Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 117 (2017) 901–908

Energy Procedia

www.elsevier.com/locate/procedia

1st International Conference on Power Engineering, Computing and CONtrol, PECCON-2017, 2-4 March 2017, VIT University, Chennai Campus

Design and Development of a Speech therapy Module Using Signal Processing Techniques

P. Mahalakshmi*, Ritwik Dhawan, Ujjwal Ashish, A. Sharmila

School of Electrical Engineering, VIT University, Vellore – 632 014, India.

Abstract

The people who have speech and language problems and have undergone surgery, trauma, etc. need the help of the speech language pathologists to improve their speech process. The pathologists many a times find it difficult to analyze the progress of the patients. So, in order to objectively evaluate the patient a system has been developed which can help the speech pathologists in analysis of the patient. A special hardware is developed for glottal vibration acquisition which is an MFCC based speech recognizer. This speech therapy module can take inputs either from the glottal vibrations or speech. All the signal processing has been done in MATLAB environment. Further, this module can extend and be used to train children to improve their vocabulary, to learn alphabets, numbers, etc.

Peer-review under responsibility of the scientific committee of the 1st International Conference on Power Engineering, Computing and CONtrol.

Keywords: Glottal Vibrations; MFCC; DCT; STFT; Signal acquisition; Signal processing.

1. Introduction

According to the World Health Organization statistics, at least 3.5% of the world population is affected by speech and language disorders which further result in impaired communication skills [1]. The disorders may vary from mild impairments like pronunciation problems to extreme cases like aphasia and crano-facial peculiarities. A speech disorder refers to a problem with the actual production of sounds, whereas a language disorder pertains to a trouble

* Corresponding author. Tel.: +91-416-2202435; fax: +91-416-2243092.
E-mail address: pmahalakshmi@vit.ac.in

1876-6102
Peer-review under responsibility of the scientific committee of the 1st International Conference on Power Engineering, Computing and CONtrol.
10.1016/j.egypro.2017.05.209

in comprehension or assembling words keeping in mind the end goal to share or express thoughts. This paper deals with the development of a speech therapy module which looks into the speech language pathology, involving prevention, screening, consultation, assessment, diagnosis and treatment of all speech and language disorders.
There is a wide range of disorders which can be treated via speech therapy namely, learning difficulties, language delay, hearing impairment, cleft palate, stammers, dyslexia and autism. People suffering Parkinson's disease, motor neuron disease, sclerosis, Huntington's disease, dementia, throat cancer etc. can undergo voice therapy to overcome the effects of the above said problems.
This paper incorporates a speech therapy system which detects glottal vibrations. It is a combination of two modules which are signal acquisition module for acquiring signal and signal processing module which consists of feature extraction and feature identification. The signal processing is done on MATLAB. MFCC based algorithm is used for speech recognition. Hence, recognition becomes efficient for speech as well as glottal vibration signal inputs.

2. Theoretical Background

Glottal vibration also known as a speech signal is imperative to the system as a whole. There have been instances wherein, due to external injuries, a hindrance is created in the pathway from the glottis (signal production center) to the lips. Some examples are nodules, polyps, cysts, etc., which concludes that speech and glottal vibrations both are important. The major sections in our work include the following two parts:

2.1. Signal Acquisition

A hardware module is developed for signal acquisition. The Fig.1 shows the components required for the acquisition model to be accomplished.

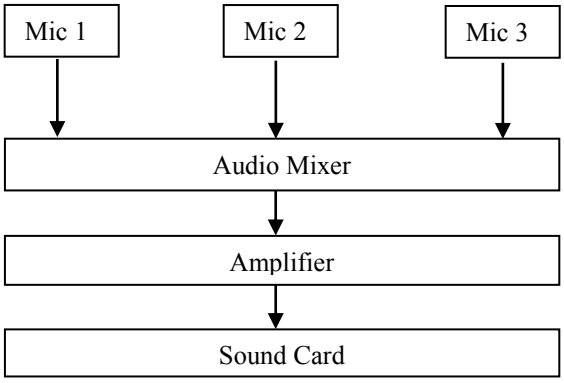

Fig. 1 Block Diagram of Signal Acquisition System.

