Vowel Analysis for Identifying Expression of Emotions in Odia- Spoken-Language

Sanghamitra Mohanty
Department of Computer Sc. & Application,
Utkal University, Bhubaneswar-751004,
India

Basanta Kumar Swain
Department of Computer Science and Engineering,
Government College of Engineering Kalahandi,
Bhawanipatna- 766001, India

ABSTRACT

Speech can be described as an act of producing voice through the use of the vocal folds and vocal apparatus to create a linguistic act for conveying intended information. Speech is a subtle and rich communication that transfers not only the linguistic information, but also information about the personality and the emotional state of the speaker. The objective of this study is to analyse the effect of vocal characteristics like duration of vowels, mean pitch, jitter and shimmer of vowels in different Oriya spoken emotional speech uttered by female & male speakers. This paper also used FFBPNN as a classifier to identify an emotion out of five emotional categories such as Normal, Santa, Karuna, Raoudra, Shringar using Jitter and Shimmer variants as feature vectors. The supervised classifier achieved 86.88% accuracy rate in vowel domain. The emotional speech corpus is recorded from several semi-professionals actors within range of 22-35 years old.

Keywords

FFBPNN, Jitter, Shimmer, Emotion

1. INTