ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY
January 2020, Vol 13(02), 218 – 228
DOI: 10.17485/ijst/2020/v13i02/149356,

Prosody Generation Using Back Propagation Neural Networks for Sindhi Speech Processing Applications

Shahid Ali Mahar[1], Mumtaz Hussain Mahar[1], Shahid Hussain Danwar[1], Javed Ahmed Mahar[1,*] and Aijaz Ali Shaikh[2]

[1]Department of Computer Science, Shah Abdul Latif University, Khairpur, Sindh, Pakistan
[2]Department of Mathematics, Shah Abdul Latif University, Khairpur, Sindh, Pakistan

Article Type: Article

Article Citation: Shahid Ali Mahar, Mumtaz Hussain Mahar, Shahid Hussain Danwar, Javed Ahmed Mahar, Aijaz Ali Shaikh. Prosody generation using back propagation neural networks for Sindhi speech processing applications. *Indian Journal of Science and Technology*. 2020; 13(02), 218-228. DOI: 10.17485/ijst/2020/v013i02/149356

Received date: December 1, 2019

Accepted date: December 17, 2019

***Author for correspondence:**
Javed Ahmed Mahar ✉ mahar.javed@gmail.com 📍 Department of Computer Science, Shah Abdul Latif University, Khairpur, Sindh, Pakistan

Abstract

Analysis and synthesis of speech to be automated still require more research efforts in general and for the development of speech processing applications based on Arabic Script like Sindh Text-to-Speech in particular. To achieve the required results from the speech processing applications prosodic features must be exercised extremely as the prosody is highly linked with the information of sounds having different characteristics like linguistic rules, complications and variations of expressions. **Objectives:** This study aims to generate and analyze the prosodic information specifically pitch and duration from the recorded Sindhi sounds using the back propagation neural network. **Methods**: Two methods are used to obtain the prosodic information of Sindhi sounds, PRAAT speech analyser is used to obtain the results and for the validation a back propagation neural network model is implemented. From the four districts of Sindh 228 speakers were chosen and the sound of different descriptive sentences was recorded for the experiments. **Finding:** After the experiments with a neural network model with multiple layers on the collected sound, 98.8% a highly acceptable level of accuracy achieved at the 18th epoch among the 100 epochs. **Application and improvements:** The generated Sindhi prosodic information and adopted research methodology will be supportive to the scholars of Sindhi speech processing applications. This research work can be considered as the first step as no work for generating Sindhi prosody is found yet.

Keywords: Sindhi Recorded Sounds, Pitch, Duration, Speech Analysis, Prosody Generation.

1. Introduction

Sindhi, being categorized into six dialects, is frequently spoken across the world with assorted accents [1]. This language has a huge sound inventory and is linguistically as well as phonologically rich as compared to other languages spoken in the subcontinent [2]. For last two decades, large research contributions regarding Sindhi language processing have been published but no worthwhile work is found, generally for speech processing and particularly prosody generation due to complex variations in Sindhi accents [3].

The objective of this research is to generate and validate the prosodic information from the units of the recorded Sindhi sounds. The fundamental and mandatory prosodic features are pitch and duration which are always considered by researchers as prerequisite parameters for the development of software applications pertaining to speech like speech recognition and text to speech [4]. It is observed that the generation and analysis process of the prosody is complicated and difficult because the prosody is connected with different levels of information having different natures. But, prosodic information is important to correct the sentence accent in the process of automatic language understanding, communications and speech synthesis [5].

Various modelling approaches like statistical data-driven [6], rule-based [7] and hybrid [8] have been proposed and implemented through which prosodic information is obtained using the recorded sounds of the languages [9]. But comprehensive research work on the automatic generation and analysis of prosodic information is not performed yet for the Sindhi language; hence, authentic Sindhi speech processing applications are not developed for common use in routine life activities. It is because of the deficiency of the prosody generation modelling and investigation which is essential for speech synthesis.