- Electret Condenser Microphone

The electret condenser microphone is used as an input transducer. It utilizes two directing plates to catch sound waves and changes it into electrical signs. This plan is found in various types of capacitor receivers with an added advantage of less power being drawn due to the presence of a conducting plate with an attached insulator. An electret is a steady dielectric material with a for all time inserted static electric charge (which, because of the high resistance and substance dependability of the material won't rot for quite a while) [2]. An electret receiver is an omnidirectional enhancer, which infers it can get sound from all bearings. The information sound waves changes the capacitance of the two leading plates. The diaphragm is the directing plate that gets the sound waves and it causes

the adjustment in capacitance. This alteration in capacitance produces contrast in voltage on the back plate and the electrical signal is conveyed forward.
Glottal vibrations are captured by a set of three microphones which are sutured on to a neck worn band. An additive mixer is used to add signals from three microphones to give a composite signal. Active components like buffer amplifiers and impedance matching. This mixer is passive in nature as it includes only resistors, and the signal is suitably amplified before feeding into the computer through the PC sound card.

- Circuit Design

The signal acquisition module is designed. The Fig. 2 displays the basic circuit diagram of the module.

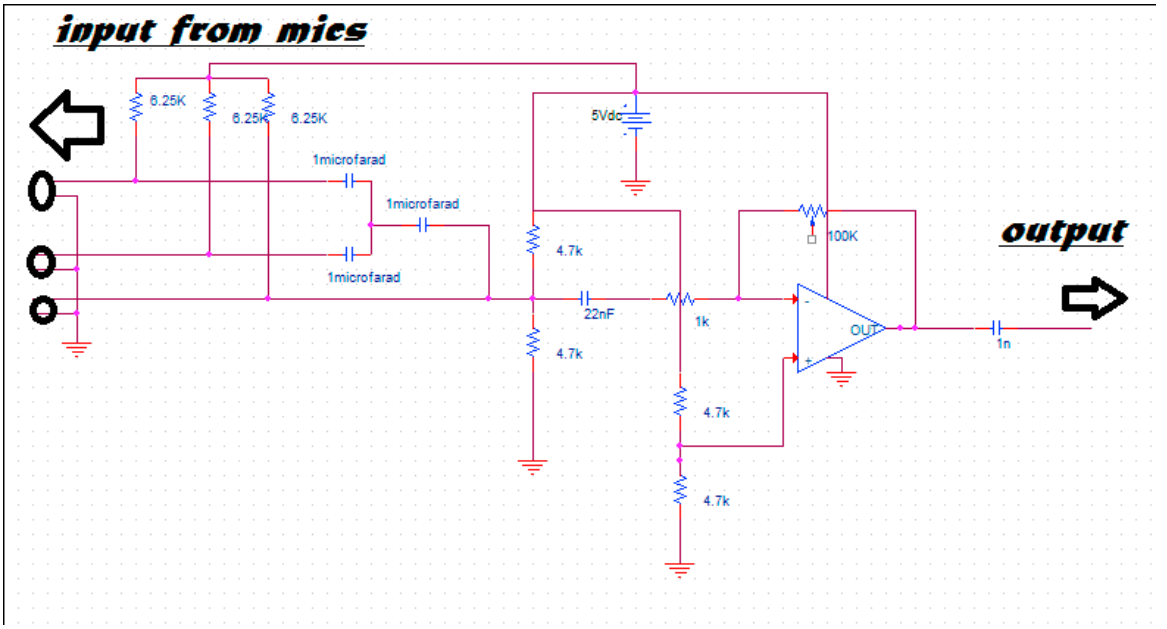

Fig. 2 Circuit diagram for acquisition system.

Circuit analysis gives us the safe current value as 0.8mA. Therefore, value of resistors attached to the microphone is 5V/0.8mA=6.25K. Capacitors are used to reduce DC components in the signal. Gain determining resistor is variable, so as to set the gain accordingly.

