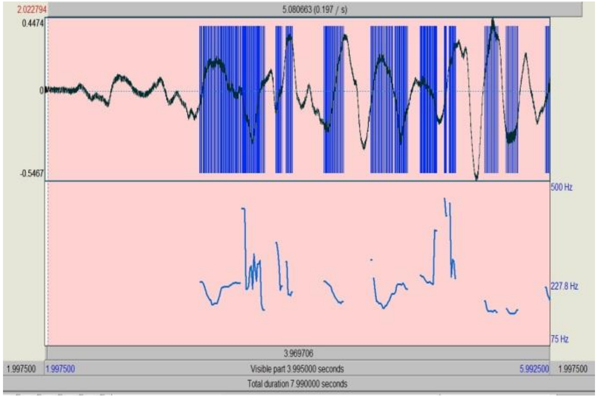

Fig.4: F0 Contour with peak shoots for ALT Speech

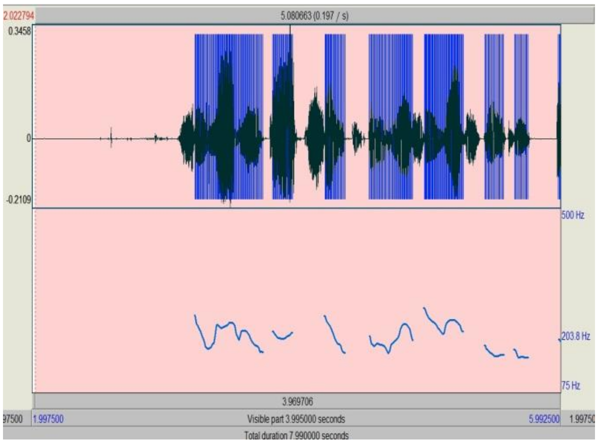

Fig.5: F0 Contour with peak shoots for HE Speech

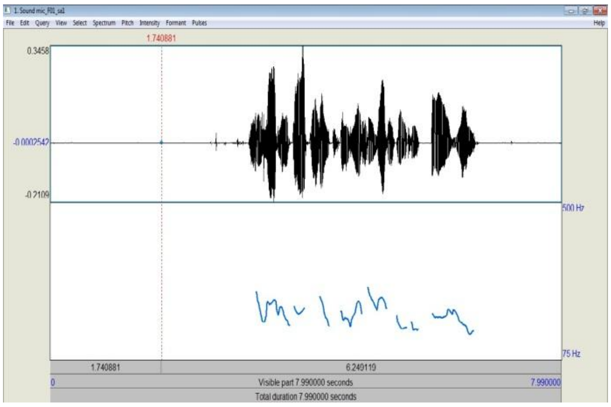

Fig.6: Modified F0 Contour using method 1

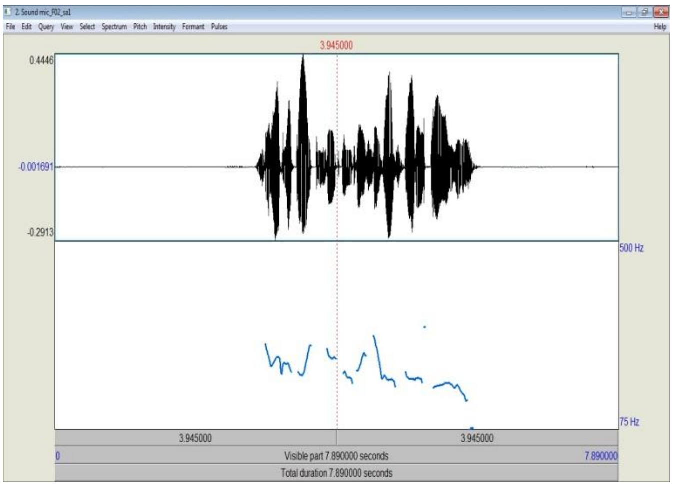

Fig.7: Modified F0 Contour using method 2

V. SPECTRAL TILT ESTIMATION

Speech is a multiple frequency convoluted signal. The frequency range from 50Hz to 4000Hz in general. During hypo utterance the density of power is comparatively low when compared to that hyper utterance. And in both the cases, power spectral density will be more at low frequency when compared to high frequency. This nature is called as spectral tilt[12]. This parameter is of significance in the recognition system of ALT speech because most of the utterance is hyper in nature involving more power. The estimation of this feature and using during distance estimation improves the capability to determine the closest match. Two methods are adapted in the estimation and both use the LP coefficients estimated from all pole model. All pole model works with the fundamental concept that current sample can be predicted from past original samples. This concept is evolved because when the articulation is done the vocal tract movement slowly move from one syllable to the other without bringing any abrupt stops in the vibration. Similarly here speech is considered in short duration. The average energy in the short duration is used as a weighting factor thereby smoothening the spectral envelop for higher order all pole system also. In the second method two stages of LP analysis happens. In the first stage the order is higher and in our work it is kept as 22 . The coefficient estimated using this method is given to inverse filter and made smoothened. The output of the inverse filter is designated as modified LP inverse coefficient. The modified coefficients is scaled and appropriately let in to the second stage of LP analysis to synthesize back the signal. This process helps in spectral tilt estimation.

VI. REFERENCE MODEL BASED SYSTEM

As discussed in the first section the recognizer system works with non characteristic value method. Here the non characteristic values are the formant frequency list, modified pitch contour, LPCC and Homomorphic coefficients. The acoustic features are segmented and given label. All these constraint values are estimated during the training phase. The sampling rate is at 16 kHz with 256 samples overlap for all calculation. During the test phase the estimated parameter of test sample is compared with training sample after gender categorization is done. In this full process no attempt of merging the label happens. We opt to store the entire database in the original order, with the advantage that the complete acoustic-phonetic context of each template remains available. In the process of local distance estimation the following method is adapted. Here the distance estimation is calculated between matrix test matrix y to that of stored reference matrix x. During this operation a set of weighted matrix is also used to compute the deviation between the column matrix [10].

