Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 07, July-2018

Artificial Neural Network Based Predicting Parameters of Emotional Speech from Neutral Speech

Niha Sultana
Dept of ECE, Mtech in Signal Processing
Vidyavardhaka College of Engineering
Mysore-570002, India

Geethashree A
Associate Professor, Dept of ECE
Vidyavardhaka College of Engineering
Mysore-570002, India

***Abstract*- Artificial Neural Network (ANN) based model is used for predicting the prosodic parameters of the target emotions from trained data. Standard multi-layer, feed-forward, back-propagation neural networks with architecture have been designed using MATLAB Neural Network tool. The work is evaluated in two phases, namely training and testing phase. Data for both training and testing are extracted from praat (A speech analyzing tool). The input parameters for the network are pitch, intensity and duration (P, I, D) of neutral speech while the (P, I, D) of emotional speech is target parameters. The predicted output plotted against target parameters given to the neural network forms the plot of regression which is later evaluated as a part of result.**

Keywords—Artificial Neural Network (ANN); Levenberg Marquant (LM) ; Multilayer Perceptron (MLP); Praat;

I. INTRODUCTION

Speech plays an important role whenever an information is being conveyed by the speaker. Human speech is a complex signal that contains information about the speaker identity, linguistic message, language type and emotion as well. Crying, screaming, dancing and laughing are the few ways in which human expresses their emotion. Expressing speech through different tone impacts the emotion of a person and it varies from person to person. Physiological changes in the body, expressive behaviors and subjective feelings also effects the human emotions. These changes in speech signal are mainly observed in prosodic parameters such as pitch, duration and intensity.

Emotional speech conversion is the process of transforming speech from one emotional state into another emotion state, thereby maintaining the speaker's identity and information being conveyed. Emotional voice conversion is one of the most challenging tasks in a speech signal processing domain. A technique to incorporate emotion into neutral voice is to alter emotion specific features to a neutral speech. Target emotions in the proposed emotion conversion framework are fear, happy, anger and sad. Multi-layer feed-forward neural network models are explored for mapping the prosodic parameters between neutral and target emotions. Predicted prosodic parameters of the target emotion are obtained once the neural network is trained. After incorporating the emotion-specific prosody, the regression plot between the actual output and predicted output is evaluated.

Emotion conversion concept has interesting applications such as making use of emotion-based robots for defense purpose and in warfare etc. thus expressing the feelings emotionally according to situation will be more effective.

II. ARTIFICIAL NEURAL NETWORK

An artificial neural network is made up of many artificial neurons that are joined together depending on the specific network architecture. The objective of the neural network is to convert the inputs into desired outputs. A neural network is a massively parallel distributed processor made up of simple processing units that have a natural tendency for acquiring, storing and utilize experiential knowledge that has been related to the networks capabilities and performances that makes it available for us.

The most useful neural networks in function approximation are Multi-Layer Perceptron (MLP). Fig. 1 shows an MLP consists of an input layer, several hidden layers, and an output layer. Neurons in input layer act as buffers for distributing the input signals.

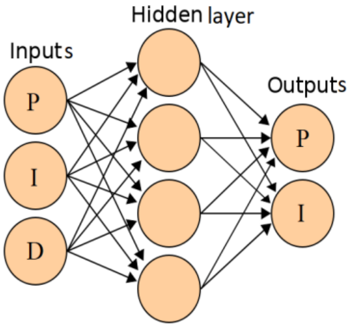


Fig. 1. Model of Artificial Neural Network.

III. SPEECH DATABASE

Different set of sentences is recorded in 5 different emotions namely Neutral, Angry, Sad, Fear and Happy. These sentences are being used for the training and testing purposes. Speech data of few female and few male speakers were recorded in noise free environment. Data used for recording are short sentences usually containing 3 to 4 phrases. Speech data are recorded with bit resolution equal to 16bits per sample and sampling rate of 44100samples/second. The prosodic parameters for these N inputs and N outputs are extracted using praat.

