2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017)

Text to Speech Synthesizer-Formant Synthesis

Sneha Lukose
Electronics and Telecommunication Dept
Fr.C.R.I.T
Mumbai,India
snehalukose1992@gmail.com

Savitha S. Upadhya
Electronics and Telecommunication Dept
Fr.C.R.I.T
Mumbai,India
savivashi@gmail.com

***Abstract*—In this paper, different methods of text to speech synthesizer techniques are discussed to produce intelligible and natural output and a vowel synthesizer using cascade formant technique is implemented. A text to speech output is based on generating corresponding sound output when the text is inputted. Wide range of applications use text to speech technique in medicals, telecommunications fields, etc. The Various speech synthesis methods that have been used for text to speech output for obtaining intelligible and natural output are Concatenative, Formant, Articulatory, Hidden Markov model (HMM).**

Keywords—Concatenative, Unit-Selection, Diaphone, Domain specific, Formant, Cascade and Parallel synthesizer, Articulatory and Hidden Markov model.

I. INTRODUCTION

Each spoken word is created from the phonetic combination of a set of vowel and consonant speech sound units. Producing an artificial human speech is known as speech synthesis.

The various speech synthesis methods used for the text to speech system are Concatenative synthesis, Formant synthesis, Articulatory synthesis and Hidden Markov model(HMM). In literature, Arafat et. al. [2] presents a grapheme to phoneme conversion followed by concatenating synthesis. Joshi, Chabbi and Kulkarni [3] investigates on Baraha software for conversion of text into phonetic sounds. Ouh-young et. al. [4] presents an approach of text written in Mandarin Phonetic Symbols II using concatenation to produce speech output. Akinwonmi et. al. [5] discusses on the HMM method by using polysyllabic and diphone units. Buza et. al. [6] explain the rule based approach of syllable concatenation synthesis. Rasekh et. al. [7] contributed on syllable-phoneme approach for reduced database and got better results. Partha Shankar Nayak [8] deals with letter to sound rules to minimize the size of audio database for an efficient web page reader.

II. TYPES OF SPEECH SYNTHESIS METHODS

In this section, types of speech synthesis methods to obtain better text to speech is presented. When both intelligibility and naturalness characterizes maximum to generate artificial speech sounds, the system likely becomes effective. The Concatenative, Formant, Articulatory and Hidden markov model are the various speech synthesis methods.

A. Concatenative Synthesis

In concatenative synthesis, the spoken sentence is broken down into words and words into syllables, demisyllables, phonemes, diaphones or triphones. Then concatenation and rearrangement of the above segments of recorded samples is done to create new words and sentences is known as concatenative synthesis. This type of synthesis provides the maximum amount of naturalness and intelligibility. It is widely used in systems where limited things are said. There are several problems encountered in concatenative synthesis such as distortions from discontinuities, high memory requirement and time consuming [1].

A.1 Unit Selection Synthesis

The unit selection synthesis uses large database of recorded speech. It provides the maximum amount of naturalness to the output which is similar to real human voices.

A.2 Diaphone Synthesis

The diaphone synthesis uses smaller database than unit selection. It provides robotic voice output than natural sound. This system creates glitches due to the concatenation of very small units. Softwares that allow free speech synthesis use diaphone synthesis.

A.3 Domain Specific Synthesis

The domain specific synthesis is used for limited output vocabulary. Such systems are widely used in a clock, medical call center, a weather report, a railway announcement, etc.

B. Formant Synthesis

The formant synthesis is based on the source-filter model. There are two types of formant synthesis structures namely Cascade and Parallel structures. A combination of the two structures also gives better performance. Formant synthesis produces infinite number of sounds than compared to concatenative synthesis. To produce intelligible speech upto three to five formants are generally required [1]. A two pole bandpass resonator is used to model each formants by passing the corresponding formant frequency and its bandwidth. The cascade structure consists of two pole resonators connected in series and for the parallel structure it is connected in parallel. A cascade structure is simpler to implement than the parallel structure. The output speech produced with the formant

978-1-5090-2794-1/17/$31.00 ©2017 IEEE

synthesis is more likely to generate robotic and unnatural voice.

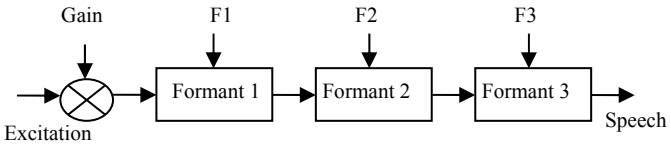

Fig.1 Cascade formant structure [1].

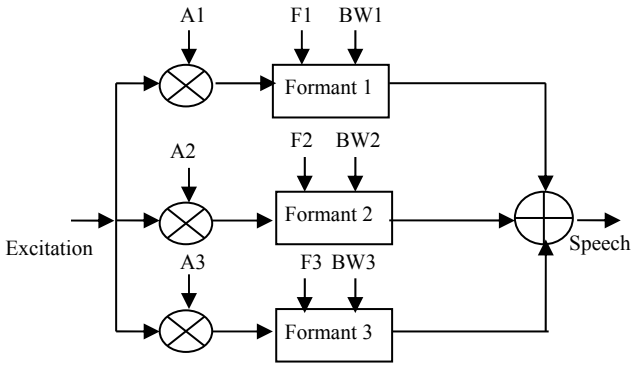

Fig.2 Parallel formant structure [1].

C. *Articulatory Synthesis*

Articulatory synthesis is based on the modeling of the human speech production system and is not commercially used as it is difficult to implement.

D. *Hidden Markov Model*

Hidden Markov model is also known as a statistical parametric synthesis. If the state sequence cannot be determined from the signal sequence, the model is said to be hidden. In any speech recognition system, the acoustic signals can only be observed and not the phonemes which are the hidden states. It is not completely automatic but requires manual help.

III. IMPLEMENTION AND RESULTS

Here, a text to speech technique is implemented using cascade formant structure for the vowels /a/, /e/, /i/, /o/ and /u/. According to [8], parameter values used for the synthesis of selected vowels as shown in Table I helps in generating the required output. The spectrogram was also plotted for verifying the formant frequency. The cascade connection has been used for the synthesis of the vowels as it has an advantage to produce an accurate model of the vocal tract transfer function during the production of non-nasal sonorants [8]. Also the output of each resonator that requires a formant frequency F and bandwidth BW come out just right without any amplitude control for each formant. The samples of the output of a digital resonator $y(nT)$ are computed from the input sequence $x(nT)$ as in (1).The sampling frequency used here was 8000Hz with sampling time T taken as the inverse of sampling frequency and using the constants A, B and C that were related to each resonator by the impulse–invariant transformation, the transfer function $T(f)$ of the digital resonator is obtained.

$$y(nT) = Ax(nT) + By(nT - T) + Cy(nT - 2T) \quad (1)$$

$$C = -\exp(-2\pi BWT) \quad (2)$$

$$B = 2\exp(-\pi BWT)\cos(2\pi FT) \quad (3)$$

$$A = 1 - B - C \quad (4)$$

$$T(f) = \frac{A}{1 - Bz^{-1} - Cz^{-2}} \quad (5)$$

TABLE I. PARAMETERS OF SELECTED VOWELS [8]

Vowels	Parameters(Hz)					
	F1	***F2***	***F3***	***BW1***	***BW2***	***BW3***
/a/	700	1220	2600	130	70	160
/e/	480	1720	2520	70	100	200
/i/	310	2020	2960	45	200	400
/o/	540	1100	2300	82	100	82
/u/	350	1250	2200	65	110	140

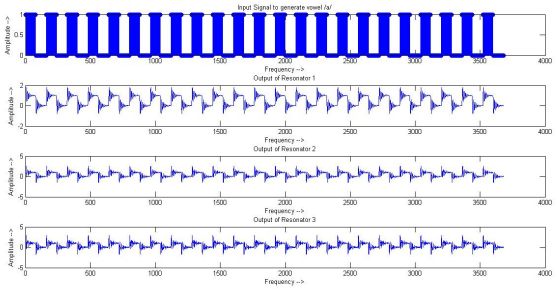

Fig.3(a) Time domain representation of Cascade formant output for Vowel /a/.

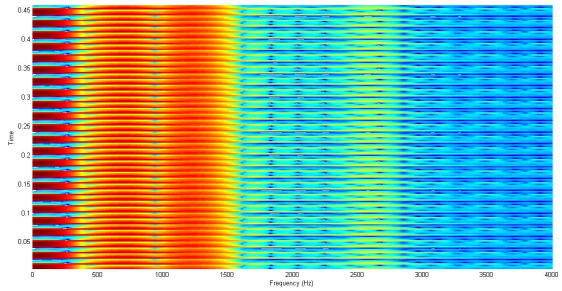

Fig.3(b) Spectrogram output for Vowel /a/.