$$d(x;y) = (x-y)^T\Lambda(x-y)$$

and as shown in the above equation by using the inverse of the covariance the Euclidean distance estimation is done. The complexity of distance estimation for the speech data increases as limitation is created because of space between the frames. This can be overcome by creating a dependency on the feature vectors that are being compared. This is achieved by partitioning it as disjoint class. In this case the distance measure is estimated using the below formula.

$$d(x;y) = (x-y)^T\Lambda_{\mathrm{k}(y)}(x-y)$$

To minimise the error in distance estimation an additive bias term is used to compensate the process of making the x and y matrix within the same class and metrics. That is mentioned in the formula below.

$$d(x;y) = (x-y)^T{\sum}_{\mathrm{k}}^{-1}(\mathrm{x}-\mathrm{y}) + \log\left|{\sum}_{\mathrm{k}}\right|$$

$$d(x;y) = \sum_{l=1}^{M}\left(\frac{x_l-y_l}{\hat{\sigma}_{k,l}}\right)^2 + \log\left(\prod_{t=1}^{M}\hat{\sigma}_{k,l}^2\right)$$

The estimate made using the above formula is validated with the performance estimation made by Parzen density estimate. The same is mentioned in the table 1.

$$\hat{f}_{Parzen}(x) = \frac{1}{|k|h^M}\sum_{y_i\in k}\mathcal{N}\left(\frac{x-y_i}{h};0,{\sum}_{\mathrm{k}}\right)$$

VII. RESULTS AND DISCUSSION

This section confirms the performance of the proposed recognizer system considering 2 methods of validation. In both the validation method, 5 sample utterances is taken. Attempt was made to ensure that both male and female voice is present and the age group is taken below 40 years and above 40 years. The table 1 represents the estimated distance of the utterance between the test and reference. Though the test sample is subjected to gender categorization and further compared with reference data, the table shows the comparison for 5 sample set only. From the table it is clear that the distance estimated using inverse covariance method yields closest match with the proper utterance. In this case UT1 to UT5 are the difference test utterance made similar to the reference utterance. In other way, UT1 and UT1-ref are the same utterance but made at different duration. Similar is the case for all the remaining articulation. So it is evident that the speech recognizer gives closest match or smallest distance with the same utterance. Above that the deviation between the best match and other utterance is more which makes the thresholding easy.