In [10] this study, various undergraduate students of our university are selected as speakers for the recording of sounds. The speakers are basically inhabitants of four districts: Sukkur, Ghotki, Shikarpur and Khairpur. The fundamental prosodic information i.e. pitches, and durations are measured from the recorded sounds using the PRAAT speech analyzer. The calculated prosodic information is mandatory for the development of speech processing software applications [11]. The back propagation Neural Network (NN) approach implemented by [12] is also used for the validation of the prosodic information. This network required three parameters such as input values, output value and the targeted values; such information is taken from the computed prosodic information of the recorded Sindhi sounds.

2. Literature Survey

Sindhi phonemes have received a great deal of research interests and undoubtedly, they have received the attention of a large population of researchers. Problems of Sindhi phonology are well observed and discussed by [13]. Their observation claims the strain and vocal inflection of Sindhi language and its 6 dialects having variation in several aspects of language. The comparison is brought about between the accents of people following different dialects. The comparison was carried out on the waveform picturing of the image to resolve the discovered problems. In [2,14] has also worked on the same subject with the addition of a letter to sound conversion. In this research, the concentration is given more the demonstration of f0 peak of variant syllables where long and short vowels are used at different places within words.

In [15–16] produced a piece of research on the analysis of the fundamental frequency of Sindhi language. The investigation is centered on the pitch working in between the intonation and stress. The accent of the pitch is examined while observing the recorded sounds of 69 words. These 69 words having different syllables were processed through digital experiments. The final outcomes of the research witnessed that the stress is directly orthogonal to f0 contours.

In [17] investigated the consonant sounds through acoustic analysis of Sindhi language. Mostly VCV formats were collected in the sound forms for the implementation of experiments. The researchers have focused mostly on the liquid consonants and the emphasis is put onto the difference between a trill and lateral consonants. In addition to this research, in [18] has examined the variation of vocalic features in vowels. Specific consideration is given to the differences among the languages spoken in Pakistan. They also include Sindhi phonology in their research. The experiments found the variation in vowels particularly comparing Sindhi ones with those of other spoken languages in Pakistan. The experiments were performed using PRAAT speech analyser.

3. Research Methodology

An accomplishment of the project of prosody generation of Sindhi language requires various steps, calculations and analysis. The core milestones of the research methodology are described in Figure 1. The first phase and foremost phase is preferences the speakers and their recorded sounds in order to get the required results. For this, we have preferred 228 graduate students of our University who belong to four different districts of Sindh province and having distinct pronunciation of Sindhi words.

Moreover, some descriptive sentences are compulsory for recordings and vice versa. Hence 81 sentences were composed, and 10 randomly selected sentences given to the selected speakers for the recording of their sounds. Furthermore, speech corpus was prepared and various segmented forms of sounds such as phoneme, syllable, words and sentences in binary formats were stored into the separate files using SQL. After that, the pitch and duration of each sound are obtained using the PRAAT speech analyzer. It was investigated digitally and measured for further use. The data sets of pitches and durations are also prepared for getting the input, output and target values for NN.

A back propagation NN model is proposed as the next step of the research methodology in which the number of inputs and the hidden layers were fixed according to the requirements of the prosody generation. The proposed network is simulated using the Matlab language. The next step is to train and test the proposed NN with different inputs of the Sindhi prosodic information as well as boundaries. The last phase of this project is to evaluate the performance of the developed NN model.

4. Speech Collection and Storing Procedure

Duration and Pitch of the recorded voices are needed to be computed. Hence, appropriate speakers required to accomplish this task [19]. The graduate students of our University

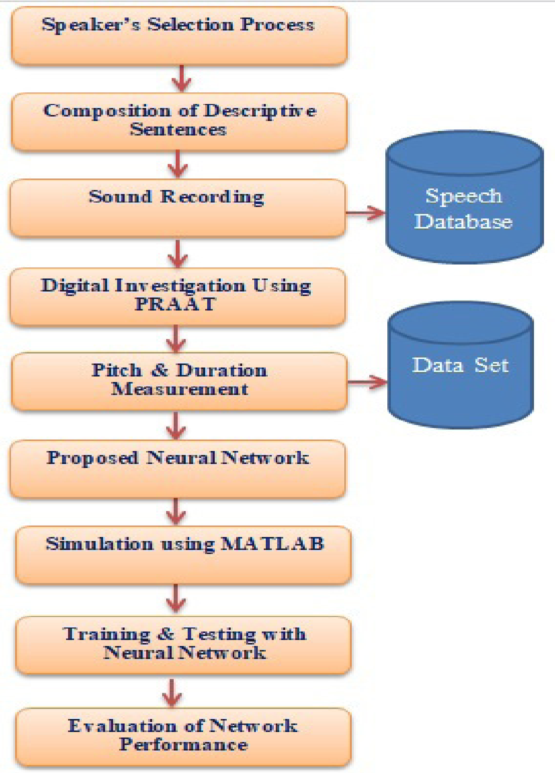

Speaker's Selection Process
Composition of Descriptive Sentences
Sound Recording
Speech Database
Digital Investigation Using PRAAT
Pitch & Duration Measurement
Data Set
Proposed Neural Network
Simulation using MATLAB
Training & Testing with Neural Network
Evaluation of Network Performance

FIGURE 1. Milestones of research methodology.

having ages of 20, 21 and 22 years are preferred for recording their voices. This research is specifically highlighted the prosodic analysis of the voices of the undergraduates studying in this university but actually belonging to four Districts of province Sindh. The number of preferred speakers of each District is given in Table 1. From District Sukkur, 51 speakers are selected. Whereas 54 and 58 speakers are chosen from the Districts Shikarpur and Ghotki respectively, 65 speakers are preferred from Khairpur.

The Radio Station Khairpur is the place where the sounds of the selected speakers have been recorded. The entire recording and storing process is identical as described by [10]. The prosodic information exists in recorded sounds intensely considered while the development of the speech corpus with the 16-bit encoding and the 16-kHz sampling rate. Among the 81 communicative composed sentences, 10 sentences were erratically given to the selected speakers along with the Sindhi prosodic restrictions. There are 228 speakers and 10 descriptive sentences so that the total obtained sounds of sentences are 228 × 10 = 2280 and the segmented words are 11348. The overall syllables are 21673, the syllables in sentences vary from 3 to 12. The number of phonemes is 53491 varies from 12 to 28. The sounds of these all segmented sounds are used for experimentation with the back propagation NN in order to generate the prosodic information.

The letter-based length of the words plays a momentous role in the computation of duration and pitch of the recorded sounds. Sindhi words which are based on two to five

TABLE 1. District and age-wise selected speakers

Districts	Age group (years)	No. of speakers
Sukkur	20	16
	21	18
	22	17
Shikarpur	20	20
	21	19
	22	15
Ghotki	20	14
	21	21
	22	23
Khairpur	20	19
	21	26
	22	20
Total number of speakers		**228**

characters are selected for the investigation. The recorded sounds are segmented and characterized by distinct values using the PRAAT speech analyzer tool commonly used by researchers like in [20] due to some specific speech analysis characteristics. In the speech database, letters based words taxonomy is essentially considered and segmented sounds are individually stored according to the mentioned taxonomy. The pitch and duration calculation process of the recorded sounds is described in [10]. One or more prosodic information is also calculated by various other researchers like in [21] for different applications. The duration and pitch of the recorded word are shown in seconds and Hz respectively.

5. Investigated Pitches and Durations

The results in terms of the pitch and duration of the recorded sounds are obtained and synthesized using the PRAAT speech analyzer. The huge numbers of calculations are received during the digital investigation of the sounds hence, the Mean values of pitches in Hz and durations in ms are computed and summarized the outcomes. The recorded sounds of the experimented words are divided into 2 Letter Words (2LW), 3 Letter Words (3LW), 4 Letter Words (4LW) and 5 Letter Words (5LW) to measure the pitch and duration of the selected speakers living from the four Districts of province Sindh: Sukkur, Shikarpur, Ghotki and Khairpur. Table 2 presents the calculated pitches and durations of all recorded sounds of words which are composed of 2 to five Sindhi letters.