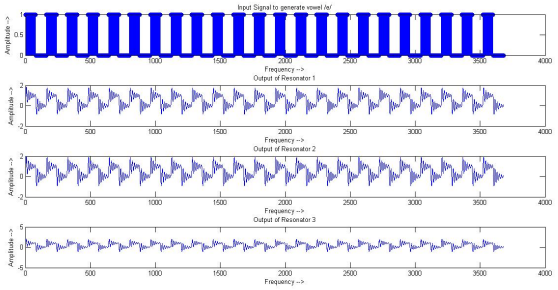

Fig.4(a) Time domain representation of Cascade formant output for Vowel /e/.

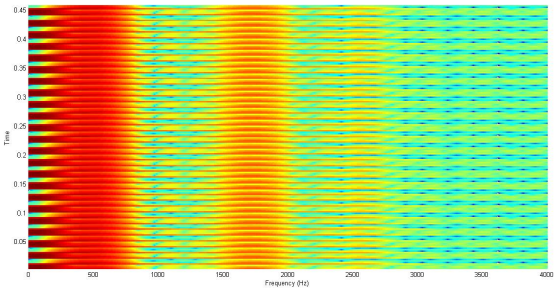

Fig.4(b) Spectrogram output for Vowel /e/.

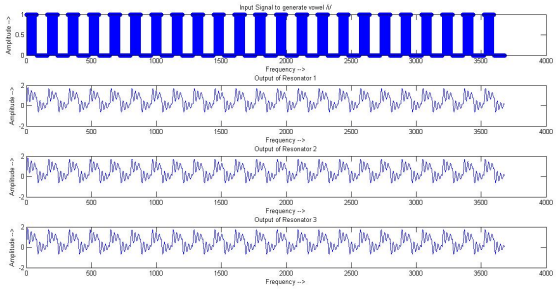

Fig.5(a) Time domain representation of Cascade formant output for Vowel /i/.

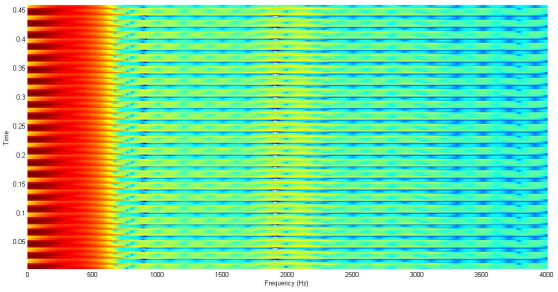

Fig.5(b) Spectrogram output for Vowel /i/.

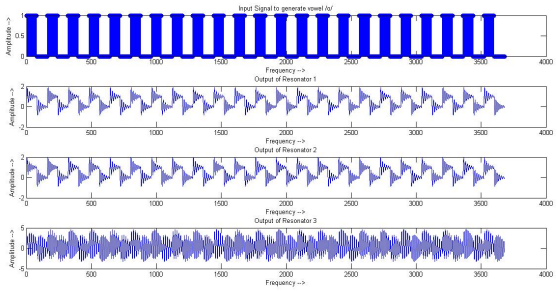

Fig.6(a) Time domain representation of Cascade formant output for Vowel /o/.

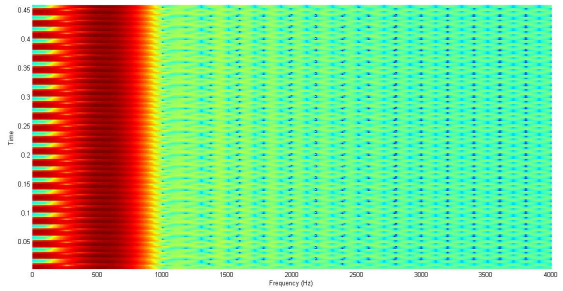

Fig.6(b) Spectrogram output for Vowel /o/.

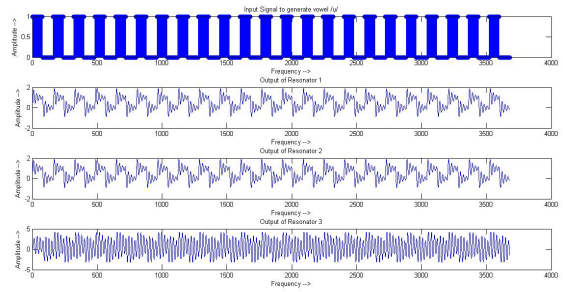

Fig.7(a) Time domain representation of Cascade formant output for Vowel /u/.

