Innovative Systems Design and Engineering
ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online)
Vol.7, No.4, 2016
www.iiste.org

Extraction of Fundamental Frequency of Human Voice by Autocorrelation Technique

Dr Karabi Ganguly Bhattacharjee
Department of Biomedical Engineering, JIS College of Engineering, Kalyani, Nadia – 741235, India

Nilotpal Manna*
Department of Electronics and Instrumentation Engineering, JIS College of Engineering, Kalyani, Nadia – 741235, India

Abstract
Human beings are blessed with a wonderful tool – voice or speaking ability that assist to communicate, talk, sing, and express emotions. Like a fingerprint, one's voice is unique and distinct, and it is different from all others. It can act as an identifier. The human voice has many components created through a myriad of muscle movements. It is specifically a part of human sound production through the vocal folds (vocal cords) as the primary sound source. It is composed of a multitude of different characteristics, making each voice different; namely, pitch, tone, and rate. [1]. Vocal folds structure and size varies from person to person making the voice as unique nature. Genetics, gender as well as the size and shape of the rest of that person's body, especially the vocal tract, and the manner in which the speech sounds are habitually formed and articulated are the governing factors to describe its uniqueness. Voice characteristics can be correlated with various electrical parameters like intensity, pitch, short time spectrum, format frequencies and bandwidth, spectral correlation etc. This paper is intended for discussion and analysis the methods of extraction of characteristics of human voice, especially the pitch frequency.
Keywords: Vocal folds, Phonation, Kymographic parameters, Pitch, Short time spectrum, Average Magnitude Difference Function.

1. INTRODUCTION

Speech is an extremely complex acoustical carrier of vocal information or voice. The human voice production mechanism is extremely difficult, as humans are capable of varying extensively the functioning of their vocal organs. Generation and production of human voice are comprising of three processes; Breathing, Phonation and Resonance. During breathing the lung acting as the pump produces adequate airflow and applies air pressure on vocal folds to vibrate. The vocal cords act as vibrating valve that chops up and control the airflow from the lungs to convert into audible pulses forming the laryngeal sound source. This process is called phonation. The length as well as tension of the vocal folds is adjusted by the muscles of the larynx to 'fine-tune' pitch and tone. Basic sound produced by the vocal folds is too weak to be heard. The upper parts of the vocal tract consisting of tongue, palate, cheek, lips, etc. above the larynx are the articulators which modify the sound to be recognized as the human voice. The articulators filter the sound emanating from the larynx and interact with the laryngeal airflow to some degree to strengthen it or weaken it as a sound source. This transformation is known as resonance. Production of a natural, effective voice depends on how well these three fundamental components are balanced or coordinated.

The fundamental frequency of speech, termed as pitch is often used as an indicator of voice development. It also indirectly indicates the hormone activity. The pitch of the voice is defined as the rate of vibration of the vocal folds. The vibrations, and the speed at which they vibrate, are dependent on the length and thickness of the vocal cords, as well as the tightening and relaxation of the muscles surrounding them. Physiologically, size and structure of vocal cords of a person is different from others; thus make the pitch frequency different from anyone. Negative correlation has been observed in the kymographic parameters with fundamental voice frequency, whereas the open quotient displayed a positive correlation with the fundamental frequency. There were significant intergroup differences in the fundamental frequency, amplitude and lateral peak [2]. In this paper an investigation has been made to extract of pitch frequency that may be used as identifier.

2. BACKGROUND

Voice is a periodical signal that presents perturbations in the glottal cycles along the frequency (jitter) and intensity (shimmer) [3]. These two parameters have been described as robust measures of the biomechanical properties of the vocal folds in situations of vocal assessment, which compares the results of therapeutic interventions [4]. For this study, the algorithms developed to calculate the jitter and shimmer values were based on the proposal of Davis [5] and Cooley et al [6]. Normality values were validated in 0.18% (standard deviation of 0.1%) for jitter and 1.08% (standard deviation of 0.37%) for shimmer [7].
There are several parameters [8] to describe the speech characteristics as follow.

- *Intensity*: This can be defined as

$$E(t) = \int_{t-T/2}^{t+T/2} s^2(\tau) d\tau$$

Where S(τ) is the speech signal and T is the averaging interval. The choice of T is somewhat arbitrary within a range of 10 ms to 30 ms. The variation in intensity of speech caused by the variation both the subglottal pressure as well as the vocal tract shape as a function of time.

