International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 1077-1084

PARAMETRIC SPEECH SYNTHESIS BY SOURCE-FILTER MODEL

[1]P. SHERLY ARUNODHAYAMARY, [2]S. SELVA NIDHYANANTHAN,
[1] PG Scholar, *Dept. of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Tamil Nadu, India*
[2] Associate Professor, *Dept. of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Tamil Nadu, India*

Abstract

Speech synthesis is a notable research field in speech processing which provides smart phone applications and also supports blind people. Process of obtaining natural speech from gadgets are gaining lots of attention due to their popularity in the market. But the major drawback is the voice quality. In this work, levels of synthesis are performed and their performance at various levels is quoted to predict the better performance among the levels of synthesis. The digits and vowels are data considered for work. The first part of synthesis is the process of generating voice from text in case 1. Recorded database is used in case 2. A level is increased with a module of source-filter synthesis. The performance of the module is evaluated and compared between the present results from case 1, case 2. To evaluate performance of the levels, measure have been taken by Mean Square Error, Peak Signal-to-Noise Ratio

Keywords: *Speech synthesis, text to speech converter, source-filter model, performance evaluation, subjective and objective evaluation method.*

1. Introduction

Generation of speech by machine with human voice characteristics are known as speech synthesis. In synthesis, the machine reads out loud in real through loud speaker. It is also defined as artificial human speech production. Here, screen reader systems integrate with synthesis module to generate user satisfying speech system to aid blind and partially-sighted people. This kind of approach makes use of unit-selection speech synthesis [1]. As of in a growing efficient world, the impact of computers is of great deal. The notable part is the human computer interaction by speech recognition and speech synthesis. In this paper, speech synthesis is given high focus. Main idea is to obtain synthesized speech that is easily understandable and indistinguishable from the human speech. The rule-based synthesis such as formant synthesis and articulatory synthesis has advantage over other such that it requires less memory and low processing cost. Formant synthesis can also be called as source-filter synthesis unlike the other, they are kind of music synthesizer. But it has a difficulty of finding parameters from the input, they can be caste out by concatenative synthesis, the best among the concatenative synthesis are the diphone synthesis. Unit Selection synthesis are also called as corpus based concatenative synthesis. It uses large set of recorded units of multiple instance of varying prosodies. Main goal of synthesis is to minimize cost function [2]. Production of continuous speech in artificial manner is quite difficult considering both the concept of synthesizer and dictionary. Speech signals are categorized as voiced and unvoiced portions. Voiced signal contains fundamental frequency (*f0*), formant (*f1, f2, f3),* harmonic components, whereas in unvoiced signal characterized by silence and unstable voice. Vibration of the vocal cord modifies the excitation signal producing formant frequency (poles) and anti-formant frequency (zeros) [3]. The design of complete source-filter model shows that the source is independent of filter. Glottal flow model is used to design glottal source which comprise the level of loudness and structure. The models are Rosenberg, Liljencrants-Fant, Rosenberg++ and so on. Voiced speech is produced by filtering from the glottal source of the excitation signal with a formant signal [4]. The excitation part of speech can actually be defined as the rapid opening and closing of the glottis which determines the characteristics of glottal

ISSN: 2005-4238 IJAST

waveform. The pulse takes a sinusoidal waveform, denoting the opening time T_p (positive slope). When there is inverse in this case, the slope of the closure increase denoting the closing time T_N (closing slope) [5]. The combination of these pulses unites to form a voiced part of the source-filter model. The pitch estimation or fundamental frequency estimation of speech signal can be estimated in time and frequency domain. Estimation by autocorrelation method in time domain is often preferred due to their simplicity. In frequency domain, PEFAC algorithm is used for high level of noise [6]. The next important calculation comes in f0 estimation. The speech data can be analyzed depending on their pitch and fundamental frequency [7].