2.2. Signal processing

The speech signal is sampled at 8000Hz. This sampled signal is passed through various processes which generates a matrix of coefficients known as Mel Frequency Cepstral Coefficients (MFCC). After scrutinizing the spectrum of voiced segments, it can be concluded that lower the frequency, lower the energy and vice versa. This drop in vitality crosswise over frequencies is brought about because of the way of glottal vibrations. Enhancing high recurrence vitality makes data from these higher formants promptly accessible to the acoustic setup and upgrades acknowledgment precision. Pre-emphasizing is the process of increasing the magnitude of high frequencies with respect to the magnitude of low frequencies within the signal bandwidth [3]. This process is similar to first order FIR filtering. It smoothen the signal and makes it less susceptible to limited impacts later in signal processing. The transfer function is given by Equation 1:

$$H(z) = 1 - \alpha z^{-1} \quad (1)$$

Where H (z) is the transfer function and α is the pre emphasis factor and the value for this is 0.95.
Speech is naturally a non-stationary signal, thus speech analysis whether FFT based or LPC-based ought to be done in short portions for which speech signal is thought to be stationary. Feature extraction is carried out on 20 to 30ms frames with a shift of 10-15ms between two consecutive frames.
In order to minimize the effect of each segment, Hamming Window has to be used as a smoothing window and is applied to each frame. This window tapers the signal to zero both at the start and at the end of each frame. Hamming Window function is written in Equation 2:

$$W(n) = 0.53836 - 0.46164\cos(2\pi n) \quad (2)$$

Where $0<n< N-1$ [4]
Figure 3 shows the plot of the hamming window to be used in the filtering process.

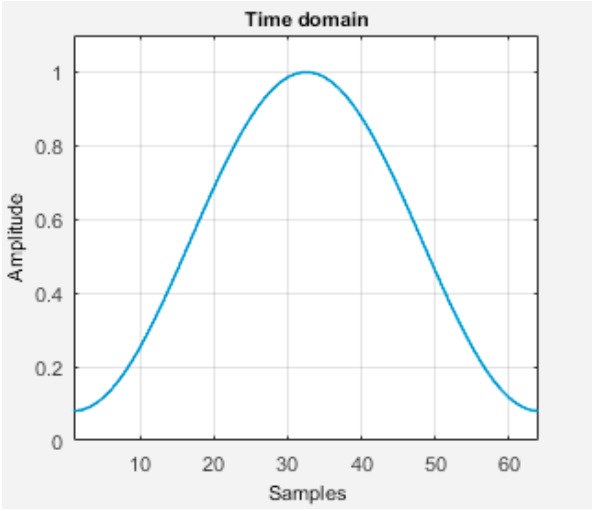

Fig.3 The plot of hamming window.

Frequency domain provides most useful parameters for speech processing. Spectral analysis of speech signals is more consistent as compared to time domain analysis. The fundamental model of speech generation with a periodic or noisy waveform that energizes a vocal tract channel compares well to separate models for the excitation and for the vocal tract. Human listening seems to give careful consideration to spectral aspects of speech as opposed to phase or timing aspects. Thus, spectral analysis is used to extract most parameters from speech [5].

3. Methodology

Figure 4 shows the block diagram of different components involved in signal processing for speech therapy module. The objective of signal processing is the extraction of important parameters namely speech signal intensity.

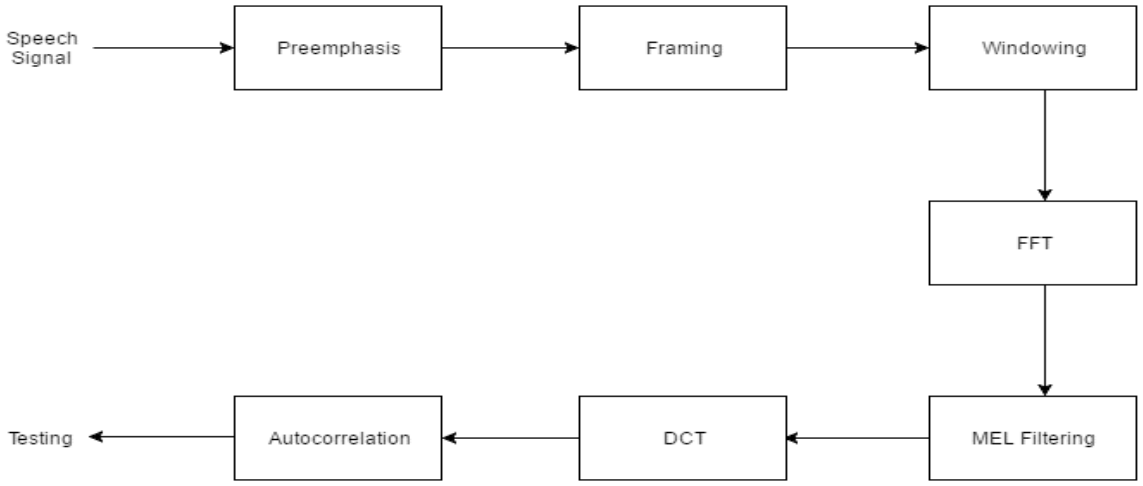

Fig.4 Block Diagram of the Signal Processing operations.

