International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Investigations of Formant Extraction of Male and Female Speech Signal Via Cepstral De-convolution Method

Sakshi Bedi[1], Randhir Singh[2], Saleem Khan[3]

[1, 2] *Dept. of Electronics and Communication Engineering, SSCET, Badhani, Punjab, India*
[3]*Dept. of Electronics, GGM Science College, J&K, India*

Abstract - *The capability of speaking the language is one of the most amazing skills human possess. It serves as a very effective way of communication, sharing experiences, thoughts, ideas among the people. In this research work, cepstral analysis of sentence level speech signals of male and female speakers is carried out. Spectrogram analysis of segmented speech is carried to determine various speech parameters. The Cepstral analysis shows the unique trait between male and female speakers in terms of low time liftered speech processing.*

Key Words: **The speech signal, sentence analysis, Hamming, Cepstral, Spectrogram.**

1.INTRODUCTION

Speech is the vocalized form of Human communication. The creation of each spoken word is in accordance with the phonetic combination of a limited set of speech sound units such as vowel and consonants. The process of generation of sound is based on three main factors for effective voice generation:

1.1 Breathing

Our purpose of producing voice is signaled to the parts of the body by taking the involvement of impulses from the brain [2]. The body shows its first response by exhalation so that to power voice there is enough air in the lungs. The exhalation from the mouth and nose passes down the wind pipe and is inhaled into the lungs. The air which we breathe in through the lungs the ribcage needs to be expanded and the diaphragm which forms the base of the chest needs to be smooth downwards. At the time we breathe in we feel most of the expansion in the lower ribs area. Once the air we inhaled down into the lungs reaches its capacity, the elastic tissue of the lungs recoils and the air we inhaled is breathed out [3].

1.2 Phonation

The production of electric supply for voice is done by breathing air out of the lungs. This process of exhalation of airflow from the lungs makes the vocal chords in the voice box vibrate to make the basic sound of the voice. This is referred as Phonation. The basic tone of sound can be differentiated in many ways, depending on the way in which we use the other parts of the voice mechanism and the vocal folds.

The basic facets of voice that can be varied are:

Pitch: Pitch refers to as the highness and lowness of the voiced sounds. It is determined by the length of the folds, the thickness of the edge of the folds and the speed of vibration of the vocal folds. The highness level of the pitch depends on the elongation and thinning the edges of the vocal folds. The less the elongation and thinning of the edges of the vocal folds the vocal folds vibrate at the slower rate. Pitch describes the main indication of gender. The average pitch of the female voice is about 200bHz where as the average pitch of the male voice is about 110 Hz. Hence pitch is the main indicator of gender. Emotion signals such as excitement, stresses that are present in voice also describes as a part carried by pitch variations. Shouting as a means of stressing on a particular point or an expression of anger describes more with rising of the pitch than loudness. More often pitch variation is often correlated with loudness variation. Happiness, distress and extreme fear in voice are signals led by the fluctuation of the pitch in voice. The accent is the part conveyed by changes in the pitch and rhythm of the speech signal. For example in more accents, such as Northern Ireland accent, at end of the sentence, the pitch signal is raised instead of being lowered [4-5].

Loudness: Loudness referred to as the how loud or soft a voice. The amount of air pressure from the lungs and muscle tension of the vocal folds decides the loudness of the voice. In the nutshell, the loudness of the sound depends on the air pressure and tension of the vocal folds. The greater the air pressure, the tenser the vocal folds so the louder the voice will be and lesser the air pressure, the slacker the vocal folds so softer the voice will be.

Quality: Quality referred to as the clarity of the voice sounds. The determination of voice quality depends on many complex factors including the relaxation of muscles of the larynx and how moist the cover of vocal folds is and how smoothly the vocal folds vibrate. The dryness of the cover of vocal folds depends on the muscles of the larynx. If the muscles of the larynx are excessively tense, the cover is dry and the folds cannot close together or move in an irregular

way. The voice quality will sound rough, strained and breathy.

Resonance: This type of sound production made by the vocal folds is too weak to be heard, and then the modification of sound takes place so that we are able to recognize it as the human voice as it moves up from larynx through throat, mouth, and nose. This transformation of sound is known as resonance.

Different types of people speak different languages according to the area in which they are born. To speak in their mother tongue they do not need a different type of training or knowledge. By understanding both Audio and visual gestures children learn to speak in their respective mother tongue at an early age of one year. The designing of these phonemes is according to the articulator movement of the vocal tract. In the English phonemes, the sounds are powered by lung air being pushed out. There are two ways for the production of sound:

1.1 By vibrating the vocal chords: Inside the throat two muscular folds of skin can be made to vibrate. The frequency of vibration can be changed within the limits.

1.2 By altering the position of components: By altering the position of components of throat and mouth in between the existing air and vocal chords. By changing the size of the cavity these alterations may modify the note produced by the vocal chords. On the other hand, they may themselves produce the noise by causing air friction.

Phonemes can be further classified into these parts:

Vowels: When the air inside the lungs passes over the vibrating chords and then freely moves out of the mouth. These types of sounds called as Vowels. Thus vowels are continued until you run out of breath. The changes in the size and shape of the resonant cavity to produce different sounds are due to the position of lips and tongues.

Some examples of the English language:

/i/ is a high, front, unrounded vowel, as in beet /bit/ or neat /nit/.
/a/ is a low, back, unrounded vowel, as in bar or bath.
/u/ is a high, back, rounded vowel, as in a spoon.

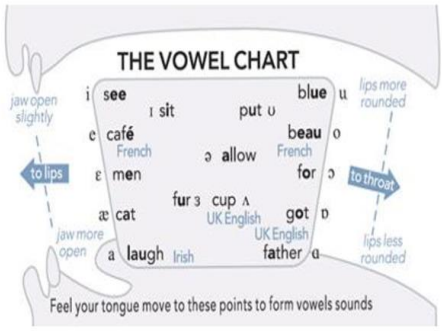

Fig-1: Formation of vowel sounds with the help of tongue.

Consonants: The second large phoneme grouping is that of consonants. The consonants consist of a number of subgroups:

Nasals: As with vowels, the source is quasi-periodic airflow puffs from the vibrating vocal folds. The velum is lowered and the air flows from the Nasal cavity, the oral tract becomes narrower; thus the voice is radiated at the nostrils. The nasal consonants are differentiated by the place along the oral tract at which tongue creates an obstruction. The two nasals that we compared are /m/ as in "**m**o". For /m/, the oral tract obstruction occurs at the lips and /n/ as "**n**o" in which the obstruction occurs within the tongue to gum ridge.

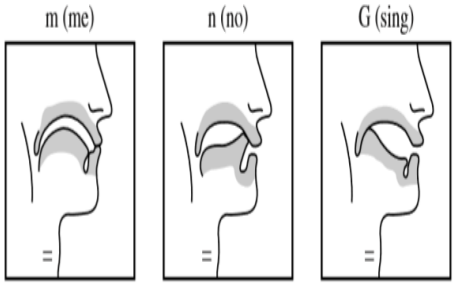

Fig- 2: Vocal tract configurations for Nasal consonants.

Fricatives: Fricatives consonants divide into two classes: Voiced Fricative and Unvoiced Fricative. In the case of voiced fricatives, there is no occurrence of simultaneous voicing with noise generation. The occurrence of voicing carried out at the stage of early frication and not at all during the time of frication. In the case of voiced fricatives, the voicing occurs sooner into the transition. For unvoiced fricatives, the vocal folds are relaxed and not vibrating. But turbulent airflow at the time of constriction that is narrower than with vowels. The example of Voiced fricative /z/ as in "**Z**ebra" and there is vibration in the vocal folds during the generation of noise. There is a comparison in the unvoiced fricative /f/ as in "**f**or" in /f/ there is no vibration in the vocal folds and the occurrence of constriction by the teeth against lips and the matching voiced fricative as in "**v**ote".

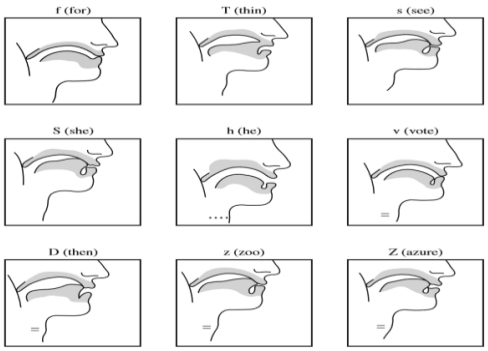

Fig-3: The vocal tract configurations for the pair of voiced and unvoiced fricatives.