After the comparison of the lowest and highest pitches of the recorded sounds, it is found that the inferior Mean Pitch of 127.76 Hz is received with the sounds of 5LW. After that, the lowest Mean Pitch of 136.51 Hz is calculated with the sounds of 2LW. Additionally, the lowest Mean Pitches of 142.54 Hz and 147.46 Hz are measured with the recorded sounds of 3LW and 4LW respectively. Furthermore, the uppermost pitches are also evaluated and observed that the highest Mean Pitch of 147.98 Hz is calculated from the recorded sounds

TABLE 2. Mean pitches and durations calculated from recorded sounds of all letter words

Districts	Age groups	n	μ Pitch in Hz				μ Duration in ms			
			2LW	3LW	4LW	5LW	2LW	3LW	4LW	5LW
Sukkur	20 Years	16	136.55	143.22	147.49	127.78	0.2162	0.2976	0.3832	0.3955
	21 Years	18	136.56	143.26	147.56	127.81	0.2162	0.2977	0.3837	0.3956
	22 Years	17	136.61	143.29	147.58	127.82	0.2163	0.2979	0.3835	0.3956
Shikarpur	20 Years	20	137.17	143.59	147.61	127.88	0.2211	0.2981	0.3943	0.3962
	21 Years	19	137.23	143.59	147.63	127.89	0.2215	0.2984	0.3945	0.3965
	22 Years	15	137.24	143.61	147.64	127.88	0.2216	0.2988	0.3946	0.3966
Ghotki	20 Years	14	137.82	144.78	147.93	128.14	0.2399	0.2993	0.3999	0.3986
	21 Years	21	137.83	144.83	147.97	128.16	0.2402	0.2993	0.4006	0.3987
	22 Years	23	137.87	144.80	147.98	128.15	0.2403	0.2994	0.4004	0.3989
Khairpur	20 Years	19	136.56	142.54	147.53	127.78	0.2162	0.2977	0.3831	0.3959
	21 Years	26	136.59	142.59	147.54	127.76	0.2164	0.2977	0.3832	0.3957
	22 Years	20	136.61	142.64	147.57	127.78	0.2165	0.2978	0.3834	0.3961

of the 4LW. The next crest Mean Pitch of 144.83 Hz is obtained with 3LW. After that, the Mean Pitches of 137.87 Hz and 128.17 Hz are received with the recorded sounds of the 2LW and 5LW respectively.

The collected Mean Durations of the recorded sounds are also investigated in terms of the stumpy and soaring positions. Among the all Mean Durations of the sounds 2LW, the minimum value of 0.2160 ms is calculated. The lowest Mean Duration of 0.2974 ms is computed from the recorded sounds of the 3LW. The inferior Mean Durations of 0.3828 ms and 0.3952 ms are received from the sounds of 4LW and 5LW respectively. With the sounds of the 2LW, the maximum Mean Duration of 0.2403 ms is assessed. The highest Mean Duration of 0.2994 ms is calculated with the sounds of 3LW. Using the recorded sounds of 4LW and 5LW the maximum Mean Durations of 0.4006 ms and 0.3989 ms are computed.

6. Performance Evaluation Using Neural Network

When the recorded sentences segmented into words, syllables and phonemes, the process of getting prosodic information and classification began using NN. Basically, two approaches were decided to implement for extracting maximum prosodic information through which comparison and statistical analysis done. One such approach used for this purpose is NN. Following the feed forward back propagation method, a neural network is developed in Matlab by setting 4 inputs, 2 output targets as depicted in Figure 2.

Order to train as well test the network, a training dataset is required which is made in Excel sheet with predefined targeted parameters based on received results of recorded speech individuals i.e. Sentence, Word, Syllable and Phoneme as shown in Figure 3. To train the network, dataset imported in Matlab by creating two arrays i.e. 4 × 68 for input values and 2 × 68 for output values of Pitch and Duration. The developed network trained on the selected parameters such as network type, number of hidden layers, number of input and output neurons and the number of iterations with the learning rate of 0.001.