IV. PROPOSED METHOD

The proposed model consists of three stages.

A. *Parameters Extraction of Speech Signal*

Recording and saving data is the first step in designing models. Thus, parameters like intensity, duration and pitch of speech signal are extracted using Praat software. Spectrum of speech signal is evaluated in praat as shown in Fig. 2.

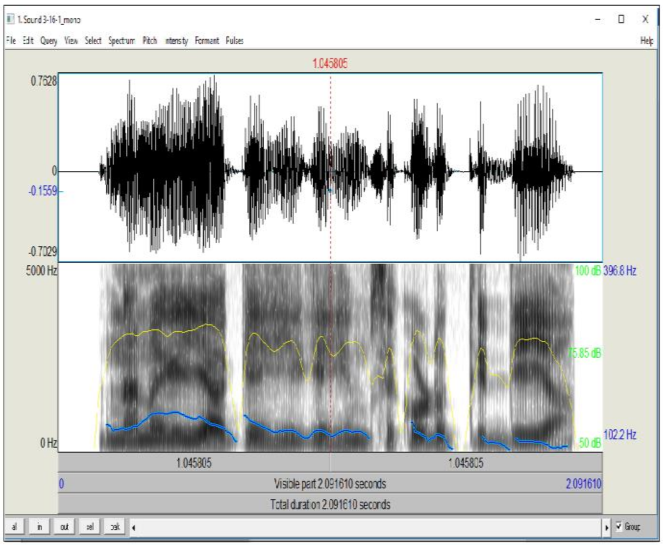


Fig. 2. Spectrogram of Speech Signal in Praat

B. *Conversion Algorithm using Matlab fitting tool*

After data collection, the procedure is carried out by three main steps in training ANN.

- Building the network

  During this stage, the designer specifies the neurons in each layer, transfer function in each layer, number of hidden layers, weight/bias learning function, training function, and performance function. Fig. 3 shows the design of a network. Multilayer perceptron (MLP) networks is used in this work.

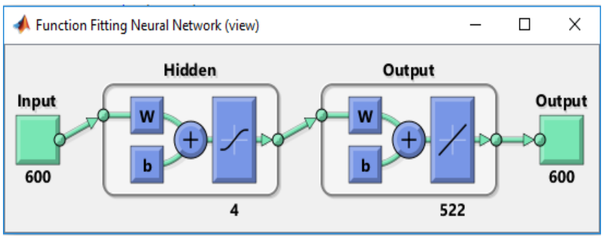
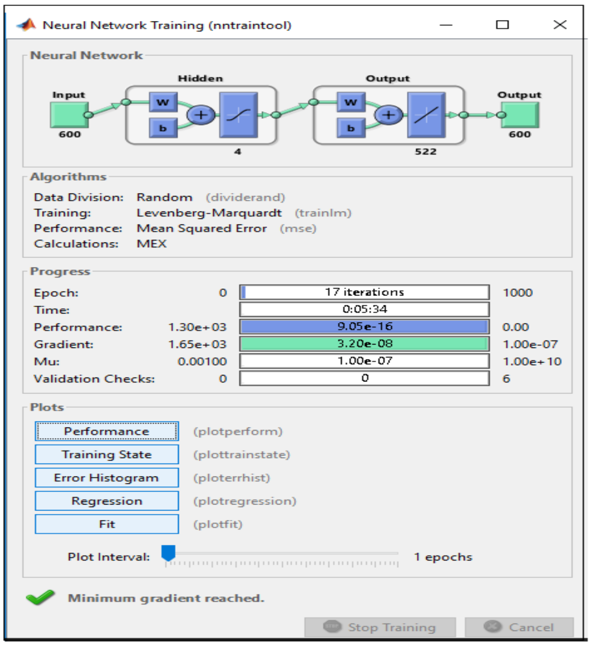


Fig. 3. Design of Neural Network

- Training the network

  Fig. 4 shows the neural network training (nntraintool). During training, the weights of the design network are made to adjust such that the predicated outputs of the network be very close to the target (measured) outputs of the network.