Plosives: Just like other sounds which can be described by steady state spectra plosives are transient phones. These can also be voiced and unvoiced. As for as unvoiced plosives are concerned, there is a generation of "Burst" at the time of the release of build up pressure behind the total obstruction of the oral tract. With the voiced plosives, behind the oral tract obstruction there is build up of pressure, but the vocal folds can also vibrate. During the occurrence of vibration, the oral tract is closed. Therefore, we hear the vibration of low frequency due to its passage through the area of walls inside the throat. This activity described as "Voice Bar".

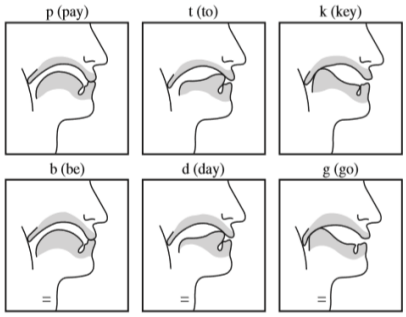

Fig-4: The configurations of vocal tract for voiced and unvoiced plosive pairs.

Affricates: This sound is produced by the combination of consonant, plosive, and fricatives and is quickly pass across from plosives to fricatives. The articulation of affricates is same as that of fricatives. The difference between them is that the fricative is superseded by the complete obstruction of the oral cavity. Examples of affricates are /tS/ as in word "**ch**ew". In this /t/ is plosive followed by fricative /s/ [6-7].

2. THE POWER CEPSTRUM

Bogert et al was the first person to describe and used the Power cepstral in 1963. To determine the echo arrival times in the composite signal is the main cause of study. Since the delayed echoes in the logarithmic spectrum of input data sequence $x(n)$ appear as ripples. Therefore, the power cepstral is the productive tool in spite of this that the frequencies of excitation function and the basic wavelet does not overlap with each other. In the nutshell, the power cepstrum of the signal is defined as the square of inverse z-transform of the logarithm of the magnitude squared of the z-transform of the data sequence.

$$x_{pc}(nT)=\left(z^{-1}\left\{\log|x(z)|^2\right\}\right)$$

$$x_{pc}(nT)=\left[\frac{1}{2\pi}\phi_c\log|x(z)|\right]^2$$

Where X (z) represents the Z-transform of the data sequence $x(nT)$. Let us suppose that data sequence consists of two convolved sequences $y(nT)$ and $v(nT)$, which represents excitation function and basic wavelet respectively.
The data sequence can be written as:

$$x(nT)=y(nT)*v(nT)$$

Then, the equation can be written as by multiplying the Fourier transform of both sequences:

$$|x(z)|^2 = |y(z)|^2 \cdot |v(z)|^2$$

Then, by taking the logarithm of both sides of operation:

$$\log|x(z)|^2 = \log|y(z)|^2 \cdot \log|v(z)|^2$$

By elaborate the power spectrum analysis; let us assume that the excitation function (signal) is given by:

$$v(nT) = \delta(nT) + c\delta(nT-n_0T)$$

Where $\delta(n)$ denotes the unit impulse function in sampled data sequence.
On the basis of this equation, we can further write as:

$$x(z)^2 = |y(z)|^2\left|1+cz^{-n_0}\right|^2$$

By taking the logarithm on both sides of the equation and by substituting $z = e^{jw}$, we expand equation as:

$$\log\left|x\left(e^{jw}\right)\right|^2 = \log\left|y\left(e^{jw}\right)\right|^2 + \log\left(1+c^2+2c\cos\left(wn_0T\right)\right)$$

$$= \log\left|y\left(e^{jw}\right)\right|^2 + \log\left(1+c^2\right) + \log\left(1+\frac{2c}{1+c^2}\cos\left(wn_0T\right)\right)$$

By taking the inverse z-transform of given equation, we now described the data sequence $x(nT)$ can in terms of its components:

$$x_{pc} = y_{pc} + v_{pc}$$

Where y_{pc} is the power cepstrum of basic wavelet and v_{pc} is the power cepstrum of the excitation signal [8].

