IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 6, Ver. II (Nov.- Dec. 2017), PP 13-25
www.iosrjournals.org

A Signal processing approach to Music tutor

Md. Tareq Hasan[1], AmathulHadi Shakara[2]
[1]*Department of Computer Science and Engineering, University of Development Alternative, Bangladesh,*
[2]*Department of Electronic and Telecommunication Engineering, University of Development Alternative, Bangladesh*

Abstract: *Computer science plays a vital role providing us different kinds of software and the facility to learn singing accurately and smartly. This paper is to provide a singing tutor which will compare the learner song with the tutor song and provide the result of accuracy in the learner song. So any one sitting in his/her own room with a laptop or desktop computer can learn singing and judge the accuracy of the song. Firstly, to detect voiced and unvoiced part of the song, linear model is used here. Apply EMD (Empirical mode decomposition) to obtain IMFs. From the first IMF which captures most of the noise, estimate the noise level in the noisy signal and "noise only" model from the confidence interval parameter for the linear model and the corresponding model for a chosen confidence interval. Then compute the EMD of the noisy signal, and compute the IMF energies by using the confidence interval as a threshold. If there is at least one IMF whose energy exceeds the threshold then the speech frame is classified as voiced. If all the IMFs lie below the threshold then the speech frame is classified as unvoiced. Pitch is estimated from the voiced part only, so after separating the voiced part, weighted autocorrelation method is used here to estimate pitch and by using Hilbert transformation, energy is estimated. Then compare the pitch and energy of tutor song and student song to observe the difference and mistakes in the learner's song. By this way any one can identify the places in the song where he/she has mistaken, identify the scale and tune through which he/she can sing more accurately.*
Keyword: *EMD, IMF, AMDF, AUTOC.*

Date of Submission: 23-10-2017 Date of acceptance: 21-11-2017

I. Introduction

Music is one of the spiritual happiness exist in the earth. Singing is the act of producing musical sounds with the voice, anyone who can speak can sing, since singing resembles sustained speech. A singing tutor or a vocal coach is a music teacher who instructs singers on how to sing a song accurately. To make the concept of singing training or learn singing more convenient and cost free, our going trend of technology and science keep a dominating roll and it provides the facility to learn singing accurately and smartly. This is really very effective way to learn singing, it is time consuming and easy also. Learner never needs to feel shy in front of the singing tutor, because the tutor is not a human. This is a repetitive way where a learner could learn a specific song accurately without any bindings. There are four physical processes involved in producing vocal sound: respiration, phonation, resonation, and articulation [1]. Singers should be thinking constantly about the kind of sound they are making and the kind of sensations they are feeling while they are singing [2]. Vocal exercises is needed to sing perfectly, it has purposes, including [3] warming up the voice; extending the vocal range; "lining up" the voice horizontally and vertically; and acquiring vocal techniques such as legato, staccato, control of dynamics, rapid figurations, learning to sing wide intervals comfortably, correcting vocal faults. Singing has a well-defined technique that depends on the use of the lungs, the larynx, the chest and head cavities and the tongue. Though these four mechanisms function independently, they are nevertheless coordinated in the establishment of a vocal technique and are made to interact upon one another [4]. Another major influence on vocal sound and production is the function of the larynx which people can manipulate in different ways to produce different sounds.[5] These different kinds of laryngeal function are described as different kinds of vocal registers [6][7]. Registers originate in laryngeal function. They occur because the vocal folds are capable of producing several different vibratory patterns. Vibrato is the pulse or wave in a sustained tone. Each of these vibratory patterns appears within a particular range of pitches and produces certain characteristic sounds [8]. The term "register" can be somewhat confusing as it encompasses several aspects of the human voice; this view is also adopted by many vocal pedagogics [9]. Vocal resonation is another important term in sound production. McKinney defines it as a process by which the basic product of phonation is enhanced in timbre and/or intensity by the air-filled cavities through which it passes on its way to the outside air. There are seven areas that may be listed as possible vocal resonators. In sequence from the lowest within the body to the highest, these areas are the chest, the tracheal tree, the larynx itself, the pharynx, the oral cavity, the nasal cavity, and the sinuses[10].

The first recorded mention of the terms chest voice and head voice [11] was around the 13th century, when it was distinguished from the "throat voice" by the writers Johannes de Garlandia and Jerome of Moravia. [12] The terms were later adopted within Bel canto, the Italian opera singing method. Another current popular approach that is based on the Bel canto model is to divide both men and women's voices into three registers. Men's voices are divided into "chest register", "head register", and "falsetto register" and woman's voices into "chest register", "middle register", and "head register". Such pedagogics teach that the head register is a vocal technique used in singing to describe the resonance felt in the singer's head [13]. The contemporary use of the term chest voice is often applied throughout the modal register. Chest timbre can add a wonderful array of sounds to a singer's vocal interpretive palette [14]. However, the use of overly strong chest voice in the higher registers in an attempt to hit higher notes in the chest can lead to forcing. Forcing can lead consequently to vocal deterioration [15]. The pitch of the voice is controlled by the cycle of the cyclic pulse string. In the case of the vocal tract resonance filter, differences in tone resulting from phonemes such as the Japanese syllables "a" or "ka" are controlled by modifying resonance characteristics (voice spectrum characteristics). The vocal tract shape is determined from the position of the vocal organs, and speech is produced by using the vocal tract area [16].

II. Related Works

There are many work exits based on signal processing about singing tutor like "A Program to Teach Sight-Singing" [17], "Application-Specific Music Transcription for Tutoring" [18], "Visual displays for the assessment of vocal pitch matching development" [19] etc. This paper will help a learner to learn any song correctly. By learning many songs in this way, the learner will be a good singer with the help of this singing tutor.

III. Role of signal processing in singing tutor

One defining characteristic of speech is its pitch, but it is oftentimes difficult to obtain this value because some segments of speech simply do not have a measureable pitch. This paper explores techniques in determining the pitch of a speaker in a noiseless utterance, starting with the classification of types of speech and automatic pitch detection.

3.1. Speech classification

Speech can be classified into two general categories, voiced and unvoiced speech. A voiced sound is one in which the vocal cords of the speaker vibrate as the sound is made, and unvoiced sound is one where the vocal cords do not vibrate. A given speech utterance could contain a mix of different voiced and unvoiced segments depending upon the. Pitch detection relies on the periodic qualities of the sound waveform, therefore any attempt to determine pitch is only valid on voiced segments of an utterance. This factor necessitates a secondary step before pitch detection is the voiced segments identification.