Algorithms

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

Progress

Epoch:	0	17 iterations	1000
Time:		0:05:34	
Performance:	1.30e+03	9.05e-16	0.00
Gradient:	1.65e+03	3.20e-08	1.00e-07
Mu:	0.00100	1.00e-07	1.00e+10
Validation Checks:	0	0	6

Plots

Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)
Fit (plotfit)

Plot Interval: 1 epochs

✓ Minimum gradient reached.

Fig. 4. Neural Network Training

The trained network is tested by evaluating the regression plot. Regression value (R-value) is a correlation coefficient which measures the correlation between predicted output and target output. Regression plot of neutral to other emotion is shown in Fig. 5 and Fig. 6 for male voice and female voice training respectively. The dashed line in each plot represents Output = (target output - predicted output). If R value is close to 1, it means exact linear relationship between the predicted and the target values. In the analysis we noticed that regression plot value for male voice is exactly equal to 1 compared to female voice.

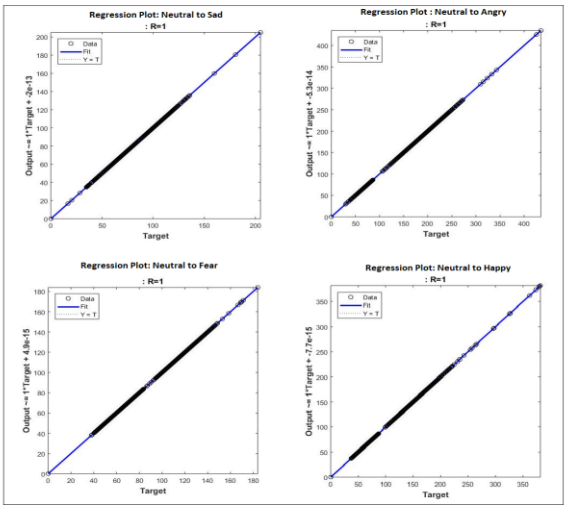


Fig. 5. Regression plot for Training data (Male voice)

- Testing the network

  During testing, different unknown data of neutral speech signal are applied to the model to obtain testing output of ANN model.

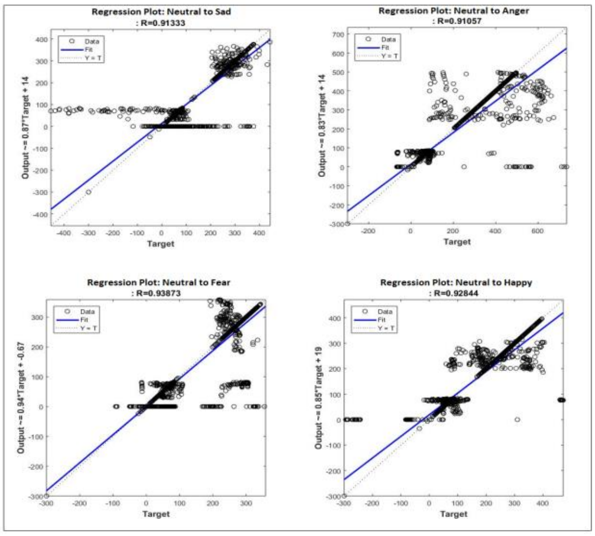


Fig. 6. Regression plot for Training data (Female voice)

C. Reconstruction of emotional speech signal

The testing output obtained from ANN model, is reconstructed into speech signal using Praat script.

V. RESULTS

In the analysis of speech signal, we observed that the duration of neutral emotion is shorter compared to fear and sad emotion and lengthy compared to angry and happy emotion.

The network is tested using 3 different sentences of male voice and 3 different sentences of female voice. Table 1 and Table 2 shows the table of regression value for these 3 different sentence.