3.1. Short Time Fourier Transform

Fourier analysis represents the speech in terms of amplitude and phase as a function of frequency. Treating the vocal tract as a linear system, the Fourier transform of speech is the product of transforms of glottal excitations and of vocal tract response. Windowing has to be used for speech as it is a non-stationary, dynamic signal [3]. Short time Fourier transform of a signal x (n) is given as:

$$X_n(e^{jw}) = \sum_{m=-x}^{-\infty}(x(m).e^{-j\omega m}.\omega(n-m)) \quad (3)$$

For our purposes, DFT is used instead of Fast Fourier Transforms. This has been done to accommodate frequency variable ω, such that it only takes N discrete values (here, N is he window of duration of DFT).

$$X_n(k) = \sum_{m=0}^{N-1}(x(m)e\frac{-j2\pi km}{N}.\omega(n-m)) \quad (4)$$

Fast Fourier Transform (FFT) is used to implement DFT. Poor resolution is given by small values of N because the window low pass filter is wide and they yield good time resolution since the speech properties are averaged only over short time intervals. Although, large N values gives good frequency resolution but poor time resolution.

3.2. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral represents the short term power spectrum of a sound which is based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency [6]. The human ear perceives and resolves frequencies in a nonlinear fashion across the entire audio spectrum. For higher frequencies, the human ear is less in comparison to the lower frequencies with the same sound intensity. MFCC's generation algorithm uses perceptual scale to reconstruct and model human perception.

3.3. Algorithm for deriving the MFCC

Mel Frequency Cepstral Coefficient's can be derived as-

- Fourier transform of the signal is taken.
- The Mapping of the powers of the spectrum is obtained above onto the Mel scale

- The log of the power of each of the Mel frequencies is taken.
- The DCT of Mel log powers is taken.
- The amplitudes of the resultant spectrum are terms as MFCCs [6].

Mel scale transformations are used in MFCC's algorithms and the following equations are used in transforming from Hertz scale to Mel scale and vice versa.

$$f_{m}el = 1127.01048 \ln (1+ fhz/700) \quad (5)$$

$$f_hz = 700(e^{\frac{fmel}{1127.01048}} - 1) \quad (6)$$

It can be assumed that Mel values beneath 1127 Hz rise linearly while above 1127Hz the Mel values rise logarithmically. Mel filter bank needs to be generated in order to split the spectrum of the signal into channels that are based on perceptual scales. The filter bank can be defined as a collection of L band pass filters, with center frequencies linearly spaced in perceptual scale. The most widely recognized shape of filters is triangular, yet in different varieties of filter usage shapes like trapezoidal, rectangular or Gaussian can be seen. Every filter in the perceptual scale has a similar width and is overlapped for half of its width with the following filter. Since the human perception is bound by a certain range, the filter bank characterizes the boundary frequencies i.e fmin and fmax. The concept of boundaries and number of filters enables algorithm to compute constant filter in perceptual scale.

$$f_w = (f_{max} - f_{min})/L \quad (7)$$

The creation of the filter bank is done prior to the processing and is done in accordance with the MFCC's algorithm. Filter bank parameters being constant are reused in each frame channeling phase. Figure 5 shows the filter banks generated for the filters [7].

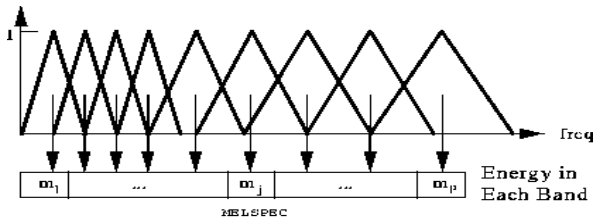

Fig.5 Mel filterbands generation.

In our case, Mel filter bank has filters of varying length. The filters are responsive to a bandwidth range of 0 to 4000Hz. The Mel filtering process is given by equation 8:

$$S(k) = \sum_{i=0}^{N/2} (Xm(i)^1 . \mathrm{Mk(i)}) \quad (8)$$

Where S is the filtered signal, Xm is the Fourier transform and Mk is the filter coefficient. As stated in the definition of MFCC, logarithm of the above signal is taken before proceeding to discrete cosine transform.