3.2. Pitch [20]

Pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. **(Verbal definition)**

Sound has a certain pitch if it can be reliably matched by adjusting the frequency of a sine wave of arbitrary amplitude. **(Operational definition)**

3.3. Fundamental frequency

The fundamental tone, often referred to simply as the fundamental and abbreviated f_0 or F_0, is the lowest frequency in a harmonic series. The fundamental frequency of a periodicsignal is the inverse of the period length. The period is, in turn, the smallest repeating unit of a signal. One period thus describes the periodic signal completely. The significance of defining the period as the *smallest* repeating unit can be appreciated by noting that two or more concatenated periods form a repeating pattern in the signal. However, the concatenated signal unit obviously contains redundant information. The fundamental frequency is the lowest frequency component of a signal that excites (imparts energy) to a system.

Fundamental frequency vs. pitch [20]

- Fundamental frequency (F0) is a physical term; Pitch is a perceptual term (perceived F0)
- Both measured in Hertz (Hz), usually pitch (perceived F0) ~ F0 (approximately equal)

3.4. Energy

Since we often think of signal as a function of varying amplitude through time, it seems to reason that a good measurement of the strength of a signal would be the area under the curve. However, this area may have a

negative part. This negative part does not have less strength than a positive signal of the same size (reversing your grip on the paper clip in the socket is not going to make you any more lively). This suggests squaring the signal or taking its absolute value, then finding the area under that curve. It turns out that what we call the energy of a signal is the area under the squared signal.

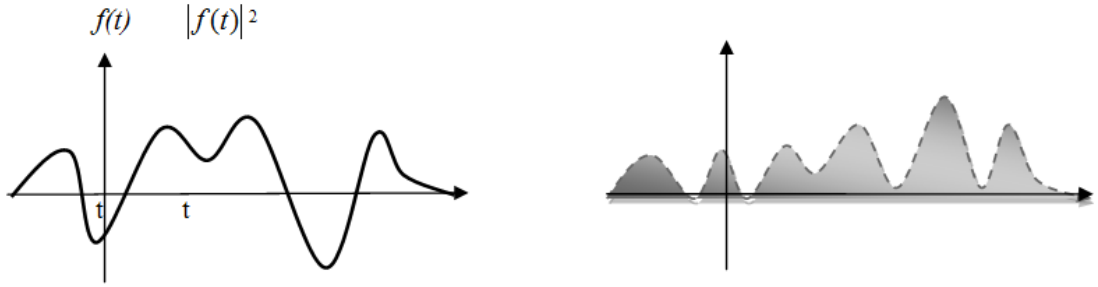


Figure 1: The energy of this signal is the shaded region.

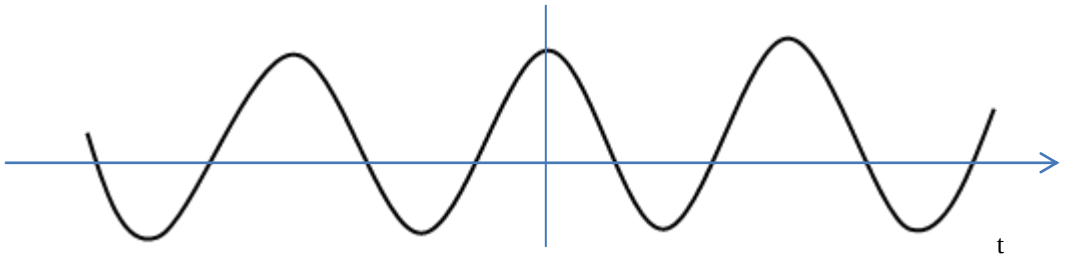


Figure 2: A simple, common signal with infinite energy.

IV. Voiced/Unvoiced detection

We need to know about EMD to detect voiced unvoiced part of the signal. A short note is given below about EMD and IMF, after that there is a big description about the classification of voiced unvoiced part and its detection.

4.1 Introduction to EMD and IMF

The fundamental part of the HHT is the empirical mode decomposition (EMD) method. Using the EMD method, any complicated data set can be decomposed into a finite and often small number of components, which is a collection of intrinsic mode functions (IMF). An IMF represents a generally simple oscillatory mode as a counterpart to the simple harmonicfunction. By definition, an IMF is any function with the same number of extremer and zero crossings, with its envelopes being symmetric with respect to zero. The definition of an IMF guarantees a well-behaved Hilbert transform of the IMF. This decomposition method operating in the time domain is adaptive and highly efficient. Since the decomposition is based on the local characteristic time scale of the data, it can be applied to nonlinear and non stationary processes.

4.2 Voiced Unvoiced Classification Methods

Voiced sounds, e.g., 'a', 'b', are essentially produced by vibrating the vocal cords, and are oscillatory. Therefore, over short periods of time, they are well modeled by sums of sinusoids. This makes short-time Fourier transform—to be discussed later—a useful tool for speech processing. Unvoiced sounds such as's', 'sh', are more noise-like, as shown in Figure 3. For many speech applications, it is important to distinguish between voiced and unvoiced speech.

4.2.1 Short-time power function

Split the speech signal *x (n)* into blocks of 10-20 ms, and calculate the power within each block:

$$P_{av} = \sum_{n=1}^{L} x^2(n) \qquad (1)$$

Typically, $P_{av,\ voiced} > P_{av,\ unvoiced}$.

Zero-crossing rate**,** "the signal *x (n)* has a zero-crossing at n_0" means that $x(n_0)\ x(n_0 + 1) < 0$. Unvoiced signals oscillate much faster, so they will have a much higher rate of zero-crossings.

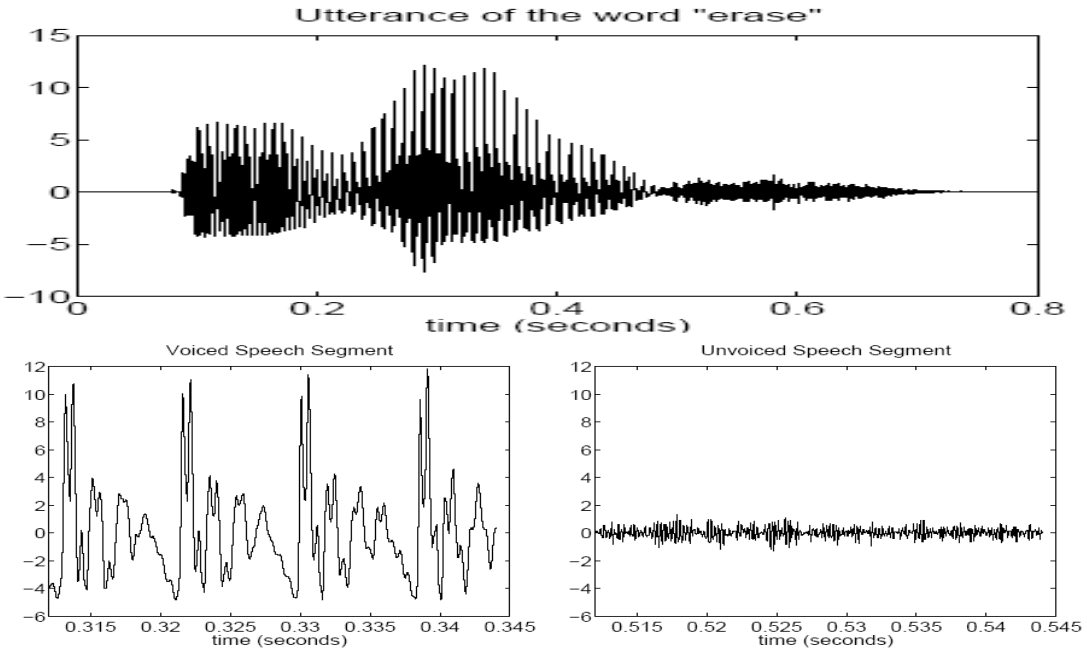


Figure 3: Distinction between voiced and unvoiced speech.

This system is illustrated in Figure 3. Its upper part (the production of voiced sounds) is very much akin to playing a guitar (Figure 3). We produce a sequence of impulsive excitations by plucking the strings, and then the guitar converts it into music. The strings are sort of like the vocal cords, and the guitar's cavity plays the same role as the cavity of the vocal tract.

4.2.2 Source-Filter Model of Speech Production

Sounds get variations in air pressure. The creation of sound is the process of setting the air in rapid vibration. Our model of speech production will have two major components:

- **Excitation**: How air is set in motion.
  Voiced sounds*:* Periodic air pulses such as in Figure 4(a) pass through vibrating vocal chords.
  Unvoiced sounds: Force air through a constriction in vocal tract, producing turbulence.
- **Vocal tract**: Guides air.

A periodic pulse train excitation is illustrated in Figure 4(a). The period *T* is called the pitch period, and 1/T is called the pitch frequency.

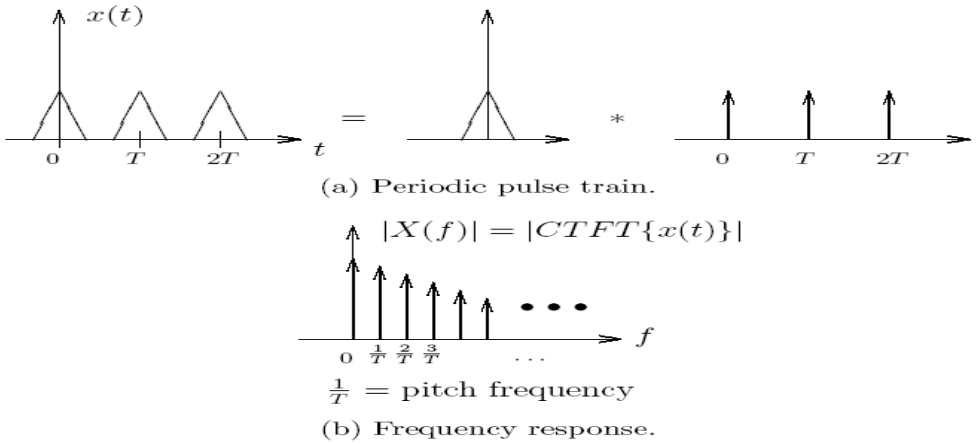


Figure 4: Time domain and frequency domain perspectives of voiced sounds.

In other words, each sound is approximately periodic, but different sounds are different periodic signals. This implies that we can model the vocal tract as an LTI filter over short time intervals. Moreover, since the vocal tract is a cavity, it resonates. In other words, when a wave propagates in a cavity, there is a set of

frequencies which get amplified. They are called natural frequencies of the resonator, and depend on the shape and size of the resonator. Therefore, the magnitude response of the vocal tract for one voiced sound (phoneme) can be modeled as in Figure 5.

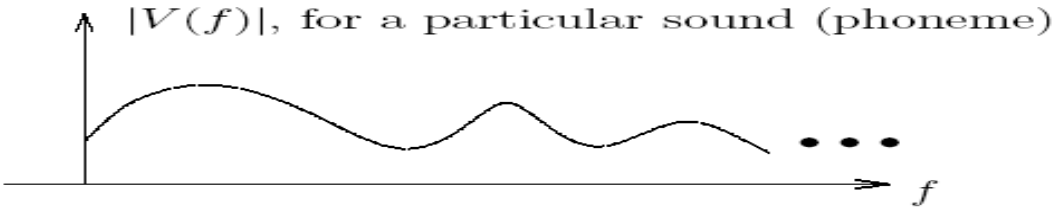
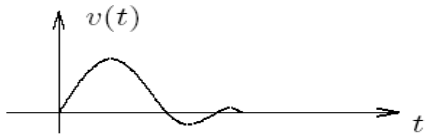
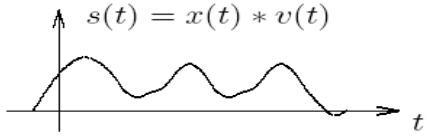


Figure 5: Magnitude response of the vocal tract.

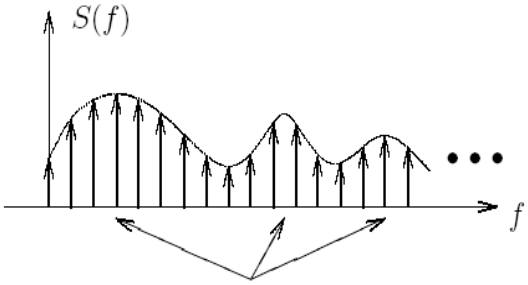
This waveform will then be the convolution of the driving periodic pulse train *x (t)* with the impulse response *v (t),* as illustrated in Figure 6 (b), and the magnitude of its spectrum |*S (f)* will be the product of *X (f)* and the magnitude response |*V (f)*|, as illustrated in Figure 6 (c).The maxima of |*S (f)*| are called the formant frequencies of the phoneme.



(a) Voice tract transfer function.



(b) Resulting sound wave.



(c) Magnitude response of the resulting sound wave.

Figure 6: The vocal tract as an LTI filter.