- *Pitch*: Pitch is the fundamental frequency of the vocal vibration. Voice pitch can be determined either in the time domain by direct measurement of the period of speech waveform or in the frequency domain computing the frequency spacing of the spectral peaks. The temporal variation of pitch or pitch contour represents the speech characteristics and has been found to be effective for speaker identification or verification.
- *Short time Spectrum*: It is the three dimensional representation of speech signal in respect of time, frequency and energy. This provides acoustical characteristics of speech completely.
- *Predictor Coefficients*: Linear prediction analysis is an important method of characterizing the spectral properties of speech in time domain. Here each sample of speech waveform is predicted as as linear weighted sum of part of few samples. The weights which minimizes the mean square error are called predictor coefficient.
- *Formant frequencies and bandwidths*: Formant frequencies are defined as the resonant frequencies of the vocal tract. These are speaker dependent, but very difficult to estimate.
- *Nasal co-articulation*: Due to slow movement of the articulators the vocal tract shape at any given time depends not only on the phoneme being spoken at that time instant but also on the neighbouring phonemes. This phenomenon is known as articulation and the nature of co-articulation in a given context is speaker dependant.
- *Spectral correlations*: Significant degree of correlation exists between the short time spectrums at different frequencies. These correlations have been found to be very consistently from one speaker to another.
- Timing and speaking rate: Relative timing of different speech events in utterances differ from speaker to speaker. This is determined by non linear deformation of the time axis of one utterance relative to that of another.

Among the above mentioned parameters, pitch contour of the voice represents represent voice characteristics most effectively. Here some methods of pitch extraction are elaborated.

3. METHODOLOGY

There are several methods to compute pitch or fundamental frequency of human voice.

1) Autocorrelation function [9] has the unique characteristics that if the signal is the period of p then autocorrelation maximas will occur at the same period of p. Therefore time differences between maximas give the period of signal,

2) Average magnitude difference function (AMDF): AMDF is also an effective algorithm [10] of pitch period measurement which can be defined by the relation

$$D_n = \frac{1}{N}\sum_{j=0}^{N-1} \left| S_j - S_{j-n} \right|,$$
$$where\ \ n = -(N-1), \dots\dots\dots -1, 0, 1, \dots\dots (N-1)$$

Where S_j are the samples of input speech and S_{j-n} are the samples shifted by n sequences. The equation can be rewritten as

$$D_n = \beta_n [\frac{1}{N}\sum_{j=0}^{N-1}(S_j - S_{j-n})^2]^{1/2}$$

β_n is the scale factor. The equation leads to the relation

$$D_n = \beta_n \, [2(R_0 - R_n)]^{1/2}$$

R_0 and R_n are the autocorrelation value at 0 and n respectively such that

$$R_n = \frac{1}{N}\sum_{-(N-1)}^{N-1} S_k . S_{k-n}$$

The properties of AMDF are accurately characterized by the equation above. It is zero at zero delay and varies with the square root of the autocorrelation function that means, null value will occur at autocorrelation maximas. Time difference of AMDF nulls is direct measurement of the pitch period.

3) Cepstrum method: A cepstrum is the result of taking the Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal. The power cepstrum in particular finds applications in the analysis of human speech. The *power cepstrum* of signal

$$= \left| \mathcal{F}^{-1} \left\{ \log(|\mathcal{F}\{f(t)\}|^2) \right\} \right|^2$$

A short-time cepstrum analysis may be used for application of pitch determination of human speech.

Here again the peaks of such processed signal are picked up as to determine the pitch period. The interaction between the formants and the fundamental frequency largely is eliminated in this method. However unfortunately

due to nonlinear logarithm operation two undesirable results are obtained: (a) the peak at the origin can no longer be considered as a reference for normalization and (b) the actual amplitude of peaks are the function not only of the number of pitch period within the window, but also of the spectral shape. Spectral shaping is largely dependent on the formant values and to a first order approximation can be considered independent of the fundamental frequency.

