Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 46 (1) 47-52 (2014)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Formant based Pashto Digits and Numbers Synthesizer

M. A. A. KHAN, S. A. ABID*, N. AHMAD[++]*, A. W. ABBAS, M. WAQAS*

Institute of Business and Management Sciences, Khyber Pakhtunkhwa Agricultural University, Pakistan

Received 9th December 2013 and Revised 27th February 2014

Abstract: This paper presents the development of Pashto digits and number synthesis system based on formant synthesis approach. The speech synthesis system generates the Pashto digits and number, form sefer (zero) to yaw zer (one thousand). The formant frequencies are extracted from the recording of isolated Pashto digits and numbers using colea tool. The Pashto digits are then synthesized by using the features extracted from the audio of these isolated Pashto digits. The training audio data used in this work has been recorded using Sony PCM-M 10 linear Recorder. The entire recording has been conducted in a noise free office environment. The recorded Pashto digits audio file is then split into isolated digits by utilizing Adobe Audition ver 1.0.

Keywords: Speech Synthesis, Formant Synthesis, Pashto Digits Synthesis, Colea

1. INTRODUCTION

Speech synthesis is the process in which text is given as input and it give us the corresponding synthesized acoustic signal as an output. The three basic approaches used for the production synthetic speech are; formant synthesis (Nishizawa, *et al.,* 2000) concatenative synthesis (Khan, *et al.,* 2013) and articulatory synthesis (Ling, *et al.,* 2013). Speech synthesis system converts normal language text, into corresponding speech signal. An inverse process of the speech synthesis is the automatic speech recognition where the input to the system is the audio signal while the output is the recognized text. Pashto spoken digits database and automatic digit recognition has been developed in (Abbas, *et al.,* 2012) while Pashto digits and numbers synthesis system based on Concatenative approach have been reported in (Abid, *et al.,* 2013). In this work, Pashto digits synthesis based on the formant approach has been presented.

The Formant synthesis is a descriptive, acoustic-phonetic approach to speech synthesis (Allen *et al.,* 1987). In the formant synthesis, parameters such as fundamental frequency and formant frequencies levels are varied over time to create a waveform of artificial speech. Formant synthesis is based on the source filter model of speech and is the most broadly used synthesis method in last two decades (Abid *et al.,* 2013). In the formant synthesis, human speech samples are not used. Synthesis of dissimilar voices and voice characteristics, and the modeling of emotive speech have kept research on formant synthesis active (Charlson *et al.,* 1991). At least three formants are required to produce an intelligible speech; however five formants have been used for producing a higher quality speech. Each formant is usually modeled with a two pole resonator which enables both, the formant frequency and its bandwidth to be specified (Donovan, 1996). Rule-based formant synthesis is based on a set of rules used to determine the parameters necessary to synthesize a desired utterance (Allen *et al.,* 1987). Infinite number of sounds provided by the formant synthesis makes it more flexible than other synthesis methods.

The paper is organized as follows. Section 2 introduces the Pashto language and its digits. Section 3 describes the Pashto digits recording and its analysis. Section 4 shows how Pashto Digits and numbers have been artificially synthesized through formant technique. Section 5 explains Pashto Digits synthesis results and discussion.

2. MATERIAL AND METHODS

Pashto Language and Its Digits

Pashto is the national language of Afghanistan and one of the widely spoken languages of Pakistan, has an estimated 50-60 million speakers all over the world (Herbert and Sloan, 2009). Pashto language has three dialects, Northern Pashto, Central Pashto and Southern Pashto (Paul, 2009). The number system in the Pashto language is similar mostly to the foreign languages such as Persian (Hejazi *et al.,* 2009), Arabic (Alotaibi *et al.,* 2010) and other languages (Herbert and Sloan, 2009). In the Pashto language all digits from Sefer (0) to Las (10) have dissimilar articulation. The numbers after Las (10), from Yawo-las (11) to Noo-las (19) have common postfix Las with every number from Yaw (1) to Naha (9) with minute deviation, while from Yaw-Vesht (21)

[++] Corresponding author: N. Ahmed email n.ahmad@nwfpuet.edu.pk Ph. 091-9216590
*University of Engineering and Technology, Peshawar, Pakistan

to Naha-Vesht (29) the common postfix is Vesht with number from Yaw (1) to Naha (9) as a prefix. The similar model is trailed in favor of every one number up to sul (100). There are a number of variations of this prototype among dissimilar areas as well as dialects. The digits and number used in this paper are based on the Yousafzai/Northern Pashto dialect shown in **Table 1**.