3.4. DCT

Mel frequency cepstral coefficients are obtained as a result of discrete cosine transform of the Mel filter output. MFCC can be obtained by the equation 9:

$$C_m(n) = \sum_{i=0}^{L-1} SLm(i).\cos(\pi n(2i+1)/2L) \quad (9)$$

Where F is the number of frames and L represents the number of filters then F*L gives us the number of Mel frequency cepstral coefficients. The discriminating power of the original glottal signal is preserved whilst reducing the size of the feature vector. For example, for a sample size of 12000, number of frames =94, so number of MFCC's = 94x25 = 2350. The next step of the process is to calculate the autocorrelation coefficients of these MFCC's. The coefficients are then compared for the trained and test word.

3.5. Short time Autocorrelation function

The cross correlation function is given by:

$$R_{x,y}(m) = Ex_{n+m}y^*_n = Ex_n y^*_{n-m} \quad (10)$$

Here, xn and yn are the jointly existing stationary processes and E (.) is an operator. If x and y are length N vectors (N>1), the length of the cross correlation sequence is 2N-1. If x is an N X P matrix, the autocorrelation matrix c will have 2N-1 rows whose p^2 columns would be having the cross correlation sequences for all the possible combinations of the columns of x.
Autocorrelation coefficients are symmetrical therefore only one part has to be considered i.e., 1175. This represents the size of the feature vector corresponding to a word size of 12000. Furthermore, we can use truncation to reduce the number of autocorrelation coefficients to 600.

3.6. Recognition

The system has 3 phases namely, training phase, practice phase and testing phase. The training phase is used to train as many number of words as we wish for, hence a database for trained words is created. During the check the autocorrelation coefficients of the given data is compared to the autocorrelation coefficients of the existing data set. The word corresponding to the minimum difference (weight) is identified. A threshold value for the weight is set by trial and error method so as to avoid false recognition. In case the minimum difference is less than the set threshold the corresponding word is recognized.

4. Results

Vocal fold diseases were diagnosed by visual inspection and speech impairment was almost impossible to diagnose via this method. The vocal fold's vibratory movement could only be detected after the development of techniques like electroglottography (ECG), photoglottography (PGG), etc [8]. The analysis of the vocal vibrations along with its spectrum via the speech therapy module may help us to derive a set of quotients which can be used to quantify various vocal pathologies.

5. Conclusion

This speech therapy module is mainly developed for the use of the speech impaired. Integration of the two modules: signal acquisition and signal processing lead to the design of the desired system. The MFCC based algorithm is responsible for speech recognition. For future reference designing of a GUI module and combining it with the above therapy module can be used for voice rehabilitation and can also be used by kids to improve their vocabulary and pronunciation.

References

[1] Oytun Turk and Levant M. Arslan., " Software Tools for Speechtherapy and Voice Quality Monitoring ", EUSIPCO-2005.

[2] Sessler G.M., West J.E., "Self-biased condenser microphone with high capacitance", Journal of the Acoustical Society of America, pp 1787-1788, vol 34: 1962.

[3] Douglas O'Shaughnessy., Speech Communication: Human End Machine, Addison-Wesley, ISBN No: 9780201165203.

[4] Oppenheim, Alan V., Ronald W. Schafer, and John R. Buck. *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice Hall, 1999, p. 468.

[5] Lawrence Rabiner., Bllng Hwang Juang., Fundamentals of speech recognition, Pearson Education, ISBN No: 81-297-0138-3.

[6] Jan G. Svec., Peter S. Popolo., Karren Rogge Miller., and Ingo R Titze., "The calibration and set up of the NCVS Dosimeter", The national centre for voice and speech online technical memo., no 2, ver 2.4, Apr 2004.

[7] http://www.ee.columbia.edu/ln/LabROSA/doc/HTKBook21/node54.html

[8] Jack J. Jiang., Shuangyi Tang., Michael Dalal., Chi-haur Wu., and David G. Hanson., "Integrated analyzer and classifier of glottographic signals", IEEE Transactions on Biomedical Engineering, no 2, vol 6, June 1998, pp 227-234.