4.3 Proposed Voiced Unvoiced Detection Method

This algorithm is based on linear model of noise filtering by empirical mode decomposition. Weighted autocorrelation is used for pitch period calculation. The general concept is, the voiced speech is quasi-periodic and unvoiced speech is non-periodic. The goal of this algorithm is to detect periodicity of each noisy speech frame. But accurate speech periodicity detection is still a challenge. Here, I newly developed linear model for noise filtering is used which is robust voiced unvoiced classification.

Basically Gaussian noise (*fGn*) is a generalization of ordinary white noise. It is a versatile model for a homogeneously spreading broadband noise without any dominant frequency band. The statistical properties of *fGn* are entirely determined by its second-order structure, which depends solely upon one single scalar parameter, *H*, its Hurst exponent. More precisely, $\{x_H[n], n=\ldots,-1,0,1\ldots\}$ is a *fGn*of index *H* (with $0 < H < 1$) if and only if it is a zero-mean Gaussian stationary process whose autocorrelation sequence $r_H[k] := E\{x_H[n]\, x_H[n+k]\}$ is

$$r_H[K] = \frac{\sigma^2}{2}(|K-1|^{2H} - 2|K|^{2H} + |K+1|^{2H}) \quad (2)$$

In linear model for filter-bank the Hurst exponent varies between0.1 to 0.9. In filter-bank structure we can improve our search by self-similarity search which means that

$S_{k',H}(f) = \rho_H^{\alpha(k'-k)} S_{k,H}(\rho_H^{k'-k} f)$(3)

for some α and any $k' > k \geq 2$. Consequently, the power spectra of all IMFs should collapse onto a single curve when properly renormalized. Indeed, the value of is set $\alpha = 2H-1$. From equation (3) of self-similarity relation for band-pass IMFs (index k>1), we can deduce how the variance should evolve as a function of *k*. Assuming that equation (3) holds for any $k' > k \geq 2$ and $\alpha = 2H - 1$, *we have*

$$V_H[k'] = var d_{K,H}[n] = \int_{-1/2}^{1/2} S_{k,H}(f)df \quad (4)$$

$$V_H[k'] = \rho_H^{\alpha(k'-k)} \int_{-1/2}^{1/2} S_{k,H}(\rho_H^{k'-k} f)df \quad (5)$$

$$V_H[k'] = \rho_H^{(\alpha-1)(k'-k)} V_H[k] \quad (6)$$

Which leads to

$$V_H[k] = C\rho_H^{2(H-1)k} \quad (7)$$

The IMF variance should be an exponentially decreasing function of the IMF index with a decay rate which is a linear function of the Hurst exponent *H*. The linear model with regards to the variability of the variance estimate, various confidence intervals is written as:

$\log_2 V_H[k] = \log_2 \hat{V}[2] + 2(H-1)(k-2)\log_2 \rho_H$ (8)

For k≥*2*, where

$$\hat{V}_H[k] = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{N}\sum_{n=1}^{N}\left(d_{k,H}^{(j)}[n]\right)^2\right] \quad (9)$$

is a function of the index *k*.

The parameters of confidence interval parameters have shown in Table 1. For white noise, the number of parameterize $T_H[k]$ is obtained by using the formula:

$\log_2(\log_2(T_H[k]/W_H[k])) = \alpha_H k + b_H$ (10)

Where$W_H[k]$ denotes the H-dependent variation of the IMF energy. The best linear fit occurs when the median of the IMF's energy is used to compute W_H over the realizations. The parameters α_H and $b_H$are obtained from Table 4.1. $W_H[1]$ can be estimated from

$$\widehat{W}_H[1] = \sum_{n=1}^{N} d_{1,H}^2[n] \quad (11)$$

and subsequent values of $W_H[k]$ are given by

$\widehat{W}_H[1] = C_H\rho_H^{-2(1-H)k}, k \geq 2$ (12)

Where $C_H = \hat{W}_H/\beta_H$.

Table 1: The Confidence Interval Parameters for the Linear Model.

H	β_H	α_H (95 %)	b_H(95 %)	α_H (99 %)	b_H(99 %)
0.2	**0.487**	**0.458**	**-2.435**	**0.452**	**-1.951**
0.5	**0.719**	**0.474**	**-2.449**	**0.460**	**-1.919**
0.8	**1.025**	**0.497**	**-2.331**	**0.495**	**-1.833**

Here in the proposed method the third row of Table 1 has been used as the confidence interval parameters for linear model. This model is applied to select the IMF whose energy exceeds the upper limit of 99% confidence interval of the Gaussian noise. From EMD domain all IMFs starting from that IMF are summed up to partial reconstruction of the signal. If there remains at least one IMF above the confidence interval of 99% then the speech frame is classified as voiced frame. If all the IMFs lie below that confidence interval then the speech frame is classified as unvoiced.

The proposed algorithm is given below:

1. From the first IMF which captures most of the noise, estimate the noise level in the noisy signal by computing $\hat{W}_H$ [1] (eq. 11).
2. Estimate the "noise only" model (by using eq. 11 and 12).
3. Estimate the corresponding model for a chosen confidence interval (eq. 11 and table 1).
4. Compute the EMD of the noisy signal, and compute the IMF energies by using the confidence interval as a threshold.
5. If there is at least one IMF whose energy exceeds the threshold then the speech frame is classified as voiced.
6. If all the IMFs lie below the threshold then the speech frame is classified as unvoiced.

Now, let observe how the proposed method works for voiced and unvoiced speech classification. For voiced speech classification I have taken the following voiced speech frame as shown in Figure 7.

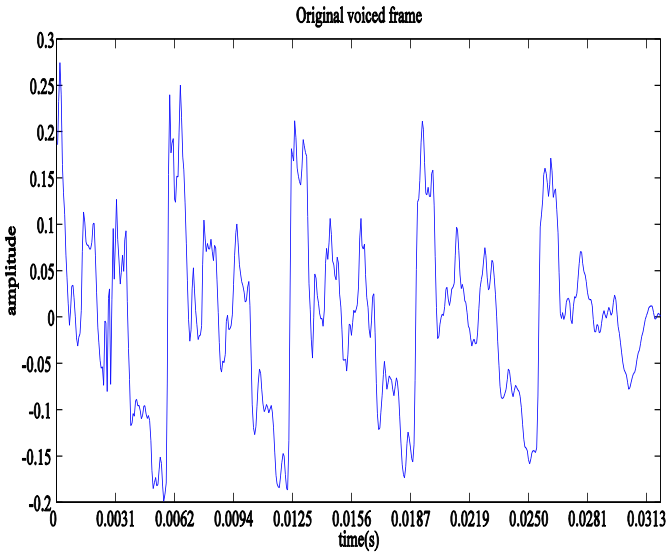


Figure 7: A voiced frame.

