Neural Network based New Bionic Electro Larynx Speech System

Ms. L. Padmini,
Assistant Professor, Department of Electronics and Communication Engineering, K.Ramakrishnan College of technology, Trichy.

Akila. S, Karpagameena. U, Kasthuri V. K
Department of Electronics and Communication Engineering,K.Ramakrishnan College of technology, Trichy.

***Abstract*—Persons who have lost their larynx and thus speech functionality need to use a substitution voice to regain speech. The electro-larynx (EL) is a widely used device but is known for its unnatural and monotonic speech quality. Previous research has addressed these problems, but until now no significant improvements could be reported. In this paper, we review existing approaches and summarize the most important findings. Subsequently, we piece together an overall speech system, which integrates several parts to significantly improve EL speech. The existing approach uses the Support vector machine (SVM) for classification. In our proposed method, we use k-NN (kernel nearest neighbor) for speech segmentation, MFCC (Mel frequency cepstral coefficients) for feature extraction and Neural Network for speech training. The accuracy obtained in our proposed method is higher than that of the existing method. Listening testsserve as a proof of concept for the resulting EL speech system, which confirm that the proposed system is very promising.**

Keywords- Electro larynx, Support Vector Machine (SVM), kernel Nearest Neighbor (k-NN), Mel frequency cepstral coefficients (MFCC), Neural Network

1. INTRODUCTION

The larynx often called the voice box, is one of the organs that helps us to speak. It contains the vocal cords. It is in the neck, above the opening of the trachea(windpipe). There, it helps keep food and fluids from entering the trachea. Cancer that starts in the larynx (laryngeal cancer) is treated differently based on which section it starts in.The larynx produces sound for speaking. The vocal cords move and come together to change the sound pitch of our voice. The laryngeal cancer starts in the lower part of the throat. Cancer starts when cells in the body begin to grow out of control. The American cancer society's most recent estimates for laryngeal cancer in the united states for 2017 are:

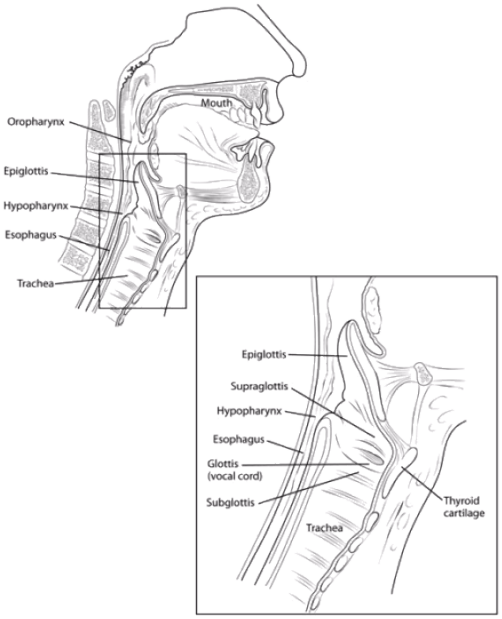


Fig. 1. Larynx parts

- About 13,360 new cases of laryngeal cancer (10,570 in men and 2,790 in women)
- About 3,660 people (2,940 men and 720 women) will die from laryngeal cancer

About 60% laryngeal cancers start in the glottis, while about 35% develop in the supragloittic area. The rate of new cases of laryngeal cancer is falling by about 2% to 3% a year, most likely because fewer people are smoking.

If people lose their larynx they will consequently also lose their ability to speak. There are three alternatives for people to regain their speech. The first method is *esophageal voice*. For this method, air is firstly gulped and then released in a controlled manner. Instead of the vocal folds, the tissue of the pharyngo-esophageal (PE) segment in the pharynx vibrates. The second method is *trachea esophageal voice*, where a shunt valve is placed between the trachea and esophagus. Due to the shunt valve, speech can be generated with the air coming from the lungs. The third method is the transcutaneous *Electro- Larynx device* (EL) (figure 2)

Fig. 2. Conventional EL device

This is a small, hand-held and battery-driven device. The vibrating coupler disk of the device is held against the neck. The signal of the coupler disk is carried into the vocal tract which filters the signal in a similar manner to healthy speech production. The disadvantage of all these substitution voice types is the poor quality of the resulting speech. The preferred substitution voice depends on the user, his/her ability to learn how to use the substitution voice and other factors such as degree of surgery (what and how much of the affected muscles and tissue needed to be removed) or the length and type of radiation therapy and ultimately the health system involved. The major drawbacks of EL speech are 1. the directly radiated electro-larynx noise (DREL) of the device itself, 2. the unnatural, monotonous quality of speech and 3. the need of one hand to operate the device. DREL reduces intelligibility and disturbs the speech quality.

2. RELATED WORKS

EL users are not at all satisfied with the quality of EL speech. The poor quality of EL speech is a result of the limited performance of conventional EL devices, the loss of the fine control of pitch, amplitude, and voice onset and offset timing. The users confirmed a deficit in voice-related segmental (e.g., voiced-unvoiced distinctions for consonants) and supra-segmental (e.g., intonation, syllabic stress) speech parameters. In the thesis of listening tests, acoustic analysis and acoustic modeling was carried out to investigate the properties of EL speech. 10 listeners judged the addition of pitch information to be the most important benefit. Removing DREL and correction for a lack of low frequency energy would also improve the speech. EL speech improvement methods based on signal processing approaches thus focus on the artificial excitation signal. In the following, we will summarize how researchers have tackled the most prominent drawbacks of the EL speech system. In order to introduce a changing fundamental frequency pattern we proposed an automatic procedure based on statistical models. The estimation of an artificial *f*0 contour using GMMs is, amongst others, inspired by who used melfrequency cepstral coefficients (MFCCs) for prediction of *f*0 and voicing in unconstrained speech. Basically statistical models are matched to static input here. This means that features are calculated on a frame by-frame basis. However, this does not correspond to the nature of speech, because phonemes are not independent but strongly depend on the phoneme before and afterward (co-articulation). This fact can be considered using dynamic first or higher-order differences of the MFCCs (Δ features). GMMs are probabilistic models. They are useful for modeling arbitrary probability distributions. A GMM consists of a weighted linear combination of *K* multivariate Gaussian probability density functions.

3. PROPOSED WORK

3.1 PREPROCESSING

Preprocessing is done to analyze the amplitude of the input and filtering the noise. Pre-emphasis is achieved with a pre-emphasis network which is essentially a calibrated filter. The frequency response is decided by special time constants. The cutoff frequency can be calculated from that value. In transmitting signals at high data rates, the transmission medium may introduce distortions, so pre-emphasis is used to distort the transmitted signal to correct for this distortion. When done properly this produces a received signal which more closely resembles the original or desired signal, allowing the use of higher frequencies or producing fewer bit errors. Pre-emphasis is a very simple signal processing method which increases the amplitude ofhigh frequency bands and decreases the amplitudes of lower bands. In simple form it can be stated as,

$yt = \alpha xt + (1-\alpha)x\ t-1$

Pre-emphasis is employed in frequency modulation or phase modulation transmitters to equalize the modulating signal drive power in terms of deviation ratio. The receiver demodulation process includes a reciprocal network, called a de-emphasis network, to restore the original signal power distribution.

3.2 SPECTROGRAM ANALYSIS

A spectrogram (fig 3) is a visual representation of the spectrum of frequencies in a sound or other signal as they vary with time or some other variable. Spectrograms can be used to identify spoken words phonetically, and to analyze the various call of animals. A common format is a graph with two geometric dimensions: the horizontal axis represents time or rpm, the vertical axis is frequency.

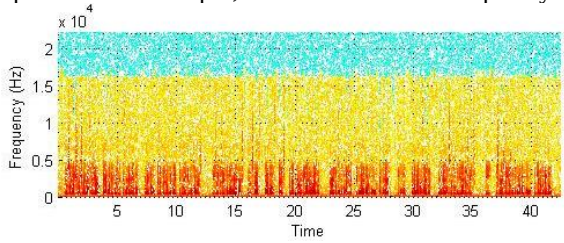


Fig.3. Spectrogram

Spectrograms are usually created in one of two ways: approximated as a filter bank that results from a series of band-pass filters or calculated from the time signal using the Fourier transform. These spectrums or time slots are then laid side by side to form the image or slightly overlapped in various ways, i.e windowing. This process essentially corresponds to computing the squared magnitude of the short time Fourier transform (STFT) of

the signal *s(t)* – that is, for a window width *ω*,***spectrogram(t,ω) = |STFT(t,ω)|².***

3.3 SEGMENTATION PROCESS

Speech segmentation is a subfield of general speech perception and an important sub-problem of the technologically focused field of speech recognition, and cannot be adequately solved in isolation. In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.

