# Speech Synthesizer for the Pashto Continuous Speech based on Formant Technique

Sahibzada Abdur Rehman Abid[1], Nasir Ahmad[1], Muhammad Akbar Ali Khan[1], Jebran Khan[1],

[1]Department of Computer Systems Engineering, University of Engineering and Technology Peshawar, Pakistan
sahib_uetian@yahoo.com,n.ahmad@nwfpuet.com,
engrakbar101@yahoo.com, engr.jebrankhan@gmail.com

**Abstract.** This paper describes formant based Pashto continuous speech synthesizer, which automatically generates the Pashto Speech from given text using formants. In this work, initially some samples of Pashto Speech are recorded in a noise free environment by using the Sony PCM-M 10 linear Recorder device. The recorded Pashto continuous speech audio file are then split into isolated Pashto sentences, using the Adobe Audition ver 1.0. After splitting the entire continuous Pashto audio file into isolated sentences the features from these isolated Pashto sentences have been extracted by utilizing the colea tool. Pashto speech sentences from given text are then synthesized through formant technique by utilizing the features extracted from the isolated Pashto audio files.

**Keywords:** Speech Synthesis, Formant synthesis, Pashto speech, Pashto speech recording, Pashto speech synthesis, Colea

## 1 Introduction

Speech synthesis is the process of converting the given input text into corresponding synthesized acoustic signal as an output. It is the reverse phenomena of automatic speech recognition which converts the given acoustic signal into text. The three approaches commonly used for automatic speech production are; formant synthesis [1] concatenative synthesis [2] and articulatory synthesis [3]. In cancatenative speech synthesis, the speech signal for the given text is achieved by concatenating the speech units already stored in a database. In the formant synthesis approach the recorded speech database is not used directly, however the speech parameter such as formant frequencies are extracted from the recorded speech files and those formant frequencies are stored in text files in the vector form. In formant synthesis approach, frequencies levels are changed continuously to produce the waveform of the synthetic speech.

The work on Pashto continuous speech synthesis and Pashto digits and numbers synthesis based on concatenative synthesis approach has been presented in [4] and [5] respectively. Besides this basic work, to the best of author's knowledge, no work has yet been done on the formant synthesis approach in the Pashto language. The formant synthesis technique that is based fundamentally on the source filter model of speech has mostly been used for the automatic speech synthesis in the last decades. The research on formant synthesis has been more active due to its potential to model emotive speech and more efficient synthesis of different voices [6]. For producing a high quality speech five formants are used however for the production of an intelligible speech the use of three formants is usually sufficient. Every formant is generally modeled with a two pole resonator that enables the formant frequency as well as its bandwidth to be specified [7]. In rule-based formant synthesis, the determination of parameters a set of rules has been used to synthesize a desired speech [7]. The production of infinite number of sounds through the formant synthesis makes it further flexible as compared to the other synthesis methods.

The rest of paper is organized as follows. Section 2 discusses the recording of Pashto continuous speech and its analysis. Section 3 shows how Pashto speech has been artificially synthesized using formant technique. Section 4 explains the results of Pashto speech synthesis and presents a discussion on the result.

## 2 Pashto Continuous Speech Recording and its Analysis

### 2.1 Pashto Continuous Speech Recording

Sony PCM-M 10 Linear Recorder has been used for the recording of Pashto continuous speech. The speaker selected for the recording of Pashto speech wee those, who could spoke the Yousafzai dialect flawlessly, smoothly and fluently. Then the recording has been performed in a noise free environment. Before starting the recording, the sensitivity level and signal to noise ratio of the recorder have been adjusted to coup with the background noise.

### 2.2 Pashto Speech Analysis

The analysis of Pashto speech is can be subdivided into two stages; the segmenting of the recorded Pashto continuous speech into isolated sentences and the extraction of formant frequencies. The process of Pashto speech analysis is depicted in Figure 1.

**Pashto Continuous speech Segmentation.** After Pashto continuous speech recording, the entire Pashto speech recorded file is split into isolated Pashto sentences using adobe edition version 1.0 software and saved in .wav format with a sampling rate of 16 kHz, 16 bit resolution and channels mono. Thus, the recorded Pashto audio file is segmented into isolated Pashto sentences.

