IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 2, Issue 5 (May. – Jun. 2013), PP 01-08
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Speaker Verification

Chandana Krishna[1], Dr. Hariprasad S.A.[2]
[1]*3rd Year B.E., Electronics and Communication Department, R.V. College of Engineering, Bangalore*
[2]*Professor, Electronics and Communication Department, R.V. College of Engineering, Bangalore*

Abstract: *Speaker verification is the method of automatically identifying who is speaking on the basis of individual information integrated in speech waves. An important application of speaker verification is for forensic purposes. Speaker verification has seen an appealing research field for the last decades which still yields a number of unsolved problems. Many algorithms have been developed to accomplish, some of which include Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Network. All the before mentioned algorithms serve the feature matching mechanism while the MFCC (Mel Frequency Cepstral Coefficients) are the features extracted of a voice signal. The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human.The simplest of the algorithms is calculating the distortion distance between the various codebooks of the speakers, but its efficiency is less compared to other algorithms. Here, we have tried to increase the efficiency of this method. The two phases of this system is the training phase and the testing phase. The training phase involves the feature extraction using MFCC and storing the codebooks in the database. The testing phase involves all these plus the distortion distance calculation using the codebook of the unknown speaker against all the speakers whose codebook is already stored in the database and is verified if the speaker matches with the claimed identity.*

Keywords: *Distortion Distance, Feature Extraction, Feature Matching, MFCC, Mel scale, Speaker Recognition, Training Phase, Testing Phase, Vector Quantisation*

I. Introduction

Speaker verification technology is the most potential technology to create new services that will make our everyday lives more secured. The MFCC technique is used for extracting the features of the voice signal, amidst which some techniques like spectral subtraction, scaling and varying the number of mel filters are incorporated here which has helped to improve the efficiency. The Euclidean distance is used for matching the speaker with one of the speakers whose features are stored in the database. The speaker must first type in the name of the personwhom he/she is claimingto be and the system verifies the match against the claimed identity. In general, the speaker verification systems can be either text dependent or text independent. Both text dependent and text independent methods share a problem however. By playing back the recorded voice of registered speakers this system can be easily deceived. So, the efficiency cannot be expected to be 100%.

The system is programmed for two phases: the training phase and the testing phase. The training phase where the features of the speakers are stored in the database. This is for later retrieval to verify the speaker. The testing phase is where the unknown speaker's features are compared to the features of the claimed identity. The training phase is only once for all the speakers where the testing phase is where the verification takes place. Thus, speaker verification, to determine whether a person (he or she) claims to be according to his/her voice sample. Here in this project, the efficiency is increased by some techniques by which this simple method can be used conveniently in many applications.

Paper Organization:

Section 2 explains briefly the mechanism behind the production of speech in a human being while the section 2.1 deals with the technical characteristics of the speech signal which helps in understanding the feature extraction detailed in the section 3. Sections 3.1 through 3.8 explain the various steps involved in the feature extraction. Section 4 talks about the distortion distance calculation and its inference with respect to the speaker which is the final step of the speaker verification system. Section 5 gives the simulation results of MATLAB.

II. Production of Speech:

While you are producing speech sounds, the air flow from your lungs first passes the glottis and then your throat and mouth. Depending on which speech sound you articulate, the speech signal can be excited in three possible ways:

• Voiced excitation: The glottis is closed. The air pressure forces the glottis to open and close periodically thus generating a periodic pulse train (triangle–shaped). This "fundamental frequency" usually lies in the range from 80Hz to 350Hz.
• Unvoiced excitation: The glottis is open and the air passes a narrow passage in the throat or mouth. This results in a turbulence which generates a noise signal. The spectral shape of the noise is determined by the location of the narrowness.
•Transient excitation: A closure in the throat or mouth will raise the air pressure. By suddenly opening the closure the air pressure drops down immediately. ("plosive burst")

With some speech sounds these three kinds of excitation occur in combination. The spectral shape of the speech signal is determined by the shape of the vocal tract (the pipe formed by your throat, tongue, teeth and lips). By changing the shape of the pipe (and in addition opening and closing the air flow through your nose) you change the spectral shape of the speech signal, thus articulating different speech sounds.