2. PROPOSED METHODOLOGY

Stages of synthesis are performed to synthesize and enhance the voice quality. The work is performed in two cases, case (i) Text input is converted to voice. Converted voice data are synthesized by Source-Filter model. In case (i), Text conversion is done by using python language. Python module used here takes in text and converts it to mp3 voice data. The data obtained are vowels and digits. The major advantage of python module is, it provides a better way for storage of audio file generated. The saved audio file can be promoted for further usage. In an audio data, Higher the sampling rate, higher the quality of audio will be. Sampling rate of voice generated by python library are of 24KHz. Voice generated in artificial manner based on their parameter varies from the natural generated speech data. Case (ii) Obtaining synthesized speech from recorded natural speech using source filter model. Recorded speech data consists of vowels and digits, recording is done in a closed glass room and noise free environment. Recording is done by Female speakers for both vowels and digit which acts as input of case (ii) in proposed methodology.

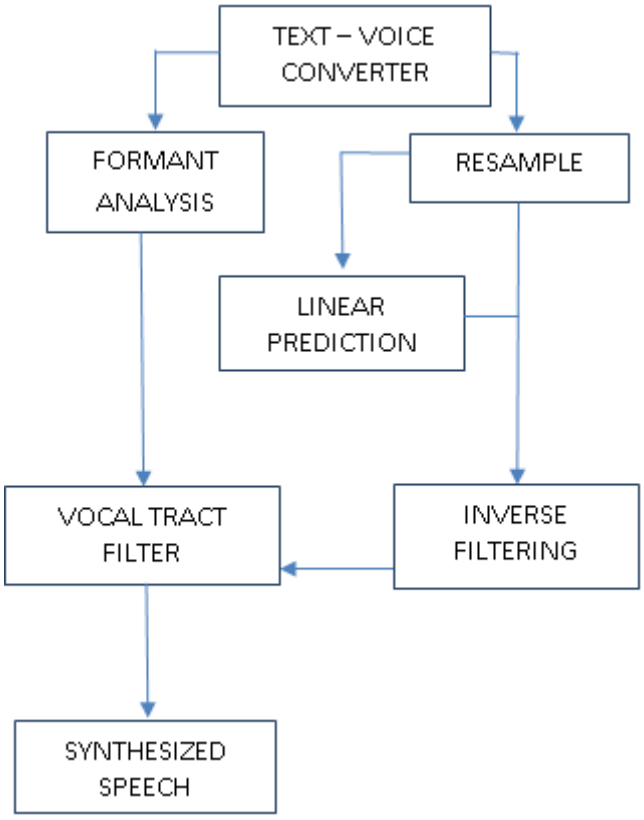

Figure. 1: Case (i). Design of Source-Filter model with text converter

Next, values of the first module from both the cases are fed to the Source-Filter model for synthesis. Source-Filter model can also be called as formant synthesis. In short Formants are removed by filtering and source of speech signal is obtained by inverse filtering. The operation starts with filtering of the formant to extract the residue which acts as the source in the synthesis. This is obtained by inverse filtering of the speech signal by Linear Prediction. The LPC concentrates on the intensity and frequency of speech data rather than the important formant information part of the speech.

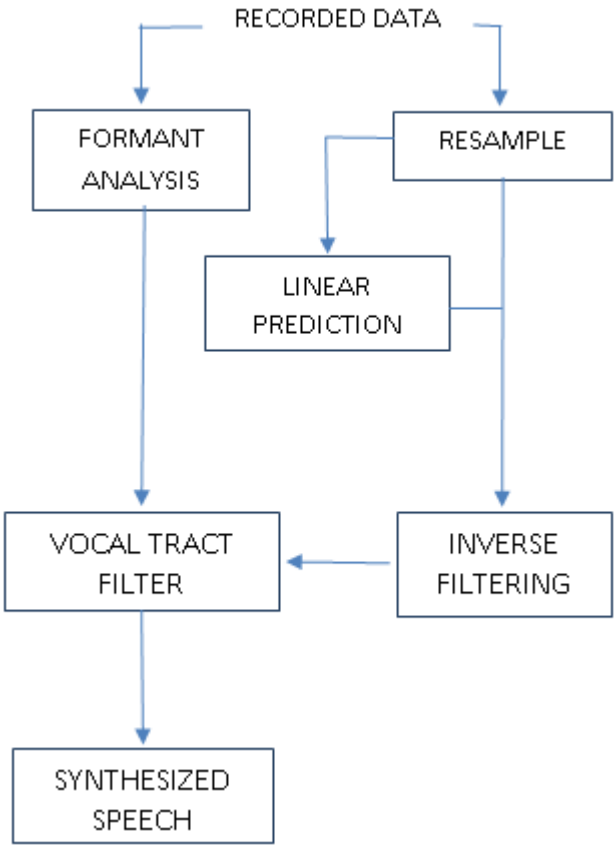

Figure.2: Case (ii). Design of Source-Filter model with recorded data

Our focus is on Burg method, this method is equivalent to the estimation of the LPC coefficients. Like in LPC, speech signals are initially resampled to the desired sampling rate. After resampling, speech signal undergoes pre-emphasis, windowing by hamming window and framing for 20msec. LP filter of order 10 is used. Higher order of filter will give higher level of distortion in synthesis. From this calculation 4 formants can be obtained from each frame. Real valued LPC coefficients have roots in complex conjugate pairs and only the roots with positive imaginary parts are considered. r_k denote the roots of the linear prediction polynomial.

$$angz = \Sigma \tan^{-1}(\mathrm{Im}\, r_k / \mathrm{Re}\, r_k) \quad (1)$$

The formant bandwidth pairs are estimated from the below formant and bandwidth equations.

$$F_K = \left(\frac{F_s}{2\pi}\right) \tan^{-1}(\mathrm{Im}\, r_k / \mathrm{Re}\, r_k) \quad (2)$$

$$B_K = - \left(\frac{-F_s}{2\pi}\right) ln\, |r_k| \quad (3)$$

Formants *f1, f2, f3, f4* from every frame are calculated. Since the burg method evaluates the formants, inverse filtering at this stage generates the residue of the signal Residue of an audio signal contains parameters other than the essential formants of the signal. Now, formants from burg method and residue from inverse filtering combine to form the synthesized speech. Speech parameters like pitch and fundamental frequency from case (i) and case (ii) are considered to understand their performance. The changes in the parameters are noted after passing through the source-filter model. Since formants of *f1, f2, f3, f4* is considered for female speaker frequency below 50Hz and above 5000 Hz are avoided in Pre-emphasis. Also, to avoid the unwanted increase in the speech signal due to the Octave Band filtering. Major advantage of Formant estimation over the residue calculation by the LPC are, the independencies for the maximum frequency consideration for the formant estimation unlike the dependencies on the Nyquist frequency for the maximum frequency range by LPC analysis window. Results of every module is discussed.

3. RESULTS AND DISCUSSION

Result of each module is given with their essential characteristic parameter. Initially a single person audio data is analyzed and finally an average data obtained from all the modules of the project is quoted in Table 1,2,3,4. Input speech represent the voice from converter module in Table 1, 2. Input speech in Table 3, 4 represent the Recorded voice of the following tabulation. The values between the case (i) and case (ii) clearly shows the clarity of the speech data is high in recorded voice rather than the computer-generated voice. The expected range of pitch for female speaker is160-350 Hz. As there are few fricatives presents in the digits the range of pitch is low compared to the voiced vowels in Table 1, 2. These data are stored for further processing by speech synthesis. Next case is the process of speech synthesis by Source-Filter model. These data are fed to the Praat software for speech synthesis by source-filter model. The output of the source filter model obtained from proposed methodology clearly shows the variations between case (i) and case (ii) from the table 1, 2, 3, 4 in the synthesized part. In the process of Pre-emphasis, the frequency content of the data tends to increase. But, the data below and above a certain level of frequency continuous to degrade in their function this is due to the selection of frequency levels by the filter between 50Hz and 5000 Hz of range. The average results of 5 speaker are given so far to analyze and understand modules of the proposed methodology. These tables give the average value of Digit and Vowel data for five persons pitch and pitch period for the text to speech converted voice and the synthesized speech of the source-filter model. Data from voice converted system case (i) continue to show degraded performance compared to the case (ii) approach from the recorded data. Since, data obtained from artificial manner does not satisfy the basic requirement of speech information it continues to degrade on further processing. Whereas upon synthesis in case (ii) it faces some reduction in levels of synthesis but stands tall to support the fortunate requirements of basic synthesis needs.