**Features Extraction.** In formant based speech synthesis approach the extraction of formant frequencies (F1, F2, and F3) is the central procedure. In this work, the colea tool [8] has been utilized for extracting formant frequencies from the isolated Pashto sentences audio files. The extraction of formant frequencies from non-stationary signals is a difficult job and as speech is a continuously changing signal so formant frequencies extraction form the speech signal has been achieved by considering the windowed portions from the speech where the signal is relatively stationary.  Significant information related to the production of speech is obtained through –by-segment analysis. Before speech analysis, the speech data has been split into frames as human ear don't respond to exceptionally rapid changes of speech data content. Thus segments of 20ms duration have been taken which is the window size adopted in this

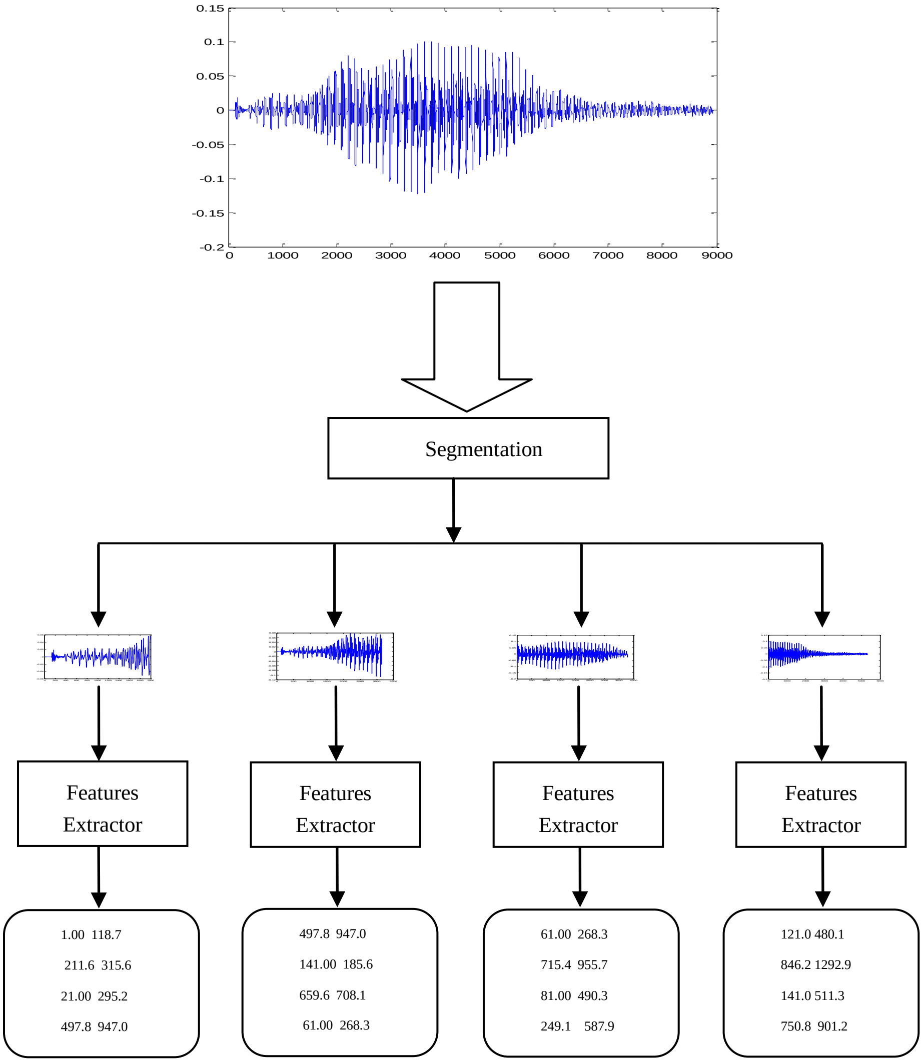

**Fig. 1.** Pashto Speech Analysis

work. Windowing is achieved by multiply the signal by a window having null values everywhere excluding the region of interest where its value is one [9]. The original signal is consider to be of infinite duration however, through the windowing operation the portion beyond the window duration is discarded and only the signal inside the windowed portion is obtained.

Different types of windows have been considered including rectangular and hamming window. The abrupt changes at the edges of rectangular window cause distortion of the signals which is one of the main disadvantages of the rectangular window. To avoid this distortion of signal, hamming window has been used here. The use of hamming window de-emphasizes the signal edges and thus the effects of sharp edge at the signal end are reduced. In several types of frequency domain analysis methods the use of hamming window has typically shown to give superior results [9]. The colea tool which has been utilized in the system proposed in this work has been based on hamming window approach to find out the formant frequencies. As, the first three formant frequencies are suitable for speech synthesis through formant technique, the spectral display of windowed Pashto speech signal is obtained and the first three formant frequencies and their related time duration are calculated using colea tool. The computed formant frequencies are then saved in text files. Some of the first three formant frequencies of the Pashto speech segment are shown in figure 2.

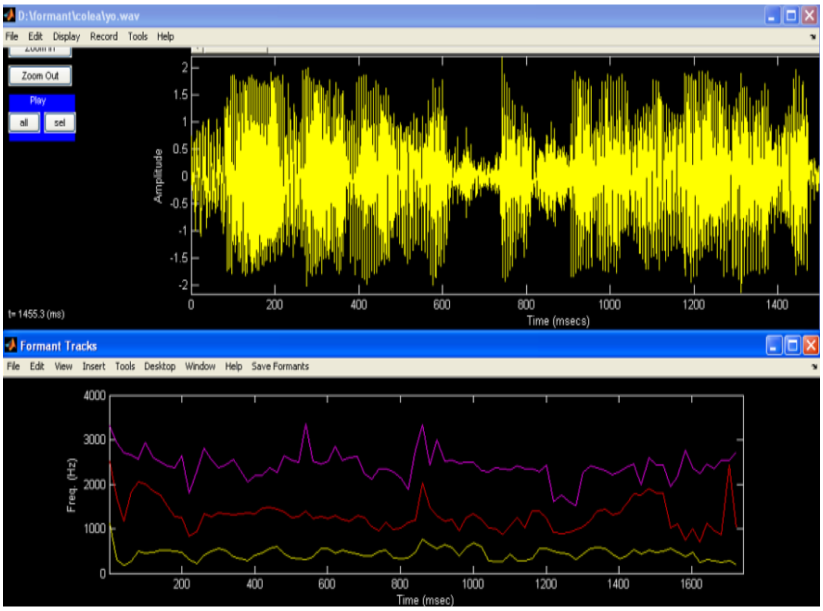

**Fig. 2.** The First three formant frequencies of the Pashto Speech Segment

In Figure 2, the frequency shown in yellow is F1, the one shown in red is F2 and while the one shown in purple is F3. A sampling rate of 22.050 kHz and LPC order of 16 is used here. As the period is predetermined at 20ms, the formants frequencies are obtained every 20ms. Some of the values obtained for different isolated Pashto sentences are shown in Table 1.

**Table 1.** The First three formant frequencies of the Pashto Speech Sentence

| t(msec) | F1(Hz) | F2(Hz) | F3(Hz) |
|---|---|---|---|
| 1.00 | 1118.751 | 2511.664 | 3315.647 |
| 21.00 | 295.264 | 1697.838 | 2947.033 |
| 41.00 | 185.681 | 1159.652 | 2708.128 |
| 61.00 | 268.331 | 1815.440 | 2655.755 |
| 81.00 | 490.301 | 2049.191 | 2557.967 |
| 101.00 | 456.627 | 2012.009 | 2947.150 |
| 121.00 | 480.172 | 1846.284 | 2592.930 |
| 141.00 | 511.325 | 1750.845 | 2501.282 |
| 161.00 | 514.689 | 1499.240 | 2405.471 |
| 181.00 | 503.884 | 1278.713 | 2367.506 |
| 201.00 | 480.305 | 1252.198 | 2637.290 |
| 221.00 | 300.322 | 822.434 | 1795.326 |
| 241.00 | 230.512 | 927.916 | 2303.689 |
| 261.00 | 409.490 | 1336.492 | 2818.182 |
| 281.00 | 485.142 | 1276.061 | 2553.558 |
| 301.00 | 554.165 | 1355.971 | 2361.235 |