TABLE I Regression value obtained for Male voice

Emotion Conversion	Male voice		
	S1	*S2*	*S3*
Neutral to Sad	0.8313	0.6560	0.8238
Neutral to Anger	0.8590	0.8972	0.8747
Neutral to Fear	0.6742	0.7243	0.7256
Neutral to Happy	0.9127	0.7532	0.8213

TABLE II Regression value obtained for Female voice

Emotion Conversion	Female voice		
	S1	*S2*	*S3*
Neutral to Sad	0.8367	0.8187	0.8743
Neutral to Anger	0.7491	0.7965	0.6574
Neutral to Fear	0.7157	0.6517	0.6642
Neutral to Happy	0.8097	0.7763	0.6832

VI. CONCLUSION

The results of this study indicate that the potential of Artificial Neural Network Fitting tool for prediction of different emotion from neutral speech using pitch, intensity and duration of desired speech signal. During reconstruction of speech signal, we noticed that male voice conversion is more accurate compared to female voice.

The correlation value R varies between 92% to 65% showing good agreement between the measured values and predicted ANN values.

REFERENCES

[1] Amrita and Bageshree Pathak, “Emotion Conversion of Speech Signal using Neural Network,” International Journal of Electrical, Electronics and Data Communication, ISSN: 2320-2084, Volume-2, Issue-9, Sept.-2014

[2] Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W Black and Kishore Prahallad, “Voice conversion using artificial neural networks”, publish in IEEE.

[3] Tao J, Yongguo K, and Li A, “Prosody Conversion from Neutral Speech to Emotional Speech”, IEEE Trans. Audio,Speech and Lang Proc., vol.14:1145–1153, 2006.

[4] Amit Kumar Yadav and Hasmat Malik, A.P. Mittal, “Artificial neural network fitting tool based prediction of Solar radiation for identifying solar power potential,” Journal of Electrical Engineering.

[5] Kurban Ubul1, Askar Hamdulla and Alim Aysa, “ A Digital Signal Processing Teaching Methodology Using Praat,” Proceedings of 2009 4th International Conference on Computer Science & Education.

[6] Md. Shafiqul Islama, Md. Monirul Kabira and Nafis Kabirb, “Artificial neural networks based prediction of insolation on horizontal surfaces for Bangladesh,” International Conference on Computational Intelligence, ELEVISER.

[7] Khaled Daqrouq, Ibrahim N.Abu-Isbeih and Mikhled Alfauori, “Speech Signal Enhancement Using Neural Network and Wavelet Transform”, 2009 6th International Multi-Conference on Systems, Signals and Devices.

[8] Jainath Yadav and K. Sreenivasa Rao, “Emotion conversion using Feedforward Neural Networks”.

[9] J. Nicholson, K. Takahashi, and R.Nakatsu, “Emotion recognition in speech using neural networks,” in 6th International Conference on Neural Information Processing, pp. 495–501, Aug. 1999.

[10] Zhaojie Luo1, Jinhui Chen, Toru Nakashika, Tetsuya Takiguchi1 and Yasuo Ariki1, “Emotional Voice Conversion Using Neural Networks with Different Temporal Scales of F0 based onWavelet Transform”. 9th ISCA Speech Synthesis Workshop • September 13 – 15, 2016 • Sunnyvale, CA, USA.

[11] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features,” Am. J. Signal Process., vol. 2, no. 5, pp. 134–138, 2012.

[12] M. S. Suri, D. Setia, and A. Jain, “PRAAT Implementation For Prosody Conversion,” pp. 1–4, 2010.

[13] S. Bhutekar and M. Chandak, “Designing and Recording Emotional Speech Databases,” Int. J. Comput. Appl. Proc. Natl. Conf. Innov. Paradig. Eng. Technol. (NCIPET 2012), vol. 4, pp. 6–10, 2012.

[14] Z. Inanoglu and S. Young, “Emotion conversion using F0 segment selection,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 2122–2125, 2008.