IJSTA
International Journal of Scientific and Technical Advancements
ISSN: 2454-1532

Effect of Linear Fitting Technique on Quality of Synthesized Speech

Akshay Kumar[1], Randhir Singh[1], Parveen Lehana[2]
[1]Sri SAI College of Engineering & Technology, Punjab, India
[2]Department of Physics & Electronics, University of Jammu, Jammu, India
Email address: [2]pklehanajournals@gmail.com

Abstract**—**The evolutionary origins of speech remain obscure. This contribution presents a novel approach to smooth the human speech using poly fitting techniques. The acquisition of speech signal is carried out using high quality system. Acoustic analysis of Hindi language female/male speaker is done via signal processing technique. An algorithm is designed to modify various speech parameters with different frame size. Various frame level ranging from 10 to 50 are taken and linear fitting of the data is carried out. Perceptual experiments were also conducted to assess the quality of the synthesized speech. Synthesized female speech showed low value of PESQ score with respect to male speech.

Keywords**—** PESQ; synthesis; linear; polyfitting.

I. Introduction

Speech is a human hallmark characteristic and its evolution is also enigmatic [1]. Speech is organized in series of open-close mouth cycles where the opened phase essentially corresponds to vowel production and the closed phase to consonant production. Characteristically, these cycles occur at 3–8 times per second, the motoric outcome of rapidly stringing together consonants and vowels to produce the syllables, words and sentences that comprise the world's spoken languages [2], [3]. The anatomy human vocal tract is shown in figure 1. The vocal tract is the space ranging from the opening between the vocal cords (glottis) to the lips. The vocal tract can be divided into the pharynx (from the esophagus to the mouth) and the oral cavity. For an average male, the length of the vocal tract is about 17 cm, and the cross-sectional area varies from zero to about 20 cm^2 (depending of the positions of the tongue, lips, jaw, and velum) [4]. The velum is also used to control the acoustical coupling of the nasal tract to the vocal tract. The sub-glottal system composed of the lungs, bronchi and trachea serves as

a source of energy for the production of speech. The sounds in speech can be divided into three categories according to their mode of excitation: Voiced, unvoiced, and plosive sounds. Voiced sounds are produced by blowing air through the glottis while the tension of the vocal cords is adjusted so that they vibrate in a relaxation oscillation. This produces quasi-periodic pulses of air flow which excite the vocal tract. Unvoiced sounds are generated by causing a constriction somewhere along the vocal tract and forcing air through it at a velocity that produces turbulence [5]. This produces a noise source to excite the vocal tract. Plosive sounds are generated by making a complete closure somewhere along the vocal tract, and building up pressure behind it. When the closure is opened, the pressure is released as a burst of air- flow excites the vocal tract [6].

The vocal tract resonances or formants provide both phonetic information i.e. identity of the intended vowel or consonant and source information. The frequencies of the lowest three formants, as well as their pattern of change over time, provide cues that help listeners ascertain the phonetic identities of vowels and consonants. Vocalic contrasts, in particular, are determined primarily by differences in the formant pattern [7-9]. The formant representation provides a compact description of the speech spectrum. Given an initial set of assumptions about the glottal source and a specification of the damping within the supralaryngeal vocal tract, the spectrum envelope can be predicted from knowledge of the formant frequencies [10]. A change in formant frequency leads to correlated changes throughout the spectrum, yet listeners attend primarily to the spectral peaks in order to distinguish among different vocalic qualities [11], [12].

In this research work formant smoothening is carried out using linear fitting of signal. The quality of speech after is estimated using Perceptual evaluation of speech quality (PESQ) method [13] is used to evaluate quality of the speech.

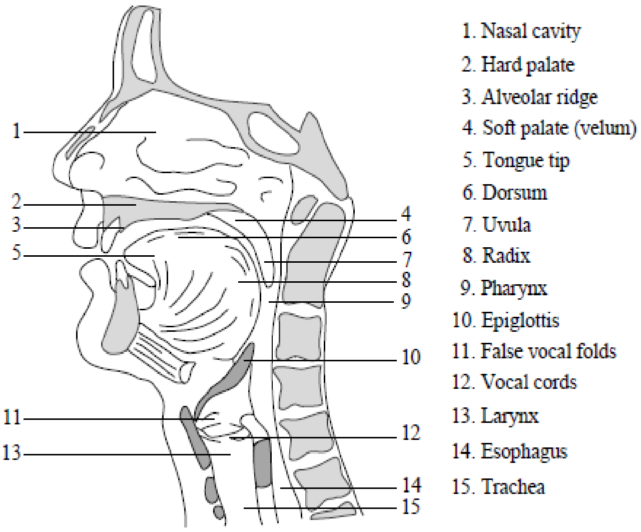

Fig. 1. The human speech production system.