2.1 Technical Characteristics of the Speech Signal

An engineer looking at (or listening to) a speech signal might characterize it as follows:
• The bandwidth of the signal is 4 kHz
• The signal is periodic with a fundamental frequency between 80 Hz and 350 Hz
• There are peaks in the spectral distribution of energy at $(2n - 1) * 500$ Hz ; n = 1, 2, 3, . . .
• The envelope of the power spectrum of the signal shows a decrease with increasing frequency (-6dB per octave)

This is a very rough and technical description of the speech signal.After passing the glottis, the vocal tract gives a characteristic spectral shape to the speech signal. If one simplifies the vocal tract to a straight pipe (the length is about 17cm), one can see that the pipe shows resonance at certain frequencies. These frequencies are called formant frequencies. Depending on the shape of the vocal tract (the diameter of the pipe changes along the pipe), the frequency of the formants (especially of the 1st and 2nd formant) change and therefore characterize the vowel being articulated

III. Feature Extraction:

The most important thing is to extract the feature from the speech signal. The features extracted discriminate among the many speakers. The speech feature extraction in a categorization problem is about reducing the dimensionality of the input vector while maintaining the discriminating power of the signal.
The extracted features should meet some criteria while dealing with the speech signal, such as:

- Easy to measure extracted speech features
- It should not be susceptible to mimicry
- It should show little fluctuation from one speaking environment to another
- It should be stable over time
- It should occur frequently and naturally in speech

In this project, we are using Mel Frequency Cepstral Coefficients (MFCC) technique to extract features from the speech signal and compare with the features already stored for the speaker in the database. The figure below shows the complete pipeline of Mel Frequency Cepstral Coefficients.

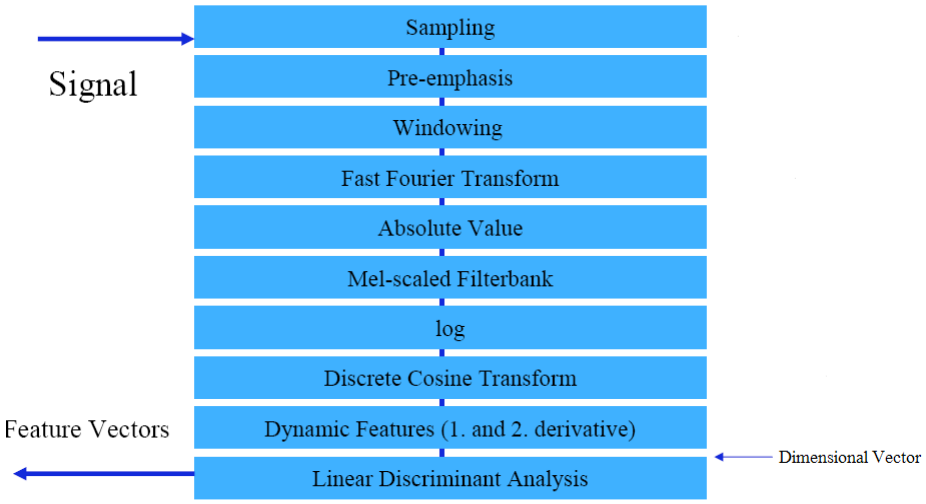


Fig.1.

3.1 Spectral Subtraction:

It is also known as Magnitude Spectral Subtraction, because it is the magnitude of the noise spectrum that is being subtracted from the noisy signal.

3.2 Scaling:

The voice signal obtained from different speakers have different ranges of amplitudes. This affects the distortion distance to be calculated at the end which directly hinders the efficiency and the error rates increase. Hence, here the voice signals are scaled or mapped to a pre-defined amplitude range. This makes sure that allthe voice signals have the same amplitude ranges and hence does not contribute towards reduction of efficiency. If scaling is not done, then if the amplitude of speaker 1 is twice that of speaker 2, then the distortion distance also happens to double, which will not let us fix a threshold value to decide if the speaker and claimed identity are matching. But if the scaling is done, then such problems won't arise and a threshold value with respect to distortion distance can be fixed and will help us distinguish between the false identity claimer and true identity claimer.

3.3 Framing and Windowing:

As shown in the figure below the speech signal is slowly varying over time and it is called quasi stationery.

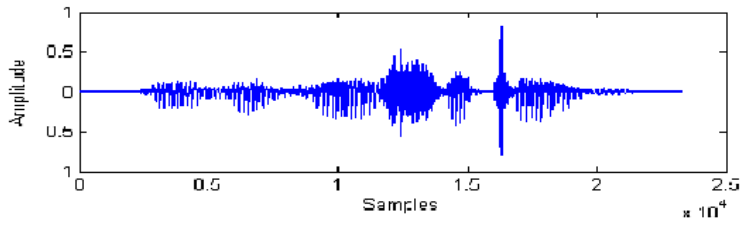


Fig.2.

Above plot shows the word spoken by speaker. The recordings were digitised at f samples is equal to 8,000 samples per second. Time goes from left to right and amplitude is shown vertically. When the speech signal is examined over a short period of time such as 5 to 100 milliseconds, the signal is reasonably stationery, and therefore this signals are examined in short time segment (this is required as already described above), short time segments is referred to as a spectral analysis. This means that the signal is blocked into 20-30 milliseconds of each frame. And to avoid the loss of any information due to windowing adjacent frame is overlap with each other by 30 percent to 50 percent. As soon as the signal has been framed, each frame is multiplied with the window function *w(n)* with length N. The function below we are using is called hamming window function
Where N = Length of the frame.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

3.4 Hamming Window:

Hamming window is also called the raised cosine window. The equation and plot for the Hamming window shown below. In a window function there is a zero valued outside of some chosen interval. For example, a function that is stable inside the interval and zero elsewhere is called a rectangular window that illustrate the shape of its graphical representation. When signal or any other function is multiplied by a window function, the product is also zero valued outside the interval. The windowing is done to avoid problems due to truncation of the signal. Window function has some other applications such as spectral analysis, filter design,and audio data compression such as Vorbis.

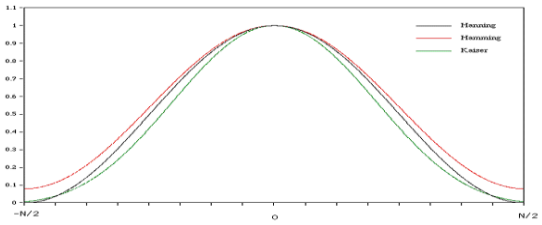


Fig.3.

3.5Cepstrum:

Cepstrum name was derived from the spectrum by reversing the first four letters of spectrum. We can say cepstrum is the Fourier Transformer of the log with unwrapped phase of the Fourier Transformer.

Mathematically we can say Cepstrum of signal = FT(log(FT(the signal))+j2_m)

Where m is the integer required to properly unwrap the angle or imaginary part of the complex log function.

Algorithmically we can say – Signal - FT - log - phase unwrapping - FT -Cepstrum.

For defining the real values real cepstrum uses the logarithm function. While for defining the complex values whereas the complex cepstrum uses the complex logarithm function. The real cepstrum uses the information of the magnitude of the spectrum where as complex cepstrum holds information about both magnitude and phase of the initial spectrum, which allows the reconstruction of the signal. We can calculate the cepstrum by many ways. Some of them need a phase-warping algorithm, others do not. Figure below shows the pipeline from signal to cepstrum.

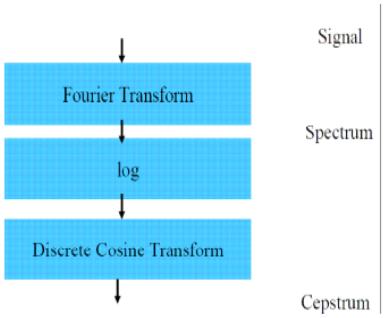


Fig.4.

As we discussed in the Framing and Windowing section that speech signal is composed of quickly varying part e(n) excitation sequence convolved with slowly varying part (n) vocal system impulse response.

$$s(n) = e(n) * \theta(n)$$

$$c_s(n) = \mathfrak{F}^{-1}\left\{ log\left|\mathfrak{F}\{s(n)\}\right| \right\}$$

Once we convolved the quickly varying part and slowly varying part it makes difficult to separate the two parts, cepstrum is introduced to separate this two parts. The equation for the cepstrum is given below:

$\mathfrak{F}$

is the Discrete Time Fourier Transformer and

$\mathfrak{F}^{-1}$

is the Inverse Discrete Time Fourier Transformer. By moving the signal from time domain to frequency domain convolution becomes the multiplication.

The multiplication becomes the addition by taking the logarithm of the spectral magnitude.

$$S(\omega) = E(\omega)\Theta(\omega)$$

$$log|S(\omega)| = log|E(\omega)\Theta(\omega)| = log|E(\omega)| + log|\Theta(\omega)| = C_e(\omega) + C_\theta(\omega)$$

The Inverse Fourier Transform work individually on the two components as it is linear.

$$c_s(n) = \mathfrak{F}^{-1}\{C_e(\omega) + C_\theta(\omega)\} = \mathfrak{F}^{-1}\{C_e(\omega)\} + \mathfrak{F}^{-1}\{C_\theta(\omega)\} = c_e(n) + c_\theta(n)$$

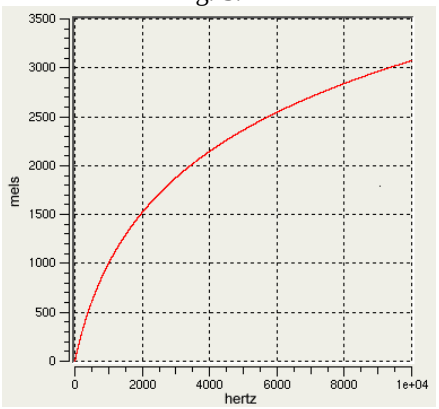
The domain of the signal *cs*(*n*) is called the quefrency-domain.

3.6 Mel Frequency Cepstral Coefficients (MFCC):

In this project we are using Mel Frequency Cepstral Coefficient. MFCC is based on the human peripheral auditory system.Mel frequency Cepstral Coefficients are coefficients that represent audio based on perception. This coefficient has a great success in speaker verification application. It is derived from the Fourier Transform of the audio clip. The human ear does not show a linear frequency resolution but builds several groups of frequencies and integrates the spectral energies within a given group. Furthermore, the mid-frequency and bandwidth of these groups are non-linearly distributed. The frequency bands are positioned logarithmically, whereas in the Fourier Transform the frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. These coefficients allow better processing of data. In the Mel Frequency Cepstral Coefficients the calculation of the Mel Cepstrum is same as the real Cepstrum except the Mel Cepstrum's frequency scale is warped to keep up a correspondence to the Mel scale.

The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units mel.The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as1000 Mels. Figure below shows the example of normal frequency is mapped into the Mel frequency.

Fig. 5.



$$m = 1127.01048 \log_e(1 + f/700)$$

$$f = 700(e^{m/1127.01048} - 1)$$

The above equationsshow the mapping the normal frequency into the Mel frequency and the inverse to get back the normal frequency respectively.

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The Mel frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz .As a reference point, the pitch of a 1 KHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. Here we are using thebank filter to warping the Mel frequency. Utilizing the bank filter is much more convenient to do Mel frequency warping, with filters centred according to Mel frequency. According to the Mel frequency the width of the triangular filters vary and so the log total energy in a critical band around the centre frequency is included. After warping are a number of coefficients. At times the results are optimized by using a certain number of mel filters (12 filters is default in case of melcepst.m MATLAB file).

Finally we are using the Inverse Discrete Fourier Transformer for the cepstral coefficients calculation. In this step we are transforming the log of the quefrency domain coefficients to the frequency domain. Where N is the length of the DFT we used in the cepstrum section.

$$Y(k) = \sum^{N/2} S(i) H_k(i)$$

$$c(n) = \frac{1}{N'} \sum_{k=0}^{N'-1} Y(k) e^{j\frac{k2\pi}{N'}n}$$

Our approach is to simulate the subjective spectrum using a filter bank, one filter for each desired Mel-frequency component. The filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant Mel frequency interval. The Mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a Mel frequency scale.

The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by the number of filters. The experiments were conducted for 12, 22, 32, 40, 48, 64, 72 filters. The efficiency is good for 12 filters but to improve the efficiency further, the results obtained by varying the number of filters are taken into account. A certain voice characteristic may be best represented by the use of say 12 mel filters while another voice may be best represented by the use of say 64 mel filters. Thus, evaluating the distance by fixing the number of filters does not yield optimum result in above such cases. Hence, this approach helps in improving the efficiency of the system. When a speaker needs to be verified, then the distortion distance is calculated for varying number of filters like in the above order and minimum distance obtained against a speaker happens to match the claimed identity, then the speaker is verified. By this technique, the efficiency improves to around 85%.

3.7 Vector Quantization:

A speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. VQ is a process of taking

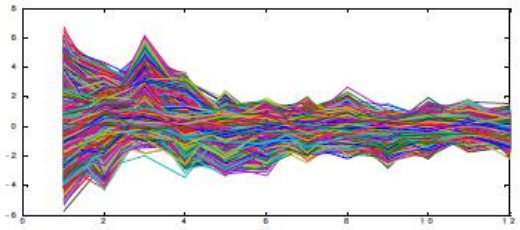


Fig.7.

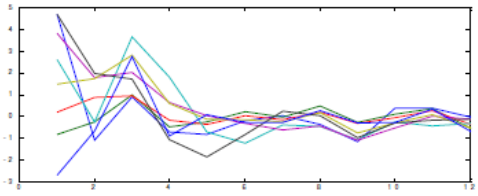


Fig 3.2 the representative feature vectors resulted after VQ

Fig.8.

a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution.