The classification process is based on four components:

1. Class

Categorical dependent variable in the form that represents the 'label' contained in the object. For example: heart disease risk, credit risk, customer loyalty, the type of earthquake.

2. Predictor

The independent variables are represented by characteristic (attribute) data. For example: smoking,
drinking alcohol, blood pressure, savings, assets, salaries.

3. Training dataset

One data set that contains the value of both components above are used to determine a suitable class based on predictor.

4. Testing dataset

Containing new data which will be classified by the model that has been made and the classification accuracy is evaluated.

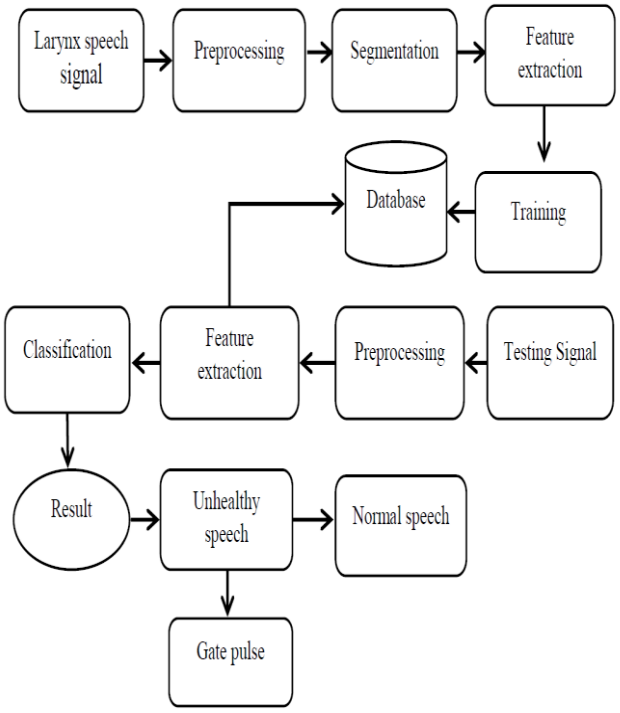
.

K-nearest neighbor is a method to perform the classification of objects based on the learning data that were located closest to the object. Learning data projected into many dimensional space, where each dimension represents the features of the data. The space is divided into sections based on the learning data classification. K value is best for this algorithm depends on the data, in general, the value k may reduce the effect of noise on the classification, but it makes the boundaries between each classification becomes more blurred.

3.4 FEATURE EXTRACTION

Mel frequency Cepstral coefficients algorithm is a technique which takes voice sample as inputs. After processing, it calculates coefficients unique to a particular sample. In this project, a simulation software called MATLAB R2013a is used to perform MFCC. The simplicity of the procedure for implementation of MFCC makes it most preferred technique for voice recognition.

BlockDiagram



Generation Of Coefficients Using MFCC

MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition..

Pre-Emphasis

The speech signal x(n) is sent to a high-pass filter :
y(n) = x(n) - a * x(n - 1)
where y(n) is the output signal and the value of a is usually between 0.9 and 1.0.
The Z transform of this equation is given by
H(z) = 1 - a * z -1

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans.

Frame Blocking

The input speech signal is segmented into frames of 15~20 ms with overlap of 50% of the frame size.

Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. If this is not the case, zero padding is done to the nearest length of power of two. If the sample rate is 16 kHz and the frame size is 256 sample points, then the frame duration is 256/16000 = 0.016 sec = 16 ms. Additional, for 50% overlap meaning 128 points, then the frame rate is 16000/(256-128) = 125 frames per second. Overlapping is used to produce continuity within frames.

Hamming Window

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by

x(n), n = 0,…N-1,
hen the signal after Hamming windowing is,
x(n) * w(n)
where w(n) is the Hamming window defined by
w(n) = 0.54 - 0.46 * cos (2πn/(N-1))
where 0 ≤ n ≤ N-1

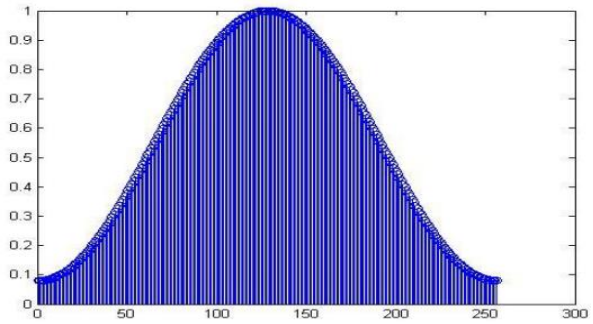


Fig. 4Plot of Hamming Window

Fast Fourier Transform

FFT is performed to obtain the magnitude frequency response of each frame. When FFT is performed on a frame, it is assumed that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, FFT can still be performed but the discontinuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we multiply each frame by a hamming window to increase its continuity at the first and last points.