Table: 1 Pashto digits of Yousafzai/ Northern Pashto dialect

Digits	Pashto	Pronunciation	Pashto
0	۰	Sefer	صفر
1	۱	Yaw	یو
2	۲	Dwa	دوَه
3	۳	Dray	درے
4	٤	Celour	څلور
5	٥	Penza	پنځه
6	٦	Shpeg	شپږ
7	۷	Owa	اووَه
8	۸	Ata	اَته
9	۹	Naha	نهه
10	۱۰	Las	لس
11	۱۱	Yawo- Las	یو لس
12	۱۲	do- Las	دوَه لس
13	۱۳	Dyar- Las	دىأر لس
14	۱٤	Swaar- Las	څوَار لس
15	۱٥	Penza- Las	پنځه لس
16	۱٦	Shparh-as	شپارس
17	۱۷	Owa- Las	اووَه لس
18	۱۸	Ata- Las	اَته لس
19	۱۹	Noo- Las	نو لس
20	۲۰	Shul	شل
21	۲۱	Yaw-Vesht	یوښت
22	۲۲	Dwa—Vesht	دوَه وښت
23	۲۳	Dray-Vesht	درے وښت
24	۲٤	Celour-Vesht	څلر وښت
25	۲٥	Penza-Vesht	پنځه وښت
26	۲٦	Shpag-Vesht	شپږ وښت
27	۲۷	Owa-Vesht	اووَه وښت
28	۲۸	Ata-Vesht	اَته وښت
29	۲۹	Naha-Vesht	نهه وښت
30	۳۰	Dairsh	دىرش
40	٤۰	Celwaikht	څلوىښت
50	٥۰	Panzoos	پنځوس
60	٦۰	Shpeta	شپېته
70	۷۰	Away	اوىه
80	۸۰	Atya	اتىه
90	۹۰	Nawi	نوى
100	۱۰۰	Sul	سل

Pashto Digits Recording And Analysis Pashto Digits Recording .

For recording the Pashto digits the Sony PCM-M 10 Linear Recorder is used and such a speaker has been chosen for recording purpose who has spoken the Yousafzai dialect smoothly and flawlessly. Then the recording has been conducted in noise free environment. The sensitivity level and signal to noise ratio of the recorder have been adjusted and then recording of Pashto digits was performed. Pashto digits from Sefer (0) to Sul (100) and then after that straightforwardly Zar (1000) has been spoken by the speaker.

Pashto Digits Analysis-

Pashto digits analysis is consisted of two parts segmenting the Pashto recorded digits into isolated digits and then formant frequencies extraction. The phenomena of Pashto digits analysis has been described in **Fig. 1.**

Pashto digits segmentation-

After Pashto digits recording, by using the adobe edition ver 1.0 software the entire Pashto digits recorded file has been split into isolated Pashto digits and have been saved in .wav format with 16 kHz sample rate, 16 bit resolution and channels mono. After segmenting the recorded Pashto digits we gained isolated digits from Sefer (0) to naha (9) then Las

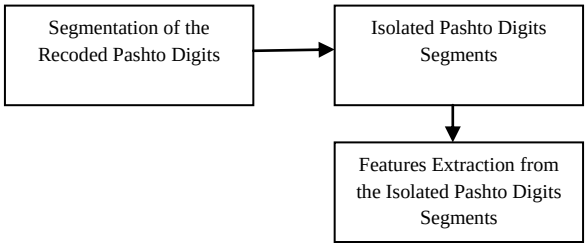

Fig.1: Pashto digits analysis

(10),Shul (20), Dairsh (30), Celwaikht (40), Panzoos (50), Shpeta (60), Awaya (70), Atya (80), Nawi (90), Sul (100), and Zar (1000).