4) Simplified inverse filter tracking (SIFT) algorithm: This method of pitch extraction retains the advantages of autocorrelation function as well as cepstral analysis. Here speech signal is prefiltered and sampled, and then first five coefficients of short time autocorrelation sequences are estimated for an appropriate time span. Thereafter a set of linear equations as below are solved.

$$\sum_{i-1}^{4} a_i p_{i-j} = -p_i$$

j= 1, 2, 3, 4 and p = autocorrelation values

Inverse filter coefficients $\{a_i\}$ are estimated where inverse filter is defined by

$$A_Z = 1 + \sum_{i=1}^{4} a_i Z^{-i}$$

The sampled signal is passed through this filter and autocorrelation of the output is computed. The largest peak is found within the specified range that yields pitch period after proper interpolation.

4. PITCH EXTRACTION SCHEME

For extraction pitch computation by auto correlation method is less complex which has been investigated in the laboratory. The autocorrelation function of a discrete time signal may be expressed as:

$$\emptyset(k) = \sum_{n=-\infty}^{\infty} x(n)x(n+k)$$

For a random signal the appropriate expression is:

$$\emptyset(k) = lim_{n\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{n=N} x(n)x(n+k)$$

The autocorrelation function has the properties that if the signal is periodic of period P, then its autocorrelation is also periodic of P, i.e., $\emptyset(k) = \emptyset(k+P)$ and $\emptyset(0)$ attains maximum value. This means that the maximum value of autocorrelation is certain interval of P. The inverse of P is fundamental frequency of signal.

The computation of autocorrelation function is not practical as the function sequence is spread over minus infinity to plus infinity. Therefore short time representation of autocorrelation is considered. Short time autocorrelation can retain all the information of speech signal and estimate accurately the vocal tract transfer function.

1. It is an even function i.e. Φ(k) = Φ(- k).
2. It attains its maximum at k = 0 i.e. Φ(k) < Φ(0)
3. The quantity Φ(0) is equal to energy for deterministic signals or the average power for random or periodic signals.

From these properties, any periodic signal attains its ACF maximas at periodically. Using this approach for practical time bound signal short time ACF may be computed as

$$R_n(k) = \sum_{n=-\infty}^{\infty} x(n)w(m-n)x(n+k)w(m-n-k)$$

Where w(m) is a window of finite length. If w(m) has segment length of N, for practical signal the equation may be rewritten as

$$R_n(k) = \sum_{n=0}^{N-1-k} x(n+m)w(n)x(n+m+k)w(n+k)$$

Experimentally it has been observed that for computing ACF rectangular window gives stronger indication of periodicity than others. Using rectangular window, where w(n) = 1 for 0 ≤ n ≤ N and w(n) = 0 otherwise, the short time ACF may be presented as

$$R_n(k) = \sum_{n=0}^{N-1-k} x(n+m)x(n+m+k)$$

Here, for this investigation, window segments lengths of 30ms are considered at 10ms interval with the overlapping period of 20ms. This process is repeated for 100 times considering each set of 30 ms of signal. Now mean and standard deviation of amplitude of autocorrelation peaks and repetition times are calculated.

One of the major limitations of ACF is that it retains too much information of signal. For processing of speech signal sometimes large peaks are observed due to damped oscillation of vocal tract response which may attribute erroneous computation largest peak. To avoid this problem speech signal may be pre-processed through clipper circuit. This will eliminate shorter signals and noise.

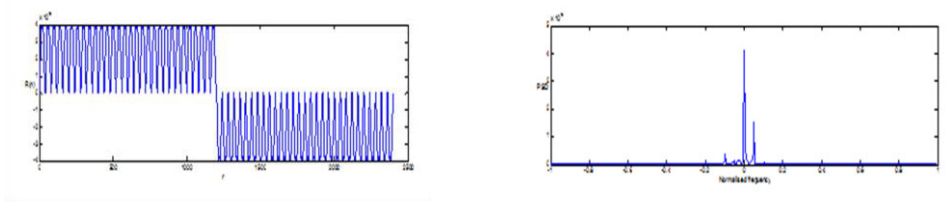
5. IMPLEMENTATION

Simple circuits of operational amplifiers with few resistors and capacitors are used in pre-amplifier section followed by a filter section. Filtered output is interfaced with microcontroller. Microcontroller type ATMEGA 8 is used that has the advantage of inbuilt analog to digital converter. Therefore use of external analog to digital converter is avoided, yielding circuit simplicity and reliability. Sampling frequency considered is 1 KHz. Microcontroller also performs all the computation for autocorrelation function and power spectral density. The schematic of experimental set up is presented at figure 1.

Figure 1: Schematic Diagram of Experimental Set up

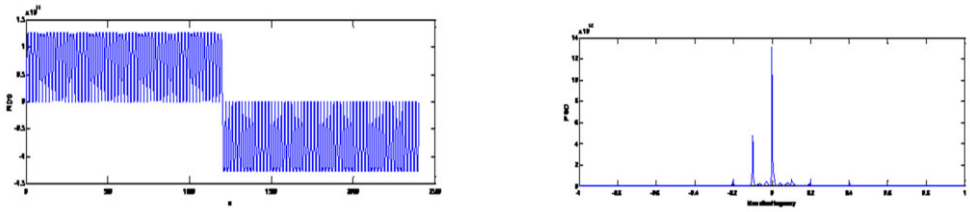
6. RESULT AND DATA ANALYSIS

Autocorrelation function describes complete features of voice signal even if is contaminated with noise and artifacts. Fundamental frequency can be extracted from this. Some of the laboratory analyses for autocorrelation as well as power spectral density are shown at figure 2 to 4. Above investigation of autocorrelation analysis has been applied for vocal fold feature extraction with 10 different people of different age group and gender. The result is promising as small deviation of fundamental frequency is detectable.



a: Autocorrelation b: Power Spectral Density

Figure 2: Autocorrelation function and Power Spectrum Density at 100Hz



a: Autocorrelation b: Power Spectral Density

Figure 3: Autocorrelation function and Power Spectrum Density at 200Hz

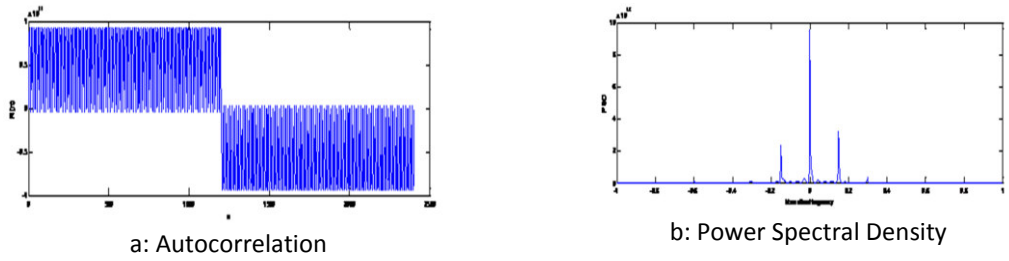


Figure 4: Autocorrelation function and Power Spectrum Density at 300Hz

7. DISCUSSION

A simple method of determining the pitch extraction has been demonstrated in this investigation. One important point of consideration is time duration of speech signal to be analyzed. In this investigation 100 segments each of 30 ms duration with the overlapping of 10 ms are considered which yielded satisfactory result. Another point is that the speech volume varies person to person as well as time to time and hence the amplitude of autocorrelation. However choice of clipper circuit threshold eliminates such problem. Clipper circuit output may be considered as:

For $x(n) \geq C_L$, $x(n) = +1$
For $x(n) \leq -C_L$, $x(n) = -1$
For $-C_L \leq x(n) \leq C_L$, $x(n) = 0$.

C_L is clipper threshold. Again clipper threshold is not unique and presetable for analysis of speech quality of all the persons.

8. CONCLUSION

The main motive of this research work is to develop a real-time pitch extraction system. There are two fold applications of this investigation. First it can automatically recognize the speech of enrolled speakers depending upon the vocal characteristics of the speakers and that may be applied for security requirement, as an essential element of the development of biometrics for speaker recognition.

Secondly, analysis of voice parameters may be applied for voice disorder analysis that may lead to clinical diagnosis and treatment of vocal folds disease conditions. Voice disorder is very common in professional speakers due to various reasons where vocal folds and larynx are affected. Investigation is being done in this direction with an objective to innovate non-invasive way of analysis of vocal folds disorder.