3. THE COMPLEX CEPSTRUM

Oppenheim developed the complex cepstrum which is an outgrowth of the homographic system theory. In spite of the fact that the power cepstrum can be used for detecting echoes since the phase information is lost it cannot be used for detecting wavelet recovery [10-11]. In the nutshell, the complex cepstrum of data sequence can be defined as the inverse z-transform of the complex logarithm of z-transform of data sequence as follows:

$$\hat{x}(z) = \log x(z) = \log y(z) + \log v(z)$$

$$\hat{x}(nT) = \frac{1}{2\pi j}\int_c \log(x(z))z^{n-1}dz$$

Where $\hat{x}(nT)$ represents the complex cepstrum and x(z) represents the z-transform of data sequence $x(nT)$

$$x(nT) = y(nT) * v(nT)$$

Where $y(nT)$ represents the basic wavelet $v(nT)$ represents the excitation function.

This can be written in z domain as:

$$x(z) = y(z)v(z)$$

The logarithm of the above equation can be written as:

$$\hat{x}(z) = \log x(z) = \log y(z) + \log v(z)$$

By taking the inverse z-transform of the equation the complex cepstrum can be estimated below:

$$\hat{x}(nT) = \hat{f}(nT) + \hat{g}(nT)$$

Where $\hat{x}(nT)$ represents the complex cepstrum of the composite signal $\hat{x}(n)$, $\hat{y}(nT)$ represents the complex cepstrum of the wavelet component, and $\hat{v}(nT)$ represents the complex cepstrum of excitation component.

To make an effort for the involvement of the excitation function in complex cepstral, we suppose that the excitation function $v(nT)$ is of the form:

$$v(nT) = \delta(nT) + c\delta(nT - n_0T)$$

By taking the z-transform and substituting $z = e^{jw}$ we have,

$$v(z) = v(e^{jwt}) = 1 + ce^{-jwn_0T}$$

And

$$X(e^{jwT}) = Y(e^{jwT})\left(1+ce^{-jwn_0T}\right)$$

Taking logarithm on both sides of the equation v(z) as below:

$$\log x(e^{jwT}) = \log y(e^{jwT}) + \log\left(1+ce^{-jwn_0T}\right)$$

Where c<1, the data sequence reveal phase characteristics to a minimum and the wavelet component dominates.

$$\log X(e^{jwT}) = \log Y(e^{jwT}) + Ce^{-jwn_0T} - c^2e^{-2jwn_0T}$$

Finally, the complex cepstrum of data sequence is obtained by taking the inverse z-transform:

$$\hat{x}(nT) = \hat{y}(nT) + c\delta(nT - n_0T) - \frac{c^2}{2}\delta(nT - 2n_0T)$$

4. EXPERIMENT AND RESULT

Speech material was collected from four speakers (2 male and 2 female, aged 23-27 years). The male speakers were represented as M1 and M2. On the other hand, female speakers were referred as F1 and F2. The speakers we use in our experiment were university students of different age groups and had English as their first language. In the first part of speaker selection, speech recording and segmentation was carried out. In this experiment, the speech of two male and female speakers was recorded in the form of sentences. In this experiment, the segmentation of speech is carried out since each speaker takes different time while speaking sentences. This process is also carried out with the use of MATLAB program which uses Hamming Window. For signals which are changing with time, we must take a sample or window of the signal over some definite interval. If we simply cut out some portion of the signal which introduces distortions in the signal, we reduce these distortions by multiplying our portion by smoothing function which reduces the size of the signal at the edges. We need the shape of that signal in such a way which has a spectrum with narrow central lobe and small side lobes. The window based on raised cosine shape is called Hamming Window.