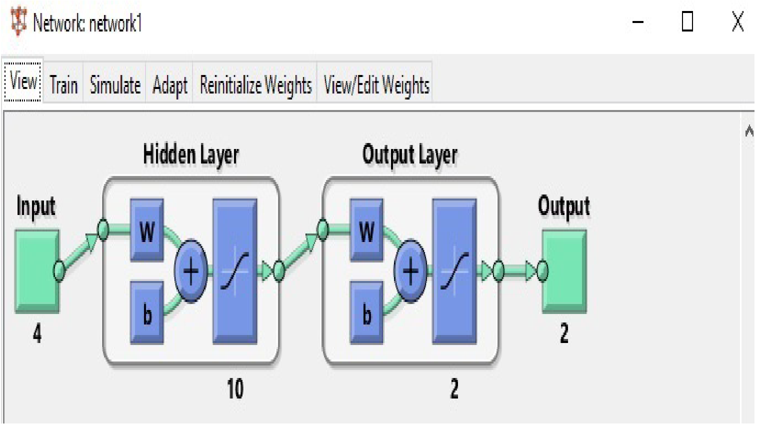
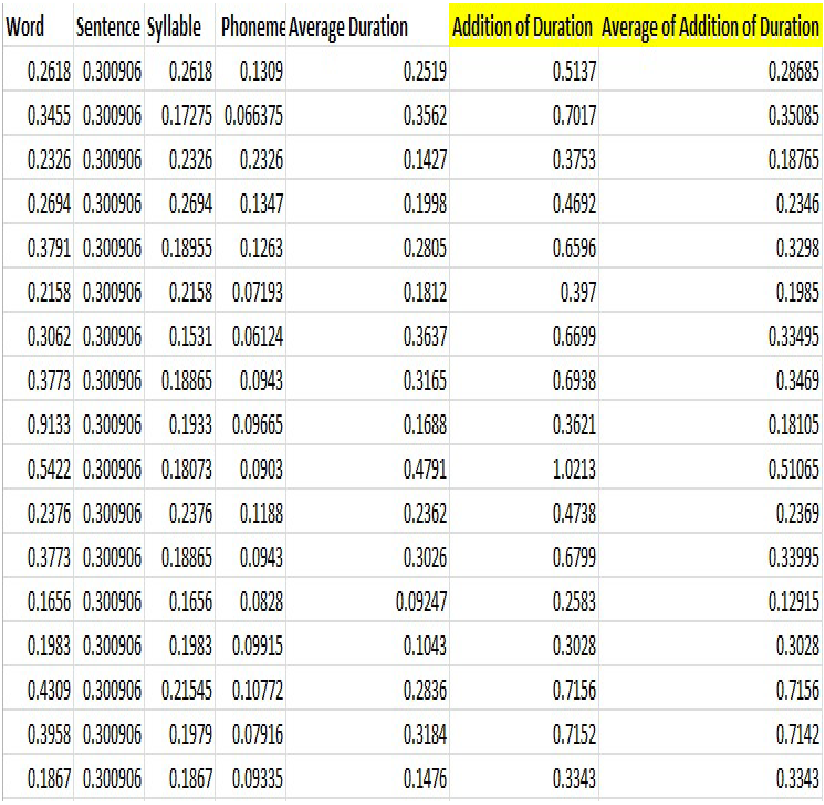


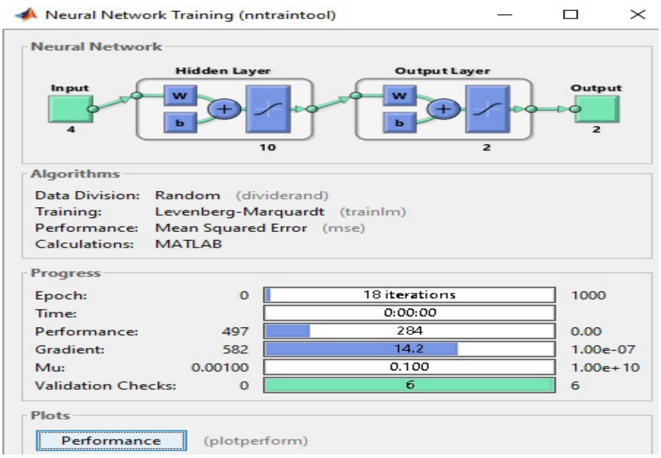
FIGURE 2. Developed a neural network.