After applying linear model to the voiced frame which contains eight IMFs, including residue as shown in

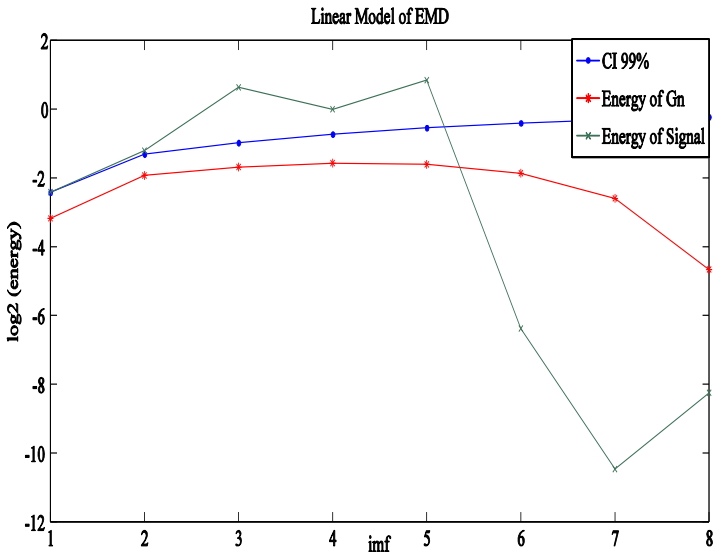


Figure 8

Figure 8: The selection of starting IMF to extract the low frequency component of the speech signal. The 3rd IMF is selected as its energy exceeds the upper limit of 99% confidence interval of the IMFs' energies of Gaussian noise (*Gn*).

From the above figure we see that more than one IMF (IMF3, 4) exceeds the confidence interval so the speech frame is voiced. From the IMFs exceed the confidence interval are summed up for partial reconstruction of the speech signal as shown Figure 9.

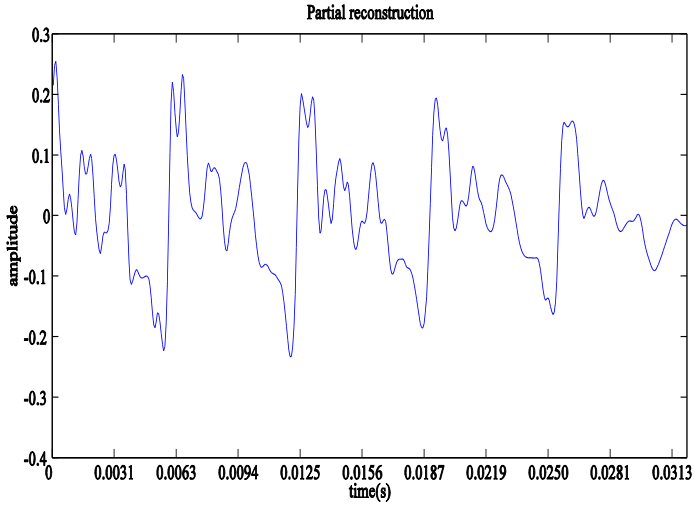


Figure 9: Partial reconstruction of speech signal from linear model.

In the next let us observe how this model works for an unvoiced speech frame and for this purpose an unvoiced frame like below is taken.

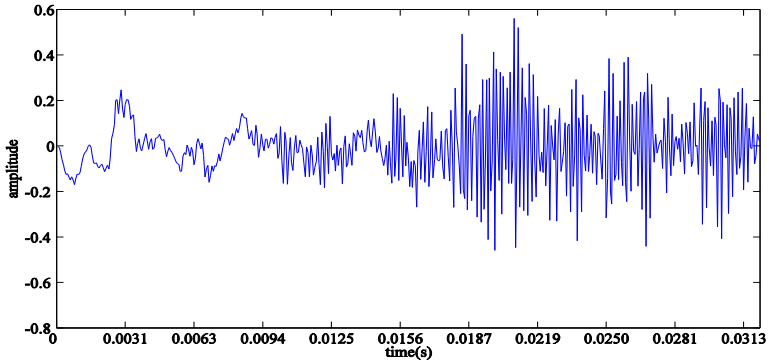


Figure 10: An unvoiced frame

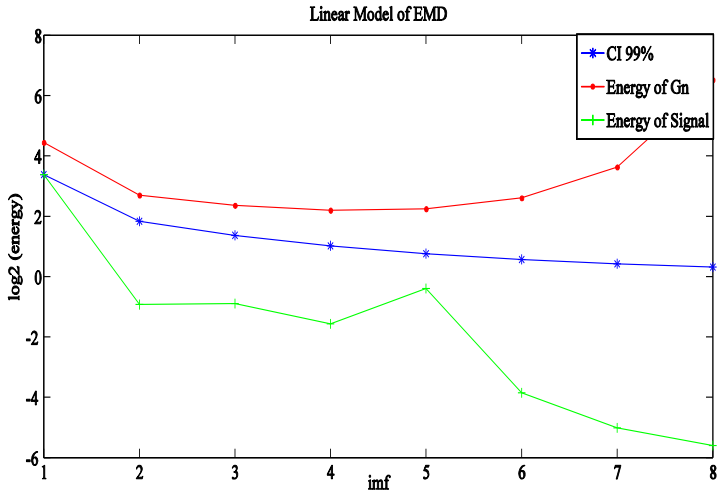


Figure 11: There exists no IMF whose energy exceeds the upper limit of 99% confidence interval of the IMFs' energies of Gaussian noise (Gn).

As we see that no IMF is selected from above the confidence interval of 99 % the frame is classified as unvoiced. Consequently no partial reconstruction has occurred.Here in this algorithm the use of the linear model is an efficient process of voiced unvoiced classification. This method works as an efficient filter-bank implementation. It is highly sensitive to noise. Due to introduce Gaussian noise we need not worried for the accuracy of voiced unvoiced classification. It is a strong mathematical theoretical foundation for noise assisted data analysis (NADA) technique for non-stationary data. It estimates the white noise added to the speech frame make a better decision to obtain the IMFs from EMD domain that contain the dominant frequency of the speech frame. The IMF selection technique is extremely a gigantic approach provided by the model. In the case shown here, we observed the "spontaneous" emergence of an equivalent filter-bank structure which has the advantage of being fully data-driven. Furthermore, because it is local in time, this structure can adopt automatically to non-stationary situations with greater flexibility than other approaches using a pre-determined decomposition scheme.