Speech is usually segmented in frames of 20 to 30ms, and the window analysis is shifted by 10ms. Each frame is converted to 12 MFCCs plus a normalized energy parameter. Assuming a sample rate of 8 kHz, for each 10ms the feature extraction module delivers 39 numbers to the modeling stage. This operation with overlap among frames is equivalent to taking 80 speech samples without overlap and representing them by 39 numbers. In fact, assuming each speech sample is represented by one byte and each feature is represented by four bytes (float number), one can see that the parametric representation increases the number of bytes to represent 80 bytes of speech (to 136 bytes).

3.5 CLASSIFICATION

"Support vector machine" is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n dimensional space with the value being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.SVM is basically a two-class classifier based on the idea of "large margin" and "mapping data into a higher di-mensional space".

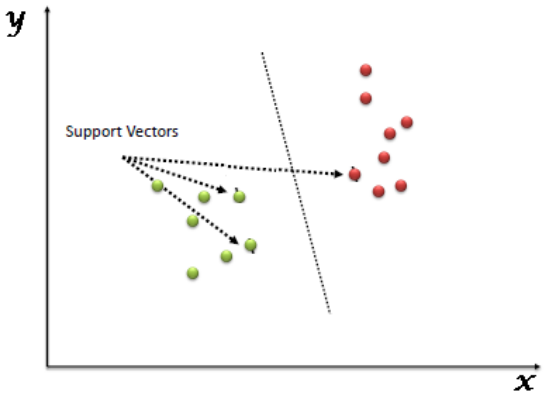


Fig. 5. Support vectors

The principle of SVM is to make minimize the structure risk, in the high dimensional feature space, find an optimal discriminant hyper plane with low VC dimension to make the distance between the twoclasses' data have large margin. When the feature space is not linear dividable, SVM maps the data into high dimensional feature space with non-linear mapping, and finds the optimal classification hyper plane in high dimensional feature space.It is a compromise between the proportion of false classified samples and algorithm complexity.

Even it is effective in high dimensional spaces and memory efficient, it does not perform well, when we have large data set because the required training time is higher. It also doesn't perform very well when the data set has more noise i.e, target classes are overlapping. SVM doesn't directly provide probability estimates, these are calculated using an expensive fivefold cross validation.

3.6 TRAINING

Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm for further iterations.Neural networksare a computational approach, which is based on a large collection of neural units, loosely modeling the way a biological brain solves problems with large clusters of biological neurons connected by axons. Each neural unit is connected with many others, and links can be enforcing or inhibitory in their effect on the activation state of connected neural units. Each individual neural unit may have a summation function which combines the values of all its inputs together. There may be a threshold function or limiting function on each connection and on the unit itself: such that the signal must surpass the limit before propagating to other neurons. These systems are self-learning and trained, rather than explicitly programmed, and excel in areas where the solution or featuredetection is difficult to express in a traditional computer program. Further error rate reduction can be obtained by using convolutional neural networks. The special structure such as local connectivity, weight sharing, and pooling in CNNs

exhibits some degree of invariance to small shifts of speech features along the frequency axis, which is important to deal with speaker and environmental variations. Experimental results show that CNNs reduce the error rate by 6%-10%.

5. RESULT

The median filtered signal (figure 6) is obtained in order to reduce the noise in the larynx speech signal. The median filter is a non-linear digital filtering technique, often used to remove noise. Such noise reduction is a typical processing step to improve the results of later processing.