The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible.

By using these training data features are clustered to form a codebook for each speaker. In the verification stage, the data from the tested speaker is compared with the codebook of the speaker whose identity is claimed upon and measure the difference. This difference is then used to make the verification decision.

3.8 K-means Algorithm:

The K-means algorithmis a way to cluster the training vectors to get feature vectors. In this algorithm clustered the vectors based on attributes into k partitions. It uses the *k* means of data generated from Gaussian

$$V = \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2$$

distributions to cluster the vectors. The objective of the k-means is to minimize total intra-cluster variance, V. Where there are *k* clusters *Si*, *i* = 1,2,...,*k* and μ*i* is the centroid or mean point of all the points.

The process of k-means algorithm used least-squares partitioning method to divide the input vectors

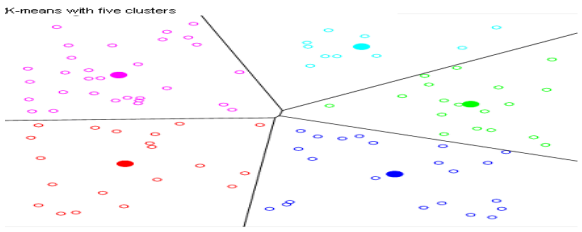
$$, x_j \in S_i$$

into k initial sets. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and

algorithm repeated until when the vectors no longer switch clusters or alternatively centroids are no longer changed.

1. Distance measure:

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of feature vector {x1, x2 ….xi), and then it is compared with the codebook of the claimedspeaker fromthe database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector



setsand identifying the minimum Euclidean distance.

Fig.9.

The Euclidean distance is the "ordinary" distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem. The formula used to calculate the Euclidean distance can be defined as following:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

i.e. the Euclidean distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn).
If the difference is below a threshold then, the speaker is verified/ matched.

2. Simulation Results:

The project is implemented on MATLAB. The screenshot of the result is given below. The 'd' variable is a 7x6 matrix which is the set of distortion distances for seven different number of filters i.e. 12, 22, 32, 40, 48, 64, 72. The six columns represent six different speakers. The speaker matches with the speaker 5 because the minimum

```
d =

   15.2315   14.2829   15.8745   28.5657   11.3137   13.8564
   14.0000   13.8564   19.1833   20.7846    6.9282   23.5797
   13.4164   14.2829    6.9282    6.9282    3.4641   20.7846
   12.0000   13.8564   15.8745    6.9282    6.9282    8.9443
   10.3923   14.2829   31.1769   38.3667   17.4356    3.4641
   18.3303   13.8564   13.2665   31.7490   21.1660   23.5797
   14.2829    3.4641    3.4641   34.6410         0   12.8062

   speaker  matches with speaker5

 >>
```

distortion distance is found against speaker 5 and this matches the claimed identity.

IV. Conclusion:

This paper has shown that the combination approach of MFCC and Distance measure can improve the performance of speaker verification systems. It helps in authentication services. The distortion distance calculation method for speaker verification is the simplest algorithm unlike Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). Techniques like Scaling and Spectral Subtraction and considering the results of various number of mel filters help in improving the efficiency of this algorithm, almost comparable to using the GMM and HMM models. By this technique, the efficiency is increased up to 88%. However the efficiency is high with less number of speakers to be verified against. Thus, this happens to be a simpler and yet efficient way for the speaker verification task and also is text independent and can be used for many applications like authenticating the user on a computer or many such simple, cost effective applications.

References:

[1]. Journal paper: Sujit Kumar Behera, Jatindra Kumar Singh, Speaker Verfication using Mel frequency cepstral coefficient and artificial neural network
[2]. Journal paper: Mohd Zaizu Ilyas, Salina Abdul Samad, Aini Hussain, Khairul Anuar Ishak, Speaker Verification using Vector Quantization and Hidden Markov Model.
[3]. Thesis submitted on Kernel Based Learning Methods for Pattern and Feature Analysis by WU Zhili, Hong Kong University.
[4]. The physiology of speech production
[5]. Paper by Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbai Md. Saifur Rahman on Speaker Identification Using Mel Frequency Cepstral Coefficients.
[6]. Review of different techniques for speaker recognition system by Bansod.N.S., Seema Kawathekar and Dabhade S.B.
[7]. ICME 2004 Tutorial on Audio Feature Extraction by George Tzanetakis, University of Victoria, Canada.
[8]. About the voice: http://www.lionsvoiceclinic.umn.edu/page2.htm (last viewed May, 2013).