4.4 Pitch estimation

In this paper, a modified version of the autocorrelation pitches extraction method well known to be robust against noise. Utilizing that the average magnitude difference function (AMDF) has similar characteristics with the autocorrelation function, the autocorrelation function is weighted by the reciprocal of the AMDF. By simulation experiments, it is shown that the proposed pitch extraction method is useful in noisy environments.

In this paper, a new pitch extraction method which uses an autocorrelation function weighted by the inverse of an AMDF is used. The characteristics of the AMDF are very similar with those of the autocorrelation function. The AMDF produces a notch, while the autocorrelation function produces a peak. However, both functions essentially have the same periodicity. The proposed method utilizes the feature that in a noisy environment, the noise components included in the autocorrelation function and AMDF behave independently (and are uncorrelated each other). This feature will be validated in this paper. By such uncorrelated properties, the peak of the autocorrelation function is emphasized in a noisy environment when the autocorrelation function is combined with the inversed AMDF. As a result, it is expected that the accuracy of pitch extraction for the AUTOC is improved.

4.4.1. Proposed Method [21]

Principle

Autocorrelation function $\phi(\tau)$ is calculated by

$$\phi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau) \qquad (13)$$

Where,

$x(n)$Speech signal;

τLag number;

nTime for a discrete signal.

The characteristic of $\phi(\tau)$ is that $\phi(\tau)$ has a large value when $x(n)$ is similar with $x(n+\tau)$. If $x(n)$ has a period of *P* then $\phi(\tau)$has peaks at $\pi = lP$ where *l* is an integer.

Essentially, $\phi(0)$ gives the largest value among $(\tau), \tau = lp$, for *l=0,1,2*..... . The second value is given by $\phi(P)$other peaks of $\phi(\tau)$ usually decrease as τ increases. Therefore, we can estimate the pitch period P from the location of the peak at $\tau = P$.

Let us assume that is a noisy speech signal given by

$$x(n) = s(n) + w(n) \qquad (14)$$

Where $s(n)$ a clean speech issignal and $w(n)$is additive white Gaussian noise. In this case, we have an autocorrelation function given by

$$\phi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}(s(n)+w(n))(s(n=\tau)+w(n+\tau))$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}(s(n)s(n+\tau)+s(n)w(n+\tau)+w(n)s(n+\tau)+w(n)w(n+\tau))$$

$$= \phi_{ss}(\tau) + 2\phi_{sw}(\tau) + \phi_{ww}(\tau) \qquad (15)$$

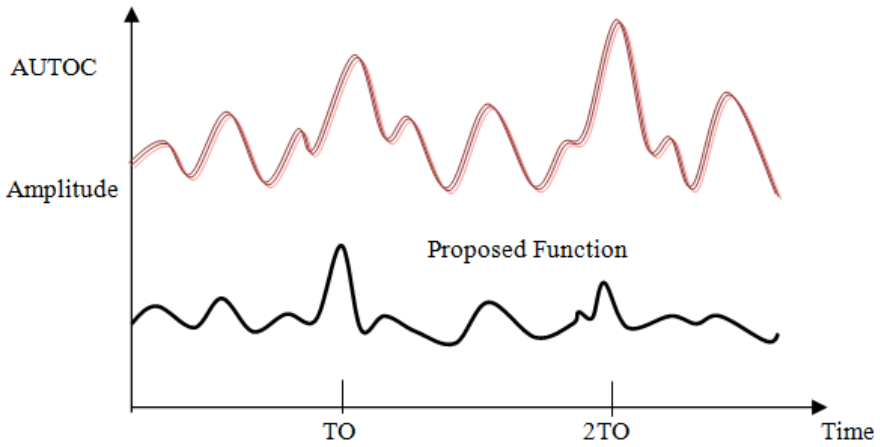


Figure 12: Autocorrelation function and proposed function, correspond to the true pitch period.

Where
$\phi_{ss}(\tau)$Autocorrelation function of $s(n)$;
$\phi_{sw}(\tau)$Cross-correlation function of $s(n)$and $w(n)$;
$\phi_{ww}(\tau)$Autocorrelation function of $w(n)$.

For large N, if $s(n)$does not correlate with $w(n)$then $\phi_{sw}(\tau) = 0$.
Furthermore, if $w(n)$ is uncorrelated, then $\phi_{ww}(\tau) = 0$ except for $(\tau) = 0$
In such a case, the relations

$$\phi(\tau) = \phi_{ss}(\tau) + \phi_{ww}(\tau)(\tau = 0) \qquad (16)$$
$$\phi(\tau) = \phi_{ss}(\tau)(\tau \neq 0) \qquad (17)$$

are valid. Based on these properties, the AUTOC provides robust performance against noise.

The autocorrelation function with the period of *P* has some peaks at the locations of *lP*. Although the maximum peak is located at except for the case of$\tau = 0$, in some cases, the peak located at $\tau = 2P$ becomes larger than that located at, as shown in Fig. 12. Then, a half pitch error occurs. On the other hand, as also shown in Fig. 12, a peak is often made at. This situation, in some cases, leads to a double-pitch error. For such reasons, if unnecessary peaks of autocorrelation function as shown in Fig. 12 are suppressed somehow, then it is expected that the accuracy of pitch extraction becomes higher.
For the purpose of emphasizing the true peak the AUTOC makes, we propose an autocorrelation function weighted by an in versed AMDF. The AMDF is described by

$$\psi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} |x(n) - x(n+\tau)|. \qquad (18)$$

The AMDF has the characteristic that when is $x(n)$ similar with$x(n+\tau), \psi(\tau)$ becomes small. This means that if $x(n)$has a period of$P, \psi(\tau)$ produces a deep notch at $\tau = P$ Therefore, $1/\psi(\tau)$makes a peak at $\tau = P$
Furthermore, the additive noise $w(n)$included in $\psi(\tau)$ behaves independently with that included in$\phi(\tau)$ (see Appendix). Hence, using the autocorrelation function weighted by$1/\psi(\tau)$, it is expected that the true peak is emphasized, and as a result the errors of pitch extraction are decreased.

The proposed function is given by

$$\eta(\tau) = \phi(\tau)/(\psi(\tau) + k) \qquad (19)$$

where*k* is a fixed number *k>0* The AMDF in (17) provides

$$\psi(0) = 0 \qquad (20)$$

which invokes a divergence of the directly inversed AMDF. For this reason, the denominator in (18) is stabilized by adding the number K.

Fig. 12 shows the autocorrelation and proposed functions obtained for a speech signal corrupted by noise. In this case, by picking the maximum amplitude of each function, the proposed function leads to the true pitch, while the autocorrelation function does an erroneous one.

4.4.2. Implementation

Pitch of the segmented speech is estimated by searching the peak of the weighted function. However, if we use the weighted function directly, the accuracy of pitch extraction is not so accurate. Therefore, the proposed system uses interpolation based on 3 points around the detected peak. It is known that such interpolation on the autocorrelation function is useful for improving the accuracy of pitch extraction. The interpolation operation used in this paper is based on Lagrange's method. The region for searching the pitch peak is set to be from 50 Hz to 400 Hz, which corresponds to the region of the fundamental frequencies of most men and women [21].

We have proposed a modified version of the autocorrelation method for pitch extraction. Based on the experimental results it has been shown that the proposed method is useful in noisy environments. Especially, it has been asserted that the proposed method provides a moderate improvement relative to the conventional autocorrelation method at very low SNR.

V. Experimental results

5.1. Data:

We have taken two songs, one is tutor song, and another is student song. Now we will compare the pitch and energy of tutor and student song, and find out all the mistakes the student has made. The lengths of these two songs are almost same. They are about 45scecond of a song, just a part of a song for analysis only; this is of course applicable for a full song.

5.2 Comparison of pitch:

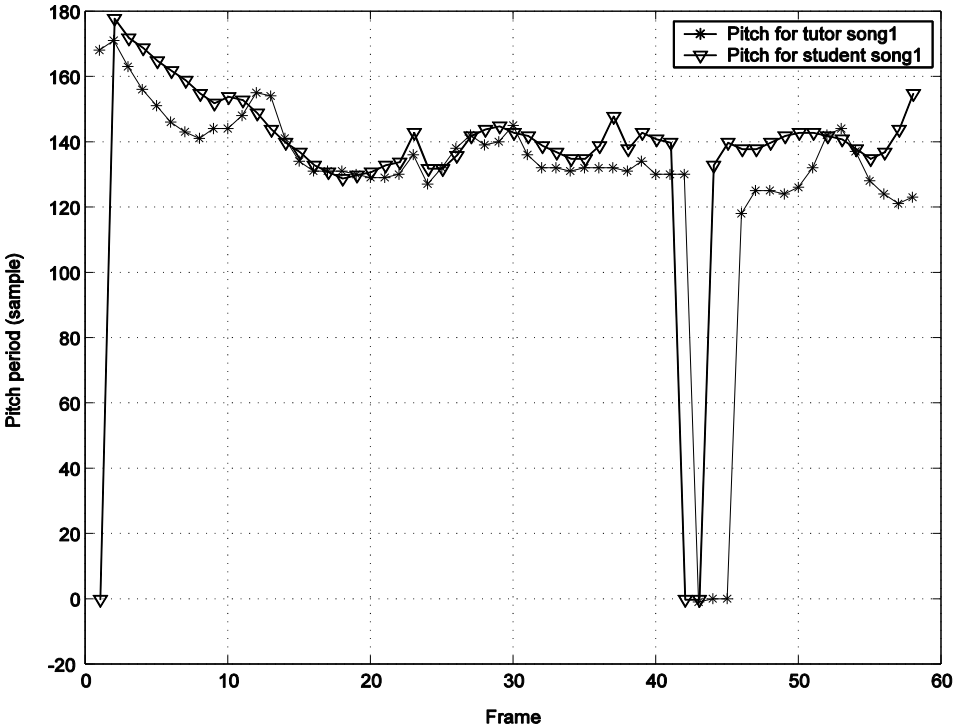


Figure 13: Pitch comparison for song 1;

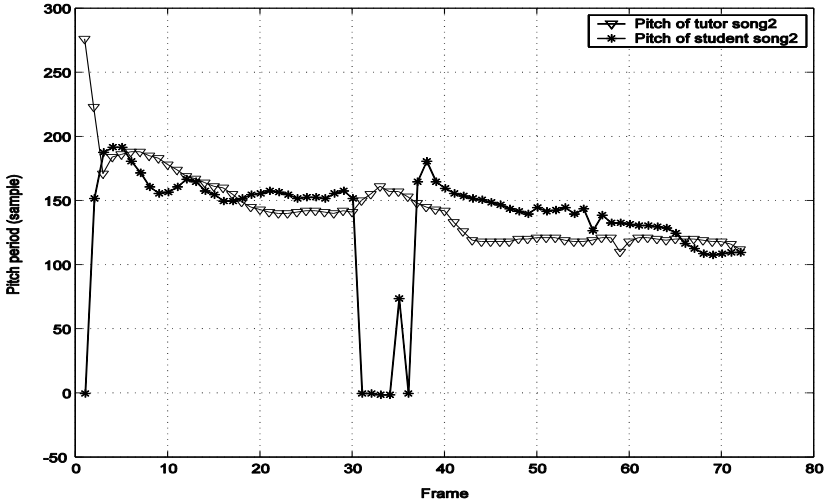


Figure 14: Pitch comparison for song 2.

5.3 Comparison of energy:

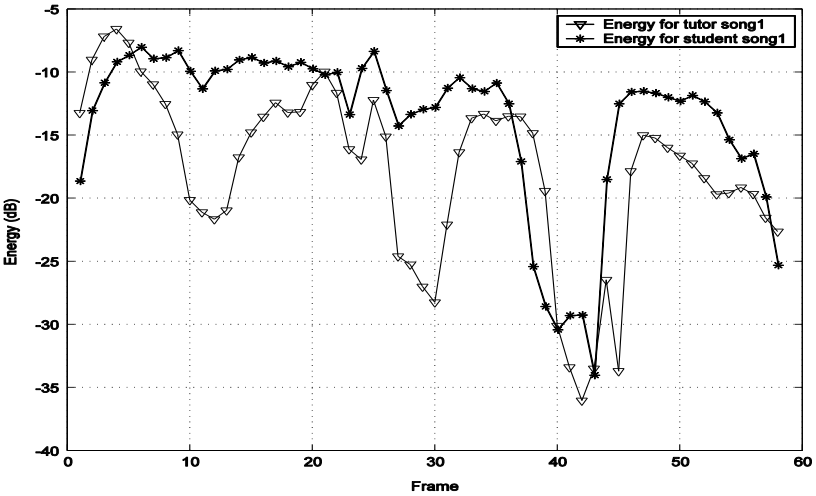


Figure 15: Energy comparison for song 1.