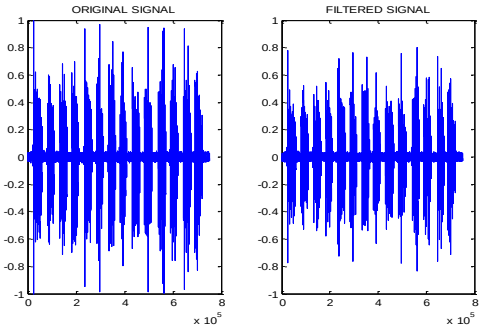


fig 6. Median filtered signal

The features such as width, magnitude, height, frame, words count, repetitions, syllables are obtained from the speech signal.(figure 7)

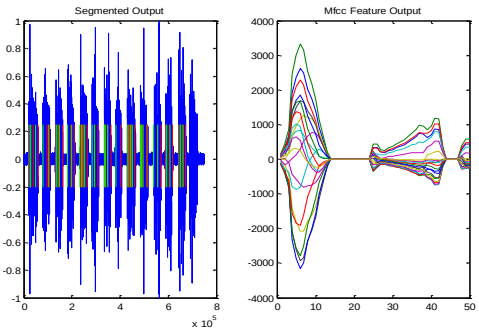


Fig. 7 MFCC feature output

Such features are given a gate pulse to obtain a high quality of speech signal using which the speech of laryngectomee patient can be obtained clearly.

6. CONCLUSION

In order to reach the goal of a new bionic EL speech system we dealt with control mechanisms of the artificial excitation signal as well as the artificial excitation source itself.We compared the healthy fundamental frequency with the estimated one. The natural fundamental frequency is the best possible reference. Nevertheless, we were able to perform to a similar level, or even better in 20% of the evaluated listening examples.This paper is based on our understanding of the importance of the electro-larynx device for social integration of laryngectomees. Concepts for moving to a human-centered computing Concepts are explained. This step is necessary for improving EL speech in terms of naturalness and intelligibility. In future research we intend to connect the proposed concepts and perform an overall evaluation.

In future work our aim is to focus on different excitation signals for female and male users in order to be able to adequately deal with the gender issue. Together with laryngectomees we will further elaborate the speech system, perform requirement analysis and evaluate the bionic EL speech system using human-centered and participatory design based research strategies.

REFERENCES

[1] F. Ahmadi, M. AraujoRibeiro, and M. Halaki, “Surface electromyography of neck strap muscles for estimating the intended pitch of a bionic voice source,” in Proc. IEEE Biomed.Circuits Syst. Conf. (BioCAS), Oct. 2014, pp. 37–40.

[2] L.Wu, C.Wan, S. Wang, and M. Wan, “Improvement of electrolaryngeal speech quality using a supraglottal voice source with compensation of vocal tract characteristics,” IEEE Trans. Biomed. Eng., vol. 60, no. 7, pp. 1965–1974, Jul. 2013.

[3] I. McLoughlin, “Super-audible voice activity detection,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 22, no. 9, pp. 1424–1433, Sep. 2014.

[4] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, “Alaryngeal speech enhancement based on one-to-many eigenvoice conversion,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 22, no. 1, pp. 172–183, Jan. 2014.

[5] Prasanta Kumar Ghosh, Student Member, IEEE, Andreas Tsiartas, Student Member, IEEE, and Shrikanth Narayanan, Fellow, IEEE,” Robust Voice Activity Detection Using Long-Term Signal Variability”, IEEE transactions on audio, speech and language processing, VOL. 19, NO. 3, March 2011

[6] Keigo Nakamura, Matthias Janke, Michael Wand, and Tanja Schultz, ”Estimation of fundamental frequencyfrom Surface Electromyographic Data: EMG-to-F_0”, Cognitive Systems Lab, Germany, 2011

[7] M.Hashiba, YasunoriSugai, Takashi Izumi, Shuichi Ino, and TohruIfukube, “Development of a wearable electro-larynx for laryngectomees and its evaluation”, Conference of the IEEE EMBS, August 2007

[8] Ehab A. Goldstein*, James T. Heaton, James B. Kobler, Garrett B. Stanley, *Associate Member, IEEE*, and Robert E. Hillman, “Design and Implementation of a Hands-Free Electrolarynx Device Controlled by Neck Strap Muscle Electromyographic Activity”, IEEE transactions on Biomedical Engineering, VOL. 51, NO. 2, February 2004

[9] Brian Madden, Mark Nolan, Edward Burke, James Condron, Eugene Coyle, “Intelligibility of Electrolarynx Speech using a Novel Actuator”, The Irish Signals and Systems Conference, June 2010.

[10] Fabian Triefenbach KrisDemuynck, and Jean-Pierre Martens, “Large Vocabulary Continuous Speech Recognition With Reservoir-Based Acoustic Models”, IEEE signal processing letters, vol. 21, no. 3, march 2014