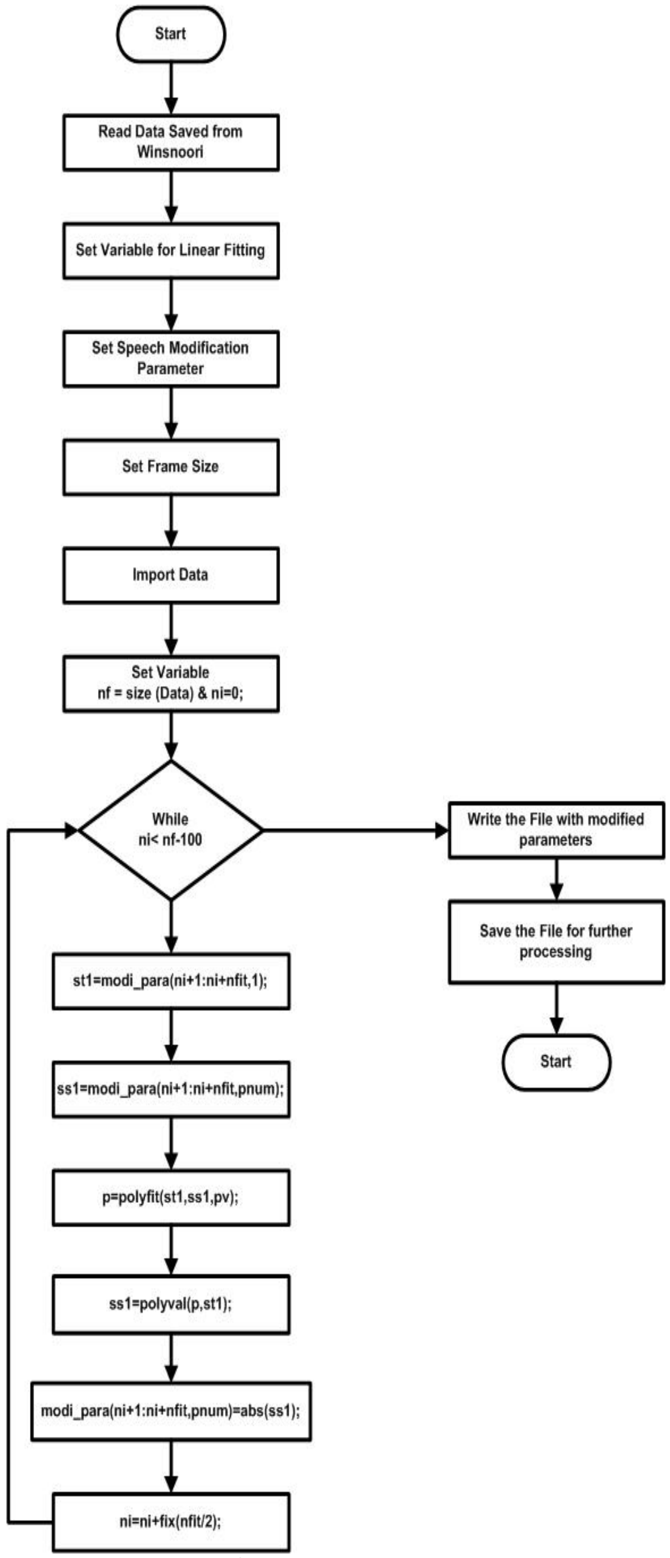
Start
Read Data Saved from Winsnoori
Set Variable for Linear Fitting
Set Speech Modification Parameter
Set Frame Size
Import Data
Set Variable nf = size (Data) & ni=0;
While ni< nf-100
Write the File with modified parameters
Save the File for further processing
Start
st1=modi_para(ni+1:ni+nfit,1);
ss1=modi_para(ni+1:ni+nfit,pnum);
p=polyfit(st1,ss1,pv);
ss1=polyval(p,st1);
modi_para(ni+1:ni+nfit,pnum)=abs(ss1);
ni=ni+fix(nfit/2);

Fig. 2. Designed algorithm for reconstructing speech with added noise.