Table 1: Estimated distance between training dataset and test data set for 5 utterances

	UT1-ref	UT2-ref	UT3-ref	UT4-ref	UT5-ref
UT1	3.3e-04	0.42	0.4	0.38	0.1
UT2	0.52	2.3e-04	0.7	0.11	0.41
UT3	0.41	0.87	1.2e-03	0.5	0.63
UT4	0.35	0.17	0.45	3.8e-04	0.12
UT5	.12	0.42	0.6	0.1	8.2e-04

Table 2 confirms the performance of the recognizer system under different combination of features. The validation is done for the system with F0 contour improvement (WF0), without F0 contour improvement (WoF0), without spectral tilt estimation (WoST) with spectral tilt estimation (WST) and using all the features. The table provides the %error introduced by the system for the above mentioned combination.

Table 2: % error in the recognition system for 5 utterance

%error	UT1	UT2	UT3	UT4	UT5
WoF0	12	14	11	128	11.4
WF0	9	7	10	12	14
WoST	13	13.53	11	14	15
WST	6	8	10	7	11
WF0+WST+Coeff	4	3	4	2	5

It is obvious that percentage error is minimum for the data set size with the combination of non characteristic parameter.

VIII. CONCLUSION

In this paper an attempt is made to create of recognizer system for ALT speech. Though the advent of HMM (Hidden Markov Model), GMM (Gaussian Mixture Model) and deep learning algorithm has improved the performance of speech recognition system considerably, improvement in ALT speech is still not appreciable. A method is proposed to identify the details spoken using ALT device in the presence of reference frame of healthy speech of same utterance. This signal processing approach has given new path in managing the constraints of poor intelligibility, monotonicity and fatigue created in hearing ALT speech with sufficient features. Evolution of HMM for disordered speech may help in achieving the milestone in this context. This open ups new avenues of research in the area of substitution speech.

REFERENCES

1. Anna Katharina Fuchs, Martin Hagmuller, “Learning an Artificial F0-Contour for ALT Speech”, Interspeech , 2012.
2. Barney HL, Haworth FE, Dunn HK,”An experimental transistorized artificial larynx”, Bell Syst Tech J 1959;38:1337–56.
3. Chih-Ting Kuo, Hsiao-Chuan Wang,”A Pitch Synchronous Method for Speech Modification”, Interspeech 2008.
4. Hillman R,Walsh M,Wolf G, Fisher S, HongW, ”Functional outcomes following treatment for advanced laryngeal cancer. Part 1. Voice preservation in advanced laryngeal cancer. Part II. Laryngectomy rehabilitation: the state-of-the-art in the VA system”, Ann Otol Rhinol Laryngol 1998;107(Suppl 172, Pt 2):1–27.
5. K. Fuchs, J. A. Morales-Cordovilla, and M. Hagmüller, “ASR for electro-laryngeal speech,” in IEEE Workshop on Automatic Speech Recognition and Understanding Workshop, Aug. 2013.
6. Karen C, Joel M, “Utilization of microprocessors in voice quality improvement: the electrolarynx”, Curr Opin Otolaryngol Head Neck 2000;8:138–42
7. Lauder E. The laryngectomee and the artificial larynx—a second look”, J Speech Hear Disord 1970;35:62–5.
8. Liang Hong, “Independent Component Analysis Based Single Channel Speech Enhancement Using Wiener Filter”, Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology ,2003.
9. Liu paer , Rothman H. “Acoustic analysis of artificial electronic larynx speech.In: Seikey A, editor. Electroacoustics analysis and enhancement of alaryngeal speech”, Springfield, IL: Charles Thomas; 1982. p. 95–118.
10. Mike Matton, Dirk Van Compernolle, Ronald Cools,” A Minimum Classification Error Based Distance Measure for Template Based Speech Recognition”, Interspeech 2008.
11. R.W. Schafer and L.R. Rabiner,”system for Automatic formant analysis of voiced speech,” J.Acoust.Soc. AM, Vol.47, pp. 634-640,February 1970
12. Sofoklis Kakouros, Okko Rasanen, Paavo Alku,”Evaluation of Spectral Tilt Measures for Sentence Prominence Under Different Noise Conditions”, Interspeech 2017.
13. http://bme240.eng.uci.edu/students/06s/paytonl/CurrentTherapies.html
14. https://www.ambulance.qld.gov.au/docs/clinical/cpg/CPG_Tracheostomy%20emergencies.pdf