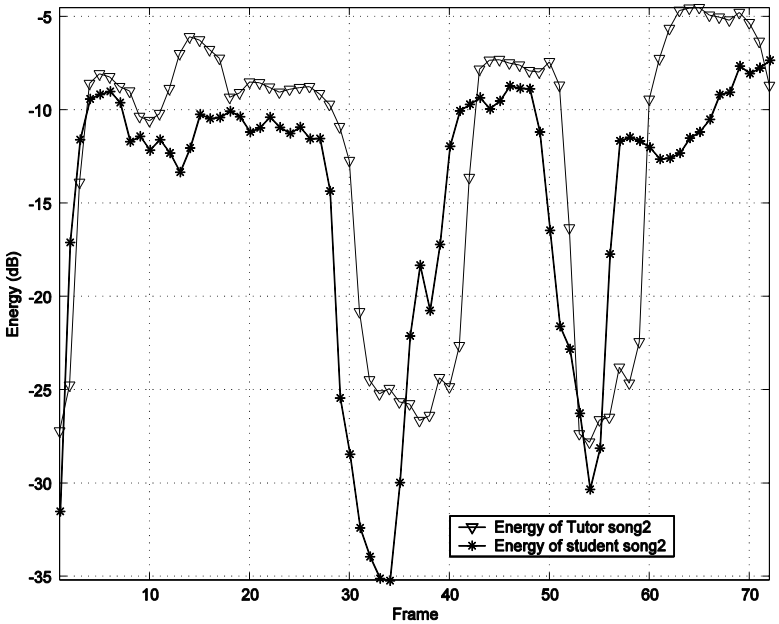


Figure 16: Energy comparison for song 2

VI. Discussion and conclusion

There are many people interested in learning singing, this is a better way to learn singing, by comparing the pitch and energy, anyone can identify the faulty places in his/her song. The energy could be plotted from different places during male and female voice comparison, we will observe their changing pattern and their distance, if these two things are same, then the song is right. This is a beginning concept of singing tutor. By developing the model and the structure of coding we can use this more conveniently and smoothly.

References

[1]. M. Honda, NTT CS Laboratories, Speech synthesis technology based on speech production mechanism, How to observe and mimic speech production by human, Journal of the Acoustical Society of Japan, Vol.55, No. 11, pp. 777-782, 1999 (in Japanese).
[2]. Appelman, Dudley Ralph (1986). The science of vocal pedagogy: theory and application. Bloomington, Indiana: Indiana University Press. pp. 434.ISBN 0253351103.OCLC13083085
[3]. McKinney, James C (1994). The diagnosis and correction of vocal faults. Nashville, Tennessee: Genovex Music Group. pp. 213.ISBN 1565939409.OCLC30786430
[4]. "Singing".Britannica Online Encyclopedia. http://www.britannica.com/EBchecked/topic/545880/singing.
[5]. Vennard, William (1967). Singing: the mechanism and the technic. New York: Carl Fischer. ISBN 978-0825800559. OCLC248006248.
[6]. Hunter, Eric J; Titze, Ingo R (2004). "Overlap of hearing and voicing ranges in singing."(PDF).J Singing61 (4): 387–392. http://web.ku.edu/~cmed/923/Hunter1.pdf.
[7]. Hunter, Eric J; Švec, Jan G; Titze, Ingo R (December 2006). "Comparison of the produced and perceived voice range profiles in untrained and trained classical singers". J Voice20 (4): 513–526. doi:10.1016/j.jvoice.2005.08.009. PMID16325373.
[8]. Large, John W (February/March 1972). "Towards an integrated physiologic- acoustic theory of vocal registers".The NATS Bulletin28: 30–35. ISSN0884-8106.OCLC16072337.
[9]. McKinney, James C (1994). The diagnosis and correction of vocal faults. Nashville, Tennessee: Genovex Music Group. pp. 213.ISBN 1565939409.OCLC30786430.
[10]. Greene, Margaret; Mathieson, Lesley (2001). The voice and its disorders (6th ed.). John Wiley & Sons.ISBN 1861561961.OCLC47831173.
[11]. Grove, George; Sadie, Stanley, eds (1980). The new Grove dictionary of music & musicians.6. Macmillan. ISBN 1561591742.OCLC191123244.
[12]. Stark, James (2003). Bel Canto: A history of vocal pedagogy. Toronto: University of Toronto Press. ISBN 978-0802086143. OCLC53795639.
[13]. Clippinger, David Alva (1917). The head voice and other problems: Practical talks on singing.OliverDitson. p. 12.Singing at Project Gutenberg
[14]. Miller, Richard (2004). Solutions for singers. Oxford: Oxford University Press. pp. 286.ISBN 0195160053.OCLC51258100.
[15]. Warrack, John Hamilton; West, Ewan (1992). The Oxford dictionary of opera. Oxford: Oxford University Press. ISBN 0198691645.OCLC25409395.
[16]. S. Saito and K. Nakata, Fundamentals of Speech Signal Processing, Ohm Publishing, 1981 (in Japanese).
[17]. Lloyd A. Smith and Rodger J. McNab, Department of Computer Science,University of Waikato
[18]. Ye Wang and Bingjun Zhang National University of Singapore
[19]. David M. Howard, Signal Processing: Voice and Hearing Research Group, Electronics Department, University of York, Heslington, York, UK YO1 5DD
[20]. AnssiKlapuri, klap@cs.tut.fi, PITCH AND MULTIPITCH ESTIMATION
[21]. Tetsuya Shimamura, Member, IEEE, and Hajime Kobayashi,Weighted Autocorrelation for Pitch Extraction of Noisy Speech
[22]. Falkner, Keith, ed (1983). Voice.Yehudi Menuhin music guides. London: MacDonald Young. pp. 26.ISBN 035609099X. OCLC10418423.
[23]. W. J. Hess, Pitch Determination of Speech Signals. Berlin, Germany: Springer-Verlag, 1983.
[24]. J. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust. Vol. AU-20, pp. 367–377, Dec. 1972.
[25]. Huang, N. E. et. al., "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis", Proc. Roy. Soc. London A, Vol. 454, pp. 903-995, 1998
[26]. Wu, B. Z. and Huang, N. E., "A study of the characteristics of white noise using the empirical mode decomposition method", in the Proc. Roy. Soc. Lond. A (460), pp: 1597-1611, 2004.
[27]. Flandrin, P., Rilling, G., and Gonqalves, P., "Empirical mode decomposition as a filter bank", IEEE signal processing letters, Vol. 11, No. 2, pp: 112-114, Feb, 2004.
[28]. Masaaki Honda,Human Speech Production Mechanisms.
[29]. https://en.wikipedia.org/wiki/Source%E2%80%93filter_model_of_speech_production.

Md. Tareq Hasan A Signal processing approach to Music tutor." IOSR Journal of Computer Engineering (IOSR-JCE) , vol. 19, no. 6, 2017, pp. 13-25.