II. Methodology

The research work is divided into two main sections. In first section speaker selection, speech recording and segmentation is done, while in the second part analysis of smoothened speech quality is carried out. Phrase in Hindi language are recorded using Audacity software at the sampling rate of 16000 Hz. The speech of two female and two male speakers were recorded in an acoustically treated environment. An algorithm shown in figure 2 is designed to smoothen the recorded signals using linear fitting techniques. Also various speech parameters such as pitch, Amplitude, formant frequency, bandwidth etc were modified in the algorithm. Various frame size ranging from 10 to 50 were taken to smooth the speech. Reconstructed speech is compared with original recorded speech signal. The deviation between synthesized speech and original speech is evaluated using PESQ method.

III. Results and Discussion

Normalized female speech and their spectrogram are shown in figure 3 and figure 4 whereas figure 5 and figure 6 shows normalized male speech and their spectrogram. Table I shows the computed PESQ score of synthesized smoothen speech having various frame size in linear fitting process with respect to original recorded speech signal of female and male speakers. Figure 7 shows the PESQ core of female and male speakers. The x-axis shows the frame size and y-axis PESQ scores. PESQ histogram result it is concluded that for female speaker as the frame size is increased PESQ attains a constant value and the maximum value for Sp1 and Sp2 is 1.425 and 1.080 respectively. It is contrary for male speaker, as the frame size is increased PESQ score shows the random deviation and the maximum value for male speaker Sp3 and Sp4 are 1.439 and 1.413 respectively.

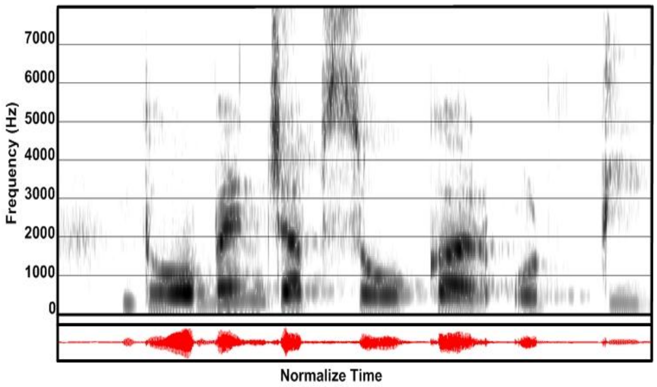

Fig. 3. Female speaker (Sp1) normalized speech signal and spectrogram.

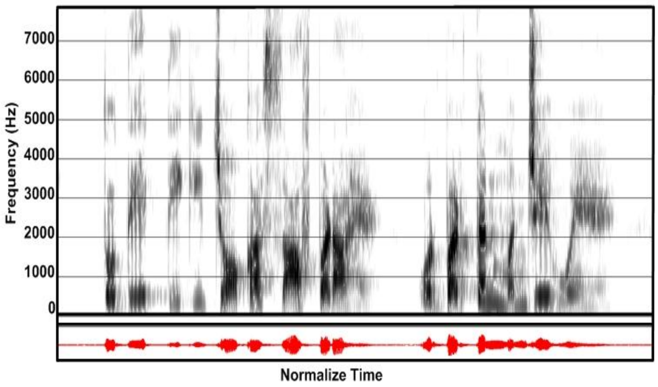

Fig. 4. Female speaker (Sp2) normalized speech signal and spectrogram.

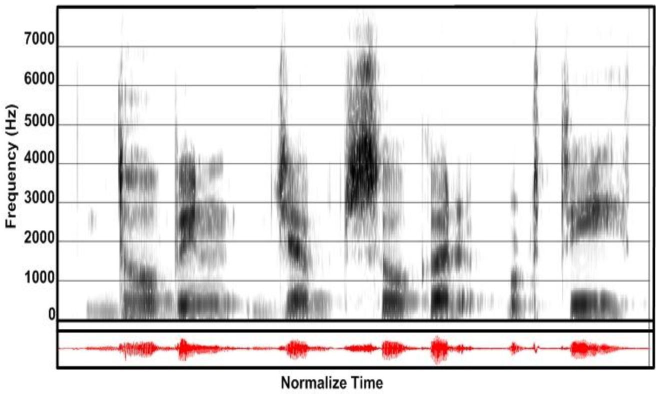

Fig. 5. Male speaker (Sp3) normalized speech signal and spectrogram.