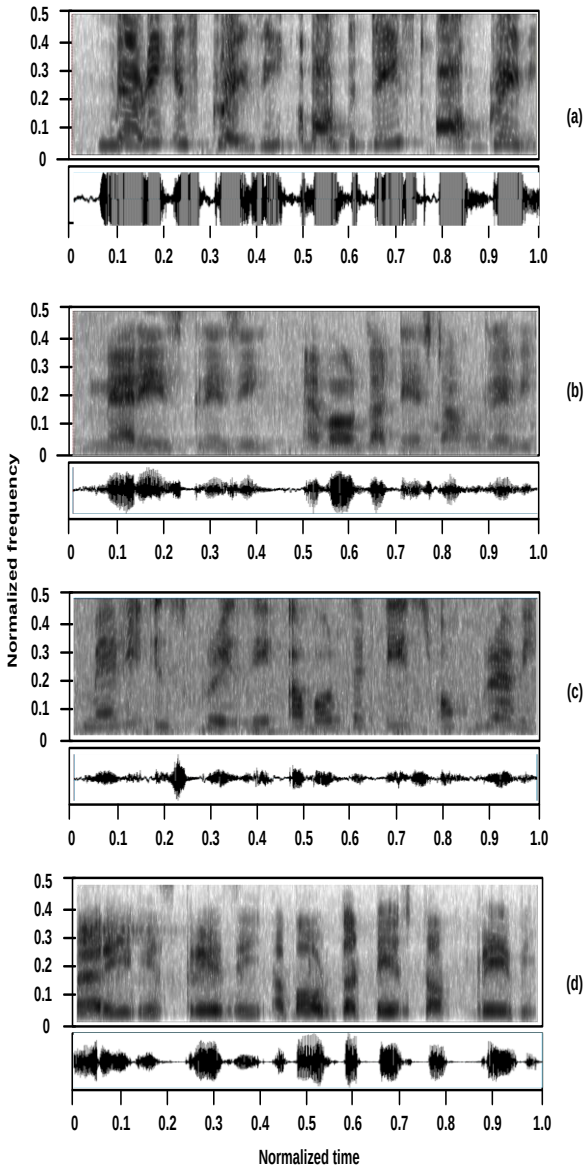

Fig-5: Normalize signal and spectrograms of "Sentence1" (a)-(b) Male speaker And (c)-(d) Female speaker.

In this experiment, the speeches of two male and female speakers were recorded in the form of sentences. This process was also carried out with the use of MATLAB program which uses Hamming Window. The output of various stages was carried out during the computation of cepstrum. Fig. 6 which is extracted from the low time liftered cepstrum the maxima at 3.9 lies at a frequency of 1700 Hz and minima at 2.4 lies at a frequency of 3700 HZ. The graph of Fig. 7 which is extracted from the low time liftered cepstrum the maxima at 1.7 lies at a frequency of 1600 Hz and minima at 0.2 lies at a frequency of 3500 Hz. The graph of Fig. 8 which is extracted from low time liftered cepstrum the maxima at 1.7 lies at a frequency of 1500 Hz and minima at 0.2 lies at a frequency of 3500 Hz. The graph of Fig. 9 which is extracted from low time liftered cepstrum the maxima at 1.5 lies at a frequency of 1500 Hz and minima at 2 lies in the frequency range of 2500 Hz.

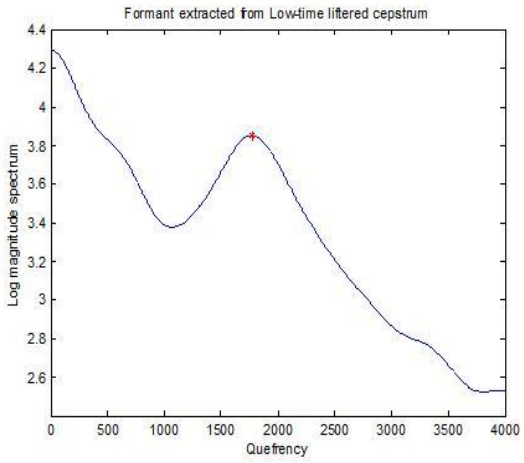

Fig-6: The low time liftered cepstrum of sp1 for "Sentence1".

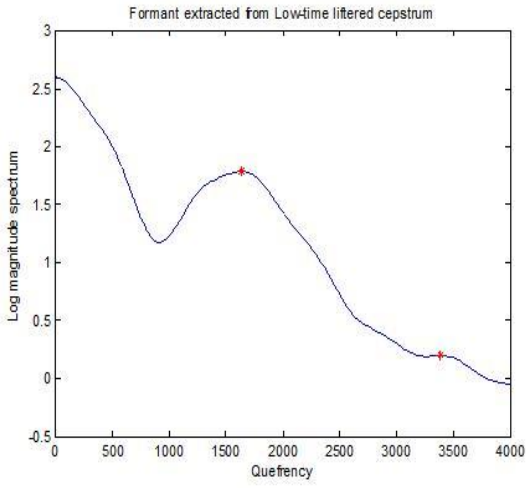

Fig-7: The low time liftered cepstrum of sp2 for "Sentence1".