Word	Sentence	Syllable	Phoneme	Average Duration	Addition of Duration	Average of Addition of Duration
0.2618	0.300906	0.2618	0.1309	0.2519	0.5137	0.28685
0.3455	0.300906	0.17275	0.066375	0.3562	0.7017	0.35085
0.2326	0.300906	0.2326	0.2326	0.1427	0.3753	0.18765
0.2694	0.300906	0.2694	0.1347	0.1998	0.4692	0.2346
0.3791	0.300906	0.18955	0.1263	0.2805	0.6596	0.3298
0.2158	0.300906	0.2158	0.07193	0.1812	0.397	0.1985
0.3062	0.300906	0.1531	0.06124	0.3637	0.6699	0.33495
0.3773	0.300906	0.18865	0.0943	0.3165	0.6938	0.3469
0.9133	0.300906	0.1933	0.09665	0.1688	0.3621	0.18105
0.5422	0.300906	0.18073	0.0903	0.4791	1.0213	0.51065
0.2376	0.300906	0.2376	0.1188	0.2362	0.4738	0.2369
0.3773	0.300906	0.18865	0.0943	0.3026	0.6799	0.33995
0.1656	0.300906	0.1656	0.0828	0.09247	0.2583	0.12915
0.1983	0.300906	0.1983	0.09915	0.1043	0.3028	0.3028
0.4309	0.300906	0.21545	0.10772	0.2836	0.7156	0.7156
0.3958	0.300906	0.1979	0.07916	0.3184	0.7152	0.7142
0.1867	0.300906	0.1867	0.09335	0.1476	0.3343	0.3343

FIGURE 3. Training dataset.

Based on 4 predefined values for input such as sentence, word, syllable and phoneme and using the set parameters, network trained to generate the outputs of pitch and duration. As shown in Figure 4, the developed network during the execution process generated the best performance result at the 18th epoch out of 100 epochs in the training process. For batch



Neural Network Training (nntraintool)

Neural Network

Algorithms

Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MATLAB

Progress

Epoch:	0	18 iterations	1000
Time:		0:00:00	
Performance:	497	284	0.00
Gradient:	582	14.2	1.00e-07
Mu:	0.00100	0.100	1.00e+10
Validation Checks:	0	6	6

Plots

Performance (plotperform)

FIGURE 4. Execution process during network training.

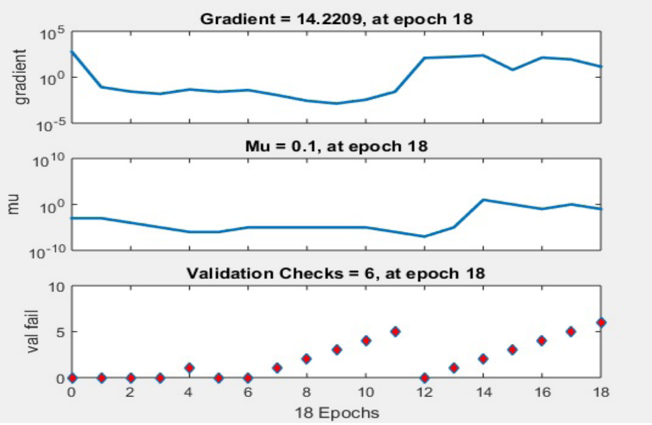


FIGURE 5. Gradient results.

training, all of the training samples pass through the learning algorithm simultaneously in one epoch before weights are updated. While Figure 5 shows the gradient ratio at the 18th epoch, Mµ defines the control parameter for the algorithm which is 0.1 used to train the neural network and total validation checks up to the 18^{th} epoch.