REFERENCES

[1] Titze IR (2006). Voice training and therapy with a semi-occluded vocal tract: rationale and scientific underpinnings. *J Speech Lang Hear Res*. 49(2):448-59.
[2] Koishi HU, Tsuji DH, Imamura R, Sennes LU (2003). Vocal intensity variation: a study of vocal folds vibration in humans with videokymography. *Rev Bras Otorrinolaringol*. 69(4):464-70.
[3] Elaine Lara Mendes Tavares, Roberto Badra de Labio, Regina Helena Garcia Martins (2010). Normative study of vocal acoustic parameters from children from 4 to 12 years of age without vocal symptoms. *Brazilian Journal of Otorhinolaryngology*. Print version ISSN 1808-8694. http://dx.doi.org/ 10.1590/ S1808-86942010000400013. 76(4):485-490.
[4} Brockmann-Bauser M, Drinnan MJ (2011). Routine acoustic voice analysis: time to think again? *Curr Opin Otolaryngol Head Neck Surg.* DOI:10.1097/MOO.0b013e32834575fe. 19(3): 165-70.
[5] Davis SB (1979). Acoustic characteristics of normal and pathological voices. *ASHA*.11:97-115.
[6] Cooley JW, Tukey JW (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation.* 19(90):297-301.
[7] Regina Aparecida Pimenta, María Eugenia Dájer, Adriana Hachiya, Domingos Hiroshi Tsuji, Arlindo Neto Montagnoli (2013). Parameters Acoustic and High-speed kymography identified effects of voiced vibration and vocal fry exercises. *CoDAS* 25(6):577-83.
[8] Samur (1975), Selection of Acoustic parameters for Speaker Identification, IEEE Transaction, Vol-ASSP-23.
[9] Rabiner (1977), Autocorrelation analysis for Pitch Estimation, IEEE Transaction, Vol-ASSP-23.
[10] Ross, Shaffer, Cohen, Freudberge and Manley (1974), Average magnitude difference function pitch extractor, IEEE Transaction, Vol-ASSP-22.
[11] Markel (1972), The SIFT Algorithm for Fundamental Frequency Estimation, IEEE Transaction, Vol-AU-20.

[12] BÖR Fritzell (1996). Voice disorders and occupation. *Logopedics Phoniatrics Vocology.* doi:10.3109/14015439609099197. 21(1):7-12.
[13] Lilia Brinca, Patrícia Nogueira, Ana Inês Tavares, Ana Paula Batista, Ilidio C Gonçalves, Maria L Moreno (2015). The prevalence of laryngeal pathologies in an academic population. *J Voice.* 29(1):130.e1-9.
[14] Amir K. Miri (2014). Mechanical Characterization of Vocal Fold Tissue: A Review Study. *J Voice.* DOI: http://dx.doi.org/10.1016/j.jvoice.2014.03.001. 28(6):657-667.
[15] Geraldo Pereira Jotz, Marco Antônio Stefani, Omero Pereira da Costa Filho, Tais Malysz, Paula Rigon Soster, Henrique Zaquia Leão (2014). A Morphometric Study of the Larynx. *J Voice.* DOI: 8. 28(6):668-67.

ABOUT THE AUTHORS

Dr. Karabi Ganguly Bhattacharjee is the Assistant Professor in the Department of Biomedical Engineering, JIS College of Engineering, Kalyani. She is having 14 years teaching experience. Karabi received her Ph.D degree from Jadavpur University, India and completed her Post Graduation and Graduation from the University of Calcutta. She has published many research papers in National / International Journals and Conference Proceedings. She has been invited as the Reviewer and Volume Editor in many International and National Journal and Conferences. Her research interest includes Engineering Physiology & Anatomy, Cellular Biochemistry and Clinical Oncology.

Nilotpal Manna obtained B.E. degree in Electronics and Telecommunication Engineering in 1979 from Bengal Engineering College, Sibpore, Kolkata, now renamed as Bengal Engineering and Science University and received M Tech degree in 1981from Indian Institute of Technology Madras (Chennai). He has wide industrial experience of twenty-two years from semi-government sectors like Instrumentation Ltd, Kota and several private industries like Toshniwal Instruments Manufacturing Pvt Ltd and others. He served mostly in the Research and Development wings and was associated in development of various electronic and communication instruments meant for military application as well as development of analytical instruments. At present he is Head of the Department of Electronics and Instrumentation Engineering of JIS College of Engineering, Kalyani, West Bengal, India. He has several research publications in national and international journals and conferences, and authored several technical books.