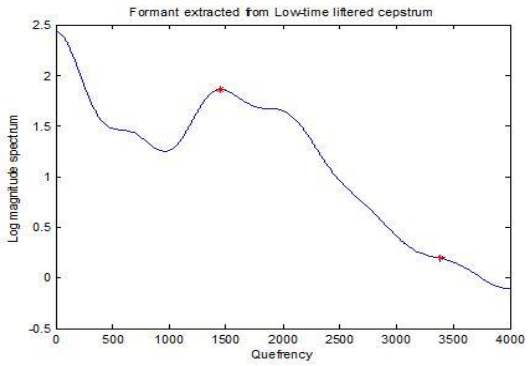

Fig-8: The low time liftered cepstrum of sp3 for "Sentence1".

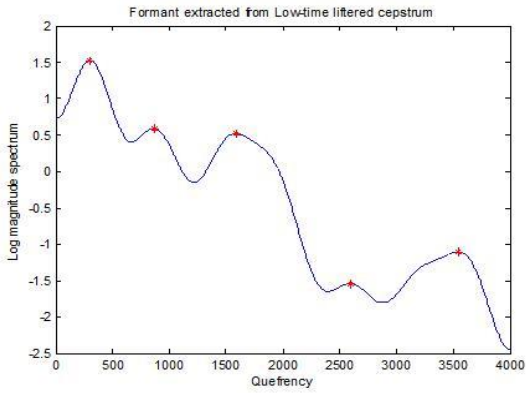

Fig-9: The low time liftered cepstrum of sp4 for "Sentence1".

Various parameters such as Pitch, Intensity and Formant frequency of male and female are given in Table1 and Table2.

Table-1 Value of speech parameters of "Sentence1" for the female speaker.

No of Parameters	Female	
	Sp1	**Sp2**
Pitch	305.38 Hz	266.64 Hz
Intensity	89.19 dB	72.96 dB
Formant Freq.	645.77 Hz	845.9 Hz

Table-2 Value of speech parameters of "Sentence1" for the male speaker.

No of Parameters	Male	
	Sp1	**Sp2**
Pitch	121.87 Hz	126.57 Hz
Intensity	74.35 dB	60.46 dB
Formant Freq.	721.78 Hz	503.10 Hz

5. CONCLUSION

In this research work, cepstral analysis of sentence for the male and female speaker is carried out. For speech analysis, it is observed female speaker has a high intensity as compared to male speaker whereas pitch and formant frequency is high for the male speaker. It is also observed from the results that female speaker have more maxima and minima in a cepstral curve as compare to the male speaker in sentence level analysis.

6. REFERENCES

[1] K. Honda, "Physiological Processes of Speech Production," Springer Handbook of speech processing.
[2] C. Darwin, "The Expression of Emotions in Man and Animals," 1872.
[3] "Species-Specific Formation of Human Vocal Apparatus," Language in the Brain: Critical Assessments.
[4] M. Akay, "PREFACE" Biomedical Signal Processing. Pp. xiii-xiv, 1994.
[5] W.H Goodenough, "Language Origin Philip Lieberman, Uniquely Human: The Evolution of Speech, Thoughts and Selfless behavior." Cambridge MA: Harvard University Press, 1991. Pp. 210.
[6] A.C Cohn, J. Clark, and C. Yallop, "An Introduction to Phonetics and Phonology Language," vol. 68, No. 1, p. 156, March 1992.
[7] Rossing. T The Science of Sound (Addison-Wesley, 1990).
[8] R. Carlson, G. Fant and B. Granstrom, "Two-Formant Models, Pitch and Vowel Perception," Auditory Analysis and Perception of Speech. Pp.55-82, 1975.
[9] H.M Teager and S.M Teager, "Evidence of Nonlinear Sound Production Mechanisms in Vocal-Tract," Speech Production and Speech Modeling, Pp. 241-261, 1990.
[10] B.H. Story and I.R. Titze, "Voice Simulation with the body-cover model of the vocal folds during pulse register phonation," The Journal of Acoustical Society of America, vol. 75, No. 4, pp. 1293-1297, Apr. 1984.
[11] Barne, C.H Shadle, and P.O.A.L Davies, "Fluid-Flow In the Dynamical Mechanical Model of Vocal Folds and Tract I. Measurements and Theory," Journal of Acoustical Society of eAmerica, vol. 105, No. 1, Pp. 343-356, Jul. 2000.