## 3 Pashto Speech Synthesis through Formant Technique

The formant frequencies extracted from the Pashto audio files in speech analysis step and saved as an inventory in the text files are used, to produce the synthetic Pashto speech. In text files the formant frequencies are stored in the form of arrays of size n x 4, where row shows indicate the frame number; the first column indicates time and the rest of three columns represent the formant frequencies. Different text files have different number of frames numbers. In this work, the number of frames in a file ranges from 1 to 100s. Some of these vectors are shown in Table 1. During speech synthesis, the formant frequencies stored in text files are accessed and converted into matrix form. The formant bandwidth, having the same size as the formant vector has also been produced for every formant frequency and stored in the vector form. The bandwidth produced for some of the formants frequencies of Table 1 have been shown in Table 2.

In the proposed formant based Pashto speech synthesizer a sampling rate of 8 kHz and frame size of 20ms has been used. The sampling rate along with the frame duration gives the total number of frames within that duration from which the Pashto speech is then synthesized. For male 100Hz and for female 200Hz fundamental frequency has been selected.

**Table 2.** Formant Frequencies an Bandwidth

| t(msec) | F1 | B1 | F2 | B2 | F3 | B3 |
|---|---|---|---|---|---|---|
| 1 | 1118.75 | 111.875 | 2511.66 | 251.166 | 3315.64 | 331.564 |
| 21 | 295.26 | 29.526 | 1697.83 | 169.783 | 2947.03 | 294.703 |
| 41 | 185.68 | 18.568 | 1159.65 | 115.965 | 2708.12 | 270.812 |
| 61 | 268.33 | 26.833 | 1815.44 | 181.544 | 2655.75 | 265.575 |
| 81 | 490.30 | 49.030 | 2049.19 | 204.919 | 2557.96 | 255.796 |
| 101 | 456.62 | 45.662 | 2012.00 | 201.200 | 2947.15 | 294.715 |
| 121 | 480.17 | 48.017 | 1846.28 | 184.628 | 2592.93 | 259.293 |
| 141 | 511.32 | 51.132 | 1750.84 | 175.084 | 2501.28 | 250.128 |
| 161 | 514.68 | 51.468 | 1499.24 | 149.924 | 2405.47 | 240.547 |
| 181 | 503.884 | 50.388 | 1278.71 | 127.871 | 2367.50 | 236.750 |
| 201 | 480.30 | 48.030 | 1252.19 | 125.219 | 2637.29 | 263.729 |

In speech processing the signal is considered to be short time stationary so Fast Fourier Transform is carried out on these short time stationary signals. The use of Fourier transform represents the signals as a sum of periodic harmonics, which helps to produce the signal through the window function having zero value outside the window. The formants that have been found are the frequency domain representation of sinusoids at integer multiples of the fundamental frequency, entitled harmonics, or harmonic partials [10]. The impulse train of a signal S[n] for interval of N samples can be expressed as:

$$s(n) = \begin{cases} 1 & multiplesofN \\ 0 & otherwise \end{cases}$$

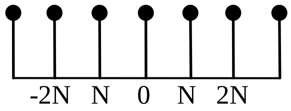

**Fig. 3.** Impulse train for interval N

The z- transform is used to transform the time domain signal into complex frequency domain. For a digital signal h[n] the z-transform [H(z)] is defined as:

$$H(z) = \sum_{n=-\infty}^{\infty} h(n) z^{-n}$$

Where z is the complex variable.

The Fourier transform of h[n] is obtained from its z-transform by substituting z=$e^{jw}$. The z-transform has been used to examine more common filter characteristics [11]. The above equation is an infinite series and its existence is not guaranteed. The following condition need to be fulfilled for the convergence of the series

$$\sum_{n=-\infty}^{\infty} |h[n]| |z^{-n}| < \infty$$

The transfer function coefficients are obtained to synthesize the signal. For each vector values of the formant frequencies and the formant bandwidth the synthesized signal is produced. The outputs of the synthesized signal have been written in .wav format and the signal plot of the synthesized speech is obtained.