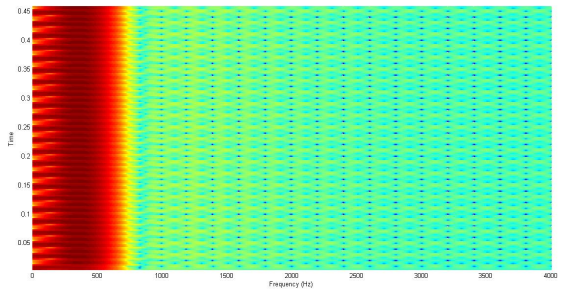

Fig.7(b) Spectrogram output for Vowel /u/.

The time domain representation of the different vowels synthesized were observed. From the spectrogram's of the different vowels synthesized, it is seen that the formant frequencies F1 and F2 obtained from the synthesized vowels matched with the formant frequencies used to synthesize these vowels. The synthesized vowels were robotic in nature.

IV. Overall Quality Evaluation

It is necessary to evaluate the speech output obtained in the text to speech system. Several methods are available to measure the overall quality of the synthesized speech. One such method widely used is the Mean opinion score (MOS) which is also the simplest method. MOS is based on how much the listener can clearly understand the output speech and rate it in a scale of 1(bad), 2(poor), 3(fair), 4(good) and 5(excellent). MOS test was performed on the vowels synthesized and the results are as shown in Table II.

TABLE II. RESULTS OF MOS

Vowels	Listeners
	MOS
/a/	4.6
/e/	4
/i/	4.5
/o/	4.1
/u/	4.5

V. SUMMARY AND DISCUSSON

A brief understanding of the speech synthesis methods for obtaining the output speech has been presented. The text to speech output is based on the amount of intelligibility and naturalness. It is observed that concatenation synthesis is the widely used speech synthesis technique as it provides high level of naturalness and intelligibility. Formant synthesis is based with the source-filter model and produces unnatural output speech. Articulatory synthesis is flexible to model the speech production system directly and is rather complex in nature. Hidden Markov model is not completely automatic. However, the selection of the appropriate speech synthesis method for the application can produce an efficient text to speech system. By considering cascade formant synthesis for implementation made it easier to generate speech waveforms for selected vowels. The speech output obtained were more robotic and unnatural in nature.

REFERENCES

[1] Lemmetty S., "Review of speech synthesis technology", Master's Thesis, Dept. Of Electrical and Communication Engineering, Helsinki University of Technology, 1999.

[2] Mohammad Yasir Arafat, Sanjana Fahrin, Md. Jamirul Islam, Md. Ashraf Siddiquee, Afsana Khan, Mohammed Rokibul Alam Kotwal, Mohammad Nurul Huda, A. Czyżewski, "Speech synthesis for Bangla text to speech conversion", *IEEE 2014 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* , 18-20 Dec, 2014.

[3] Anusha Joshi, Deepa Chabbi, Suman M and Suprita Kulkarni, "Text to speech system for Kannada language", *IEEE 2015 International Conference on Communications and Signal Processing (ICCSP)*, 2-4 April, ICCSP 2015.

[4] Ming ouh-young, Chin-jiting Shie, Chiu-yu Tseng and Lin-shan Lee, "A Chinese text-to-speech system based upon a syllable concatenation model", *IEEE 1986 International Conference on Acoustics, Speech, and Signal Processing, ICASSP 86,* (Vol:11), *TOKYO,* 7-11 April 1986.

[5] Akintoba Emmanuel Akinwonmi, Boniface Kayode Alese, "A prosodic Text-to-Speech system for Yorùbá language", *IEEE 2013 8th International Conference for Internet Technology and Secured Transactions (ICITST),* 9-12 Dec., 2013.

[6] Ovidiu Buza, Gavril Toderean, Jozsef Domokos, "A rule-based approach to build a text-to-speech system for Romanian", *IEEE 2010 8th International Conference Communications (COMM),* 10-12 June 2010.

[7] Partha Shankar Nayak, "Bangla web page reader - an approach to Bangla text-to-speech conversion", *IEEE 2014 International Conference on Recent Advances and Innovations in Engineering (ICRAIE),* May 09-11, 2014.

[8] Dennis H. Klatt, (1980),"Software for a Cascade/Parallel Formant Synthesizer", *Journal of the Acoustical Society of America, JASA*, Vol. 67: 971-995,1980.