4. PERFORMANCE MEASURE

To evaluate the quality of speech obtained from the synthesis process. Performance measure can be performed by subjective evaluation and objective evaluation. Subjective evaluation of data depends on the individual marking for the audio file. Objective evaluation depends on the mathematical evaluation criteria. Listening are made in the clean environment. Listeners are given instruction to award the intelligibility and quality of the speech by comparing to the original audio file. The ratings are between 1-5. Range of 1 is unacceptable and worse whereas the range of 5 stands for imperceptible and clear.

ISSN: 2005-4238 IJAST

4.1 MEAN OPINION SCORE

MOS is calculated as arithmetic mean over the rating performed by human subjects for a given audio. This comes under subjective evaluation method.

$$\text{MOS} = \frac{\sum_{n=1}^{N} R_n}{N} \quad (4)$$

N, the total number of elements. R_n, the score of the listeners.

4.2 PEAK SIGNAL TO NOISE RATIO

Defined as the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of the representation. Maximum power of a signal is *m*.

$$\text{PSNR} = 10*\log_{10}\left(\frac{(m)^2}{mse}\right) \quad (5)$$

4.3 MEAN SQUARE ERROR

Mean squared error or mean squared deviation of an estimator measures the average squared distance between the estimated value and original value. When values of MSE increases it represents that the data values are dispersed widely, and smaller MSE is preferred and shows that the data values are dispersed closely MSE is always non-negative and close to zero.

$$\text{MSE} = \sum_{M,N} \frac{(y-\hat{y})^2}{M \cdot N} \quad (6)$$

M, *N* represent the rows and columns of data. *y* is the clean signal and$\hat{y}$ is the synthesized speech signal.

Table. 1 Average Digit data - case (i) of proposed methodology

DIGIT DATA	PITCH		PITCH PERIOD	
	INPUT SPEECH (Hz)	SYNTHESIZED SPEECH (Hz)	INPUT SPEECH (ms)	SYNTHESIZED SPEECH (ms)
Person_1	54	31	114	56
Person_2	105	54	66	96
Person_3	100	99	28	16
Person_4	93	71	73	82
Person_5	94	77	91	56

Table. 2 Average Vowel data - case (i) of proposed methodology

VOWEL DATA	PITCH		PITCH PERIOD	
	INPUT SPEECH (Hz)	SYNTHESIZED SPEECH (Hz)	INPUT SPEECH (ms)	SYNTHESIZED SPEECH (ms)
Person_1	60	27	11	83
Person_2	120	129	35	29
Person_3	121	102	16	22
Person_4	95	111	49	19
Person_5	107	130	29	20

ISSN: 2005-4238 IJAST

Table. 3 Average Digit data - case (ii) of proposed methodology

DIGIT DATA	PITCH		PITCH PERIOD	
	INPUT SPEECH (Hz)	SYNTHESIZED SPEECH (Hz)	INPUT SPEECH (ms)	SYNTHESIZED SPEECH (ms)
Person_1	179	200	64	64
Person_2	164	169	20	19
Person_3	151	171	20	11
Person_4	176	199	15	11
Person_5	249	199	39	15

Table. 4. Average Vowel data - case (ii) of proposed methodology

VOWEL DATA	PITCH		PITCH PERIOD	
	INPUT SPEECH(Hz)	SYNTHESIZED SPEECH (Hz)	INPUT SPEECH (ms)	SYNTHESIZED SPEECH (ms)
Person_1	207	188	10	11
Person_2	204	102	5	41
Person_3	194	169	9	40
Person_4	181	142	10	40
Person_5	231	151	11	15

Table. 5 Subjective evaluation of speech from case (i) and case (ii) – Digit, Vowel

CASE	DATA TYPE	RANKING OF LISTENER					MOS
		1	2	3	4	5	
Case(i)	DIGIT DATA	3	2.9	3.1	2.9	3.2	3.02
	VOWEL DATA	3.1	3.2	2.9	2.9	2.9	3
Case(ii)	DIGIT DATA	3.6	3.5	3.4	3.5	3.4	3.4
	VOWEL DATA	3.3	3.7	3.5	3.3	3.4	3.4

ISSN: 2005-4238 IJAST

Table. 6. Performance of speech case (i) – Digit, Vowel

Case (i)	DIGIT DATA		VOWEL DATA	
	PSNR	MSE	PSNR	MSE
Person_1	4.804	0.061	3.522	0.103
Person_2	9.34	0.121	8.50	0.143
Person_3	9.464	0.13	7.11	0.208
Person_4	8.665	0.15	6.25	0.25
Person_5	7.934	0.173	8.42	0.164

Table. 7. Performance of speech case (ii) – Digit, Vowel

Case(ii)	DIGIT DATA		VOWEL DATA	
	PSNR	MSE	PSNR	MSE
Person_1	15.39	0.018	10.72	0.11
Person_2	8.99	0.020	9.34	0.11
Person_3	7.09	0.026	10.14	0.16
Person_4	8.50	0.009	8.42	0.15
Person_5	9.22	0.129	10.15	0.10

5. CONCLUSION

The work shows the difference between the computer synthesized speech and recorded natural voice. This work also depicts the working of source-filter model using Praat software in the case of both artificial sound and natural sound. The synthesis obtained from natural speech stands high compared to the results of artificially obtained data. The synthesized data obtained from both the cases of speech are analyzed by the performance measure like subjective evaluation. Objective evaluation by Peak signal-to-noise ratio and mean square error.

REFERENCES:

1. AimiliosChalamandaris; Sotiris Karabetsos; PirrosTsiakoulis; Spyros Raptis(2010), 'A unit selection text-to-speech synthesis system optimized for use with screen readers', IEEE Transaction on Consumer Electronics, vol:56,issue:3
2. Youcef TABET, Mohamed BOUGHAZI (2011),'Speech Synthesis techniques. a survey' international workshop on systems, signal processing and their application (WOSSPA) 2011.
3. Nwakanma Ifeanyi, AluigboLkenna and OkpalaIzunna, (2014),"Text-To-Speech synthesis", International Journal of Research in Information Technology (IJRIT),Volume 2,Issue 5, May 2014, pp: 154-163.
4. Q.Fu, P.Murphy, (2006)"Robust glottal source estimation based on joint source-Filter model optimization", IEEE Transaction on Audio, Speech and Language Processing, vol:14, issue:2, pp: 492-501.
5. A.E.Rosenberg ,(2005), "effect of glottal pulse shape on the quality of natural vowel", The Journal of the acoustical Society of America, vol: 49, issue: 2B

6. Geliang Zhang, Simon Godsill, (2016),'fundamental frequency estimation in speech signal with variable rate particle filters', IEEE/ACM Transaction on Audio Speech and Language processing, vol:24, issue: 5, pp: 890-900.
7. P. Jayawardhana, A. Aponso, N. Krishnarajah and A. Rathnayake, "An Intelligent Approach of Text-To-Speech Synthesizers for English and Sinhala Languages," 2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT), Kahului, HI, USA, 2019, pp. 229-234.doi: 10.1109/INFOCT.2019.8711051