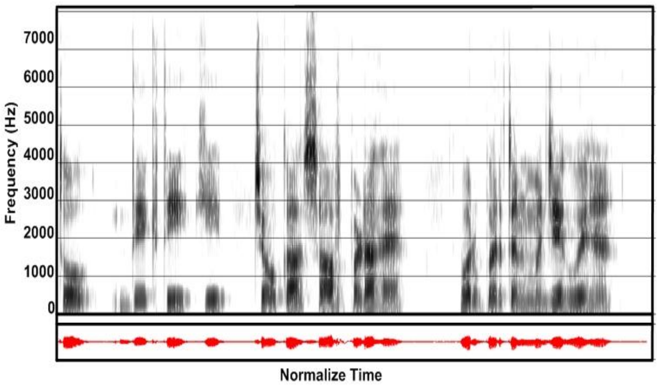

Fig. 6. Male speaker (Sp4) normalized speech signal and spectrogram.

Table I PESQ score of female and male speaker with various frame size.

Frame Size	PESQ Score			
	Female Speaker		Male Speaker	
	Sp1	Sp2	Sp3	Sp4
10	1.28	1.00	1.34	1.07
20	1.42	1.08	1.32	1.23
30	1.42	1.08	1.43	1.20
40	1.42	1.08	1.34	1.16
50	1.42	1.08	1.43	1.41

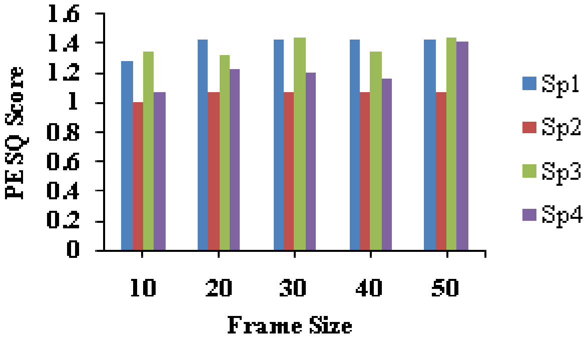

Fig. 7. PESQ score of female and male speaker.

IV. Conclusion

Research work is carried out to investigate the linear fitting of signals on male/female Hindi language speaker. A phrase is recorded using high quality system and professional sound recording software. An algorithm is designed to linear fit the data using poly fitting technique. PESQ score of 1.280 and 1.008 is obtained for frame size 10 and 1.425 and 1.080 at frame size 50 for female speakers. Similarly for male speaker the PESQ score comes out to be 1.349 and 1.075 at 10 and 1.439 and 1.413 at 50 frame size respectively.

References

[1] A. R. Lameira, M. E. Hardus, A. M. Bartlett, R. W. Shumaker, S. A. Wich1, and S. B. J. Menken, Speech-Like Rhythm in a Voiced and Voiceless Orangutan Call. New York: Cambridge University Press, 1984.

[2] A. A. Ghazanfar and D. Y. Takahashi, "Facial expressions and the evolution of the speech rhythm," *J Cogn Neurosci*, vol. 26, pp.1196–1207, 2014.

[3] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," *in Proc. IEEE Acoustics, Speech and Signal Processing*, pp. 4609-4612, 2008.

[4] L. Rabiner and R. Schafer, Digital Processing of Speech Signals, Prentice Hall, 1978.

[5] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, and Language Processing*, pp. 153-165, 2011.

[6] G. E. Peterson and H. L. Barney, "Control methods used in a study of vowels," *J Acoust Soc Am*, vol. 24, pp.175–184, 1952.

[7] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *J Acoust Soc Am*, vol. 85, pp. 2088–2113, 1989.

[8] P. F. Assmann and W. F. Katz, "Time-varying spectral change in the vowels of children and adults," *J Acoust Soc Am*, Vol. 108, pp.1856–1866, 2000.

[9] G. Fant, Acoustic Theory of Speech Production. Mouton: The Hague, 1960.

[10] R. Carlson, B. Granstrom, and D. Klatt, "Vowel perception: the relative perceptual salience of selected acoustic manipulations," Speech Transmission Laboratories Quarterly Progress Report, pp. 73–83, 1979.

[11] C. J. Darwin, "Perceiving vowels in the presence of another sound: constraints on formant perception," *J Acoust Soc Am*, vol. 76, pp. 1636–1647, 1984.

[12] M. Sommers and P. D. Kewley, "Modeling formant frequency discrimination of female vowels," *J Acoust Soc Am*, vol. 99, pp. 3770–3781, 1996.

[13] S. Pennock, "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm," *in Proc. Measurement of Speech and Audio Quality in Networks Line Workshop*, 2002.