Features Extraction-

For the isolated Pashto digits the formant frequencies extraction has been conducted via using the colea tool (Loizou, 1999). Feature taking out for instance formant frequencies (F1, F2, and F3) is the central procedure within formant based speech synthesis technique. As speech signal is continuously changing and formant frequencies extraction is difficult from such a signal so for this purpose the speech signal is considered being stationary in appropriate time window and then formant frequencies have been extracted from the speech signal. When analysis is conducted on the origin of segment-by-segment, important information concerning the generation of speech is pulled off. As human ear cannot get act toward remarkably quick modification of speech data content, therefore earlier than the analysis speech data is split in frames. In our proposed work the 20ms frame i.e. window size has been taken. Windowing might be observed since multiply a signal through a window that has null value everywhere excluding the region in favor of attention where its value is one (Cassidy, 2002). Therefore the signals to which we utilized are imagined to be infinite

and those portions of the signals that have zero value are discarded and give attention on immediately the windowed segment of the signal. The type of window consists of only zeros and ones a value is known as rectangular window. The main disadvantage of such type window is abrupt change at edges occurs that is the main mean of distortion of the signals that have been examined. Actually each windowing process carries some distortion, because the signal is varied by the window. To drop off such distortion hamming window has been utilized. Such type of window has zero values at the edges and gradually ascends and at the middle it raises to one.

The signal edges are de-emphasized along with the edge effects are reduced via using such kind of a window in several kinds of analysis Hamming window consumption is superior typically within the frequency domain methods (Cassidy, 2002). Whereas rectangular window is the best alternative in support of time domain method. In our proposed system for finding out the formant frequencies colea tool has been utilized that is based upon hamming window approach. The taking out of the formant frequencies from the Pashto digit speech signal, firstly the speech units are separated in 20 ms small frames of samples. After that via time window the Pashto digit speech signal is expressed to get spectral display of the unit. The first three formant frequencies are suitable for speech synthesis through formant technique. The first three formant frequencies values as well as their related time duration of the isolated Pashto digit speech signal have been computed via the colea tool and have been saved in text file. Some of the first three formant frequencies of the Pashto digit speech segment have been shown in **Fig 2.**

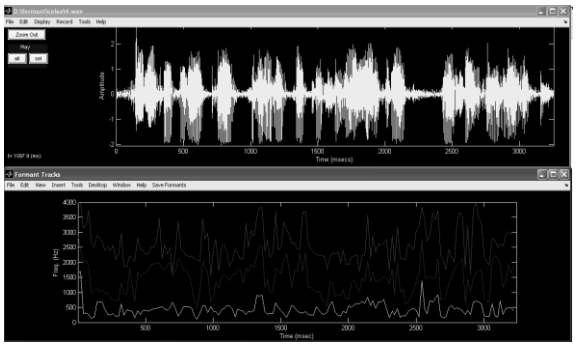

Fig 2. The first three formant frequencies of the Pashto digit speech segment.

In **Fig 2** the frequency pointed out at bottom is F1, the frequency pointed out at the center is F2 and the third one at the top is F3. 22.050 kHz sampling rate as well as LPC order of 16 is utilized. As we have predetermined the period as 20ms then the formants will be verified within every 20ms period, the values that we obtain for different isolated Pashto digits among them some of the values have been shown in **Table 2.**

Table 2: The first three formant frequency of the Pashto digit segment

t(msec)	F1(Hz)	F2(Hz)	F3(Hz)
1.00	520.409	1742.303	3222.140
21.00	297.272	1686.036	3831.855
41.00	457.226	1624.961	3519.254
61.00	485.375	1593.490	2755.578
81.00	321.815	1780.557	2647.248
101.00	393.965	1744.455	2613.837

Pashto Digits And Numbers Synthesis

Pashto digits synthesis is demonstrated by F**ig.3**

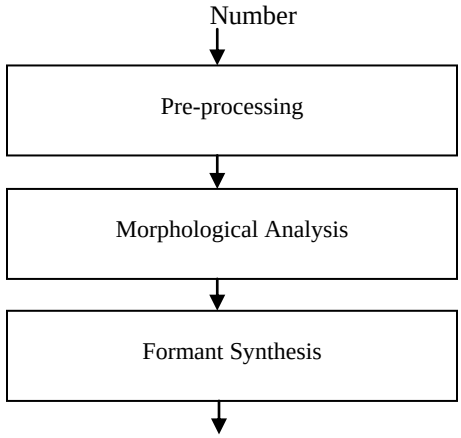

Fig. 3: Pashto Digits Synthesis

Pre-processing-

The input number is given in string format by the user. The number string is then separated into each digit. Which is based on the digit value and its position, sampled data of respected speech is selected. As for (45), it will be read as Penza Celwaikht.

Morphological Analysis -

Pashto digits can be explained in terms of a morphological dictionary, which gives a list of 0 to 9 digits.

Pashto Digits Formant Synthesis-

Formant frequencies have been extracted in speech analysis process, so by using those formant frequencies Pashto digits and numbers have been produced artificially. The formant tracks of the Pashto digits required for the formant synthesizer have collected in the inventory. The formants have been

stored in vector form that consist of n rows which shows frames number and four columns, the first column shows the time and remaining three columns represents formant frequencies. Different files have different numbers of rows, the least number of rows would be one and can be go up to 100s of lines of rows as some of the rows and columns combination have been shows in table 1. The formant frequencies collected in the inventory in the vector form accessed by the code have been changed into the matrix form and for every formant the corresponding formant bandwidth has been formed having the same size as the formant vector and stored in vector form just like the formant vector. The bandwidth generated for some of the formants frequencies of table 2 have been shown in **Table 3**.

Table 3: Formant frequencies and bandwidths

t(msec)	F1	B1	F2	B2	F3	B3
1	520.4	52.04	1742.3	174.23	3222.1	322.21
21	297.2	29.72	1686.0	168.60	3831.8	383.18
41	457.2	45.72	1624.9	162.49	3519.2	351.92
61	485.3	48.53	1593.4	159.34	2755.5	275.55
81	321.8	32.18	1780.5	178.05	2647.2	264.72
101	393.9	39.39	1744.4	174.44	2613.8	261.38

In our proposed system 8 kHz sampling rate and frame size of 20ms have been considered. The sampling rate along with the frame duration gives us the total number of frames within that duration from which the Pashto digit and number is then artificially synthesized. For male 100Hz and for female 200Hz fundamental frequency has been selected.

The signal is supposed to be short time stationary in speech processing; on these short time stationary signals Fast Fourier transform is carried out. The Fourier transform usage makes it probable to demonstrate every function, such as the non-periodic ones as a sum of periodic sinusoids, further decomposition of which is not possible. This facilitates to create the signal through the window function which has zero value outside some define range. The formants what we found are in fact the frequency domain representation of sinusoids at integer multiples of the fundamental frequency, called harmonics, or harmonic partials (Schwarz, 1998). The impulse train of a signal S[n] for all interval time interval of N can be expressed as:

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

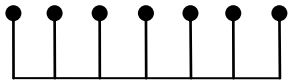

Fig. 4: impulse train for interval N

The time domain signals that are a series of real numbers have been changed into complex frequency domain by using of the z-transform. The z-transform is simplified form of the Fourier transform. For a digital signal h[n] the z-transform [H(z)] can be defined as:

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

Where z is complex variable, the Fourier transform of h[n] equivalents its z-transform at $z=e^{jw}$. The Fourier transform has used to plot the filter's frequency response, while to examine more common filter characteristics the z-transform has been used (Huang *et al.*, 2001). As the above equation illustrates infinite sum and such a sum existing is not assured. A sufficient circumstance utilized for convergence is:

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

This is correct simply in favor of a Region of Convergence (ROC) within the complex z-plane. Getting the values of the radius, pole locations along with angular radius which have been shown by z, j, and w symbols respectively. The achievement of the transfer function coefficients makes it possible to synthesize the signal. Then for every vector values of the formant bandwidth along with the formant frequencies the synthesized signal will be build up. At the end, the outputs of the synthesized signal have been saved in the .wav format to the specified directory and can also obtain the signal plot of the speech that has been synthesized.

3. RESULTS AND DISCUSSIONS

Fig. 5 shows the signal plot of Pashto digit Owa (seven), and **Fig. 6** shows signal plot of Panzoos (fifty), the parameters of Owa (seven) and Panzoos (fifty), have extracted and saved it in the vector form in the text file present in the inventory, but parameter of Owa Panzoos (fifty seven) is not available in the inventory by combining the parameters of Owa (seven) and Panzoos (fifty), Owa Panzoos (fifty seven) is produce and similarly other Pashto numbers too from Yaw Las (11) upto Sul (100).

For comparing the produce Pashto digits Owa Panzoos (fifty seven) by formant thenique plot of which is shown in **Fig. 7** to the plot of recoded Owa Panzoos

(fifty seven) shown in **Fig.8** spoken by person. At some points both signal plots of Pashto number Owa Panzoos (fifty seven) shows similarity and at some points they are not similar this is due to we have used formant frequencies that have been extracted from the Pashto isolated digits so the artificially produced Pashto Digits and numbers where the vowels have been pronounced sounds natural to that of the recorded and where consonants have been pronounced at some points that produced sounds noisy.

Similarly Pashto digits and numbers from Sefer (0) to yaw-sul (100) have been generated both speech wise as well as plot wise by concatenating few isolated Pashto digits formant frequencies present in the inventory in vector form.

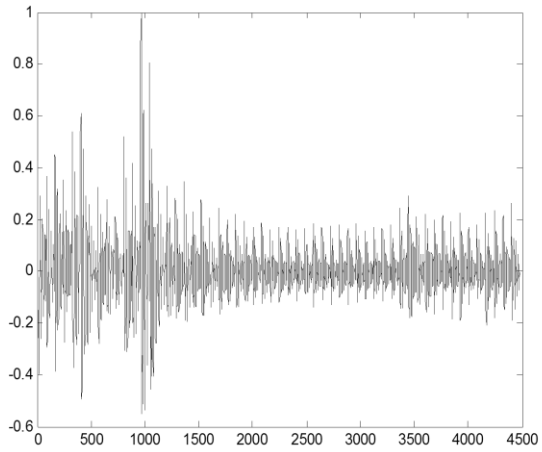

Fig.5: Signal Plot of Artificially Synthesized Pashto Digit Owa (Seven)

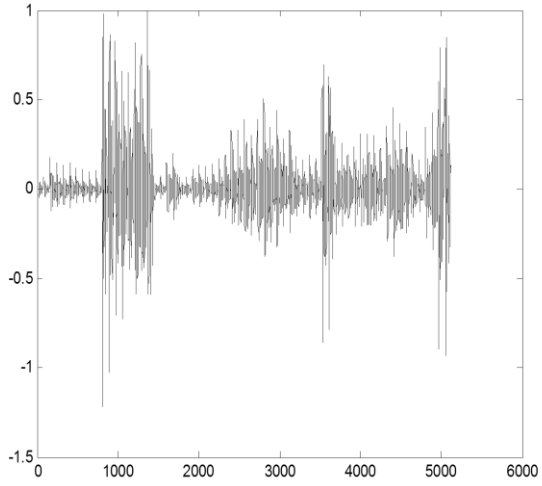

Fig. 6: Signal Plot of Artificially Synthesized Pashto Number Panzoos (Fifty)

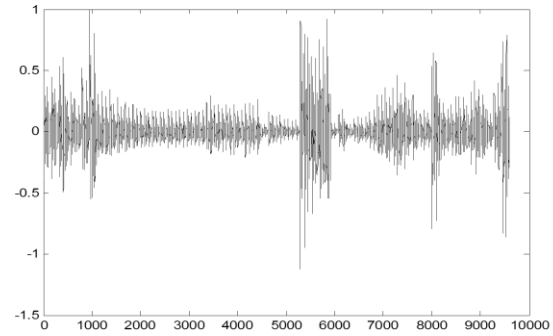

Fig.7: Signal Plot of Artificially Synthesized Pashto Number Owa Panzoos (Fifty Seven)

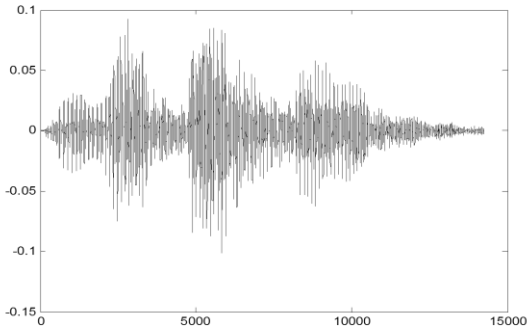

Fig. 8: Signal Plot of the Recorded Pashto Number Owa Panzoos (Fifty Seven)

4. CONCLUSION AND FUTURE WORK

In this research the formant based Pashto digits and numbers have been presented. We have recorded the Pashto digits and then exratcted formant frequencies of few isolated digits from Sefer (0) to Naha (9), Las (10), Shul (20), Dairsh (30), Celwaikht (40), Panzoos (50), Shpeta (60), Awaya (70), Atya (80), Nawi (90) and Sul (100) and stored the extracted formant frequencies in the vector form in the inventory. Then we have synthesized Pashto digits and numbers from yaw (1) to yaw-sul (100) by concatenating the formant frequencies of the isolated digits speech wise as well as plot wise from that limited isolated Pashto digits formant frequencies stored in the inventory.

REFERENCES:

Abbas, A. W., N. Ahmad. and H. Ali. (2012), "Pashto Spoken Digits Database for Automatic Speech Recognition Research", 18th International Conference on Automation & Computing, Loughborough University, UK, 8 September 2012, 363-366.

Abid, S. A., N. Ahmad, M.A.A. Khan, and F. T. Zuhra. (2013), "Concatenative based Pashto Digits and Numbers Synthesizer", International Journal of Computer Applications, 72 (6), 39-42.

Allen, J., M. S. Hunnicutt, and D. Klatt, (1987), "From text to speech the MITalk system", MIT Press, Cambridge, Massachusetts.

Alotaibi, Y. A., M. Alghamdi, and F. Alotaiby, (2010), "Speech Recognition System of Arabic Digits based on A Telephony Arabic Corpus", ICISP 2010, Canada.

Carlson, R., B. Granström, and I. Karlsson, (1991), "Experiments with voice modelling in speech synthesis", Speech communication, 10, 481–489.

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

Donovan, R. (1996), "Trainable Speech Synthesis", PhD. Thesis, Cambridge University, England.

Hejazi, S. A., R. Kazemi, C. Ghaemmaghami, and S. Sharif (2009), "Isolated Persian digit recognition using a hybrid HMM-SVM", International Symposium on Intelligent Signal Processing and Communications Systems, (ISPACS 2008), Bangkok, Thailand, 1-4.

Herbert, and Sloan, I. (2009), "A Grammar of Pashto a Descriptive Study of the Dialect of Kandahar, Afghanistan", Ishi Press International, 210.

Huang, X., A. Acero, H. Hon, (2001) "Spoken Language Processing", Prentice Hall, New Jersey.

Khan, M. A.A., S. A. Abid, F. T. Zuhra, and N. Ahmad. (2013), "The Development of Pashto Speech Synthesis System" International Jjournal of Computer Applications, 71 (24), 49-53.

Ling, Z., K. Richmond, and J. Yamagighi, (2013) "Articulatory Control of HMM-Based Parametric Speech Synthesis Using Feature-Space-Switched Multiple Regression", IEEE Transctions on Audio, Speech, and Language Processing, 21(1), January 2013.

Loizou, P., (1999) "COLEA: A Matlab Software Tool for Speech Analysis", from http://www.xmarks.com/site/www.utdallas.edu/~loizou/speech/colea.htm. [accessed 4/25/2014, 11:10 AM]

Nishizawa, N., N. Minematsu, and K. Hirose, (2000) "Development of a formant-based analysis-synthesis system and generation of high quality liquid sounds of Japanese", INTERSPEECH 2000, 725-728.

Paul, L. M., (ed.), (2009), "Ethnologue: Languages of the World", (16th eddition), Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com. [accessed 4/25/2014, 11:10 AM]

Schwarz, D, (1998) "Spectral Envelopes in Sound Analysis and Synthesis", Diplomarbeit Nr.1622, Fakultät Informatik, Universität Stuttgart, Stuttgart, Germany, 1998.