Admirable results received during the training process of a developed network. The testing phase performed after the training process on the selected dataset and predefined parameters. It is observed that the network generated the best validation performance result of 198.8672 at the 12^{th} epoch against the Mean Square Error during the testing phase as depicted in Figure 6. The performance and results throughout the process using an above-defined dataset for input and desired targeted output are depicted in Figure 7 as it shows that the developed system produces 98.9% accuracy in the training process and 99% generated during the system validation process. However, the result generated during the testing phase is 98.8% which is quite similar to the training results and the overall accuracy achieved by the developed system is also 98.8%. The obtained results of pitch and duration

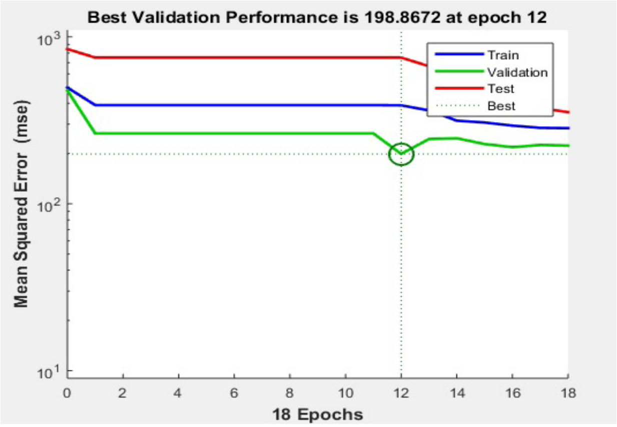


FIGURE 6. Best performance result.

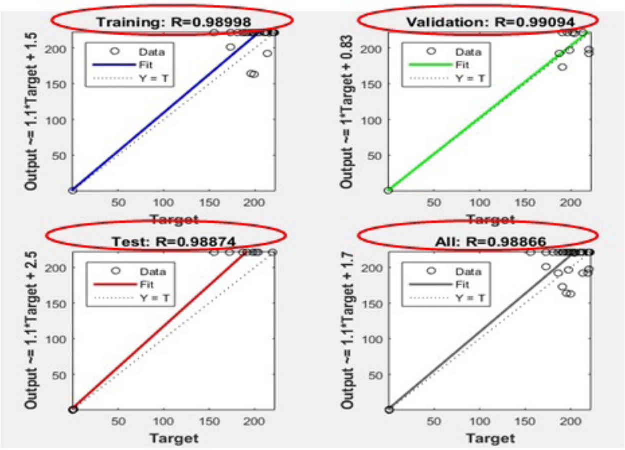


FIGURE 7. Overall system accuracy.

have been evaluated with the error rate provided by the developed Neural Network and with the standard deviation technique to find the variation of pitch between the male and female genders having different age groups.

7. Conclusion

Automatic speech analysis and synthesis still required some research efforts particularly for the development of Arabic Script-Based speech processing applications. The prosody is correlated with the information having dissimilar temperaments such as linguistic rules, complications and variations in the manifestation of the sounds hence, prosodic features should be exercised at the maximum level to achieve the required results from the speech processing applications. The Sindhi prosodic information is generated and analyzed specifically pitch and duration from the recorded Sindhi sounds using the back propagation

NN. The prosodic information of Sindhi sounds is obtained with two methods. At first, the PRAAT speech analyzer is used to obtain the results. Secondly, a back propagation NN model is proposed and implemented on the 2280 sounds of the sentences collected from 228 speakers living in the four districts of province Sindh to validate the generated Sindhi prosodic information. The experiments performed on the pitch and duration data sets collected from the recorded sounds. For getting the acceptable level of accuracy, multiple layers and 100 epochs were used which gives 98.8% accuracy at the 18th epoch. The generated Sindhi prosodic information and adopted research methodology will be supportive to the scholars of Sindhi speech processing applications.