## 4 Results and Discussions

The signal plot for the recorded Pashto sentence is shown in Figure 4 while the signal plot for the synthesized speech is shown in Figure 5. Comparing the signal plots of the recorded Pashto sentence and that of artificially produced Pashto sentence, it was found that two signals are more similar in the region where a vowel sound is pronounced however they differ slightly in the regions where consonant sounds are pronounced. The same phenomenon was observed in listening of the recorded and artificially produced Pashto sentences. The same pattern was observed for all other Pashto sentences synthesized through the formant synthesis technique. The earlier work on the Pashto speech synthesis is based on the concatenative synthesis technique which is though more natural as it is produced by the concatenation of units obtained from actual speech, however it needs an enormous amount of linguistic resources and are less flexible while producing different speaking styles. The proposed formant based technique on the other hand  require less linguistic resources and can produce the output Pashto speech in various speaking styles.

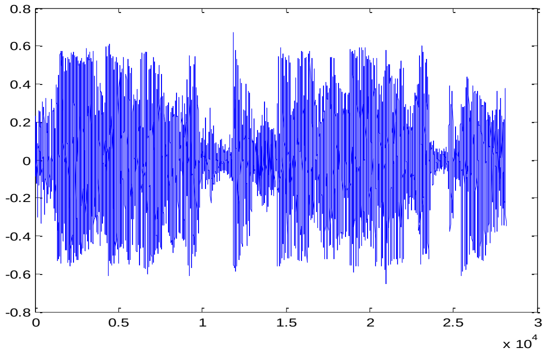

**Fig. 4.** Signal Plot of the Recorded Pashto Speech Sentence

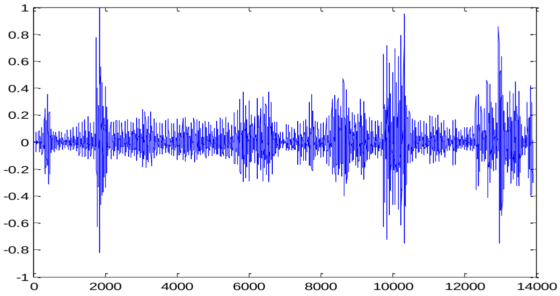

**Fig. 5.** Signal Plot of Synthetic Pashto Speech Sentence

## 5 Conclusion

In this research the formant based Pashto continuous speech synthesis has been presented. Pashto continuous speech has been recorded and split into isolated sentences. The formant frequencies have been extracted from the isolated Pashto speech sentences and stored in inventory in the vector form. The Pashto speech sentences are then synthesized and their results compared with the recorded speech both in terms of signal plots and listening.

## References

1. Anberbir. T., Takara. T. : Development of an Amharic Text-to-Speech System Using Cepstral Method, Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 46–52, Athens, Greece, 31 March 2009.

2. Beller. G.: Gestural Control of Real Time Concatenative Synthesis, ICPhS XVII, Hong Kong, 17-21 August 2011.

3. Palo. P. : A Review of Articulatory Speech Synthesis, MSc. Thesis. Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, (2005).

.

4. Khan. M. A. A., Abid. S. A. R, Zuhra. F. T., and Ahmad. N.: The Development of Pashto Speech Synthesis System, International journal of Computer Applications, 71(24):49-53. Published by Foundation of Computer Science, New York, USA, (June 2013).

5. Abid. S. A. R., Ahmad. N., Khan. M. A. A., and Zuhra. F. T.: Concatenative based Pashto Digits and Numbers Synthesizer, International journal of Computer Applications, 72(6):39-42. Published by Foundation of Computer Science, New York, USA, (May 2013).

6. Öhlin D., and Carlson R., “Data-driven formant synthesis” Proceedings, FONETIK 2004, Dept. of Li nguistics, Stockholm University (2004).

7. Donovan. R.: Trainable Speech Synthesis, PhD. Thesis. Cambridge University Engineering Department, England. (1996).

8. Loizou. P.: COLEA: A Matlab Software Tool for Speech Analysis, from http://www.utdallas.edu/~loizou/speech/colea.htm, (1999).

9. Cassidy. S.: Speech Recognition, Department of Computing, Macquarie University, Sydney, Australia, (2002).

10. Schwarz. D.: Spectral Envelopes in Sound Analysis and Synthesis, Version 98.1 p1 release, March 2nd, (1998).

11. Huang. X., Acero. A., Hon. H.: Spoken Language Processing. Prentice Hall, Upper Saddle River, New Jersey, (2001).