References

1. Hassan A. Assimilation and incidental differences in Sindhi language. *Eurasian Journal of Humanities*. 2016; 2(1), 1–16.
2. Mahar JA, Memon GQ. Phonology for Sindhi letter-to-sound conversion. *Journal of Information & Communication Technology*. 2009; 3(1), 11–21.
3. Shaikh H, Mahar JA, Mahar MH. Statistical approaches to instant diacritics restoration for Sindhi Accent prediction. *Sindh University Research Journal (Science Series)*. 2017; 49(2), 363–366.
4. Hoffmann S, Pfister B. Employing sentence structure: syntax trees as prosody generators. 13th Annual conference of the international speech communication association. 2012, 470–473.
5. Dasgupta PB. Detection and analysis of human emotions through voice and speech pattern processing. *International Journal of Computer Trends and Technology*. 2017; 52(1), 1–3.
6. Liu H, Lu H, Shao X, Xu Y. Model-based parametric prosody synthesis with deep neural network. INTERSPEECH: San Francisco, 2016, 2313–2317.
7. Thakuria LK, Acharjee P, Das A, Thakdar PH. Integrating rule and template based approaches to prosody generation for emotional BODO speech synthesis. 4th international conference on communication systems and network technologies, Bhopal, India. 2014.
8. Sheikhan M. Hybrid of evolutionary and swarm intelligence algorithms for prosody modeling in natural speech synthesis. *International Journal of Information & Communication Technology Research*. 2016; 8(2), 33–44.
9. Chen J, Liu Y, Zhang Z, Fan C, Ding Y. Text-Driven Visual Prosody Generation for Embodied Conversational Agents. 19th ACM international conference on intelligent virtual agents. 2019, 108–110.
10. Mahar SA, Mahar MH, Danwar SH, Mahar JA. Investigation of pitch and duration range in speech of sindhi adults for prosody generation module. *International Journal of Advanced Computer Science and Applications*. 2019; 10(9), 187–195.
11. Waghmare K, Kayte S, Gawali B. Analysis of pitch and duration in speech synthesis. *Communications on Applied Electronics*. 2016; 4(4), 10–18.
12. Magsi A, Mahar JA, Danwar SH. Date fruit recognition using feature extraction techniques and deep convolutional neural network. *Indian Journal of Science and Technology*. 2019; 12(32), 1–12.
13. Shaikh H, Mahar JA, Malah GA. Digital investigation of accent variation in Sindhi dialects. *Indian Journal of Science and Technology*. 2013; 6(10), 5429–5433.
14. Mahar JA, Memon GQ, Shah HA. Perception of syllables pitch contour in Sindhi language. IEEE international conference on natural language processing and knowledge engineering. 2009, 593–597.

15. Abbasi AM, Hussain S. The role of pitch between stress and intonation in Sindhi. *ELF Annual Research Journal*. 2015; 17, 41–54.
16. Abbasi AM, Pathan H, Channa MA. Experimental phonetics and phonology in Indo-Aryan & European languages. *Journal of Language and Cultural Education*. 2018; 6(3), 21–52.
17. Keerio A, Channa N, Malkani YA, Qureshi B, Chandio J. A. Acoustic analysis of the liquid class of consonant sounds of Sindhi. *Sindh University Research Journal (Science Series)*. 2014; 46(4), 505–510.
18. Farooq M. Acoustic analysis of corner vowels in six indigenous languages of Pakistan. *Journal of Research in Social Sciences*. 2018; 6(2), 2305–6533.
19. Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*. 2018; 21(1), 93–120.
20. Lathadevi LT, Guggarigoudar SP. Objective acoustic analysis and comparison of normal and abnormal voices. *Journal of Clinical and Diagnostic Research*. 2018; 12(12), 1–4.
21. Sarasola X, Navas E, Tavarez D, Serrano L, Saratxaga I, Hernaez I. Application of pitch derived parameters to speech and monophonic singing classification. *Applied Science*. 2019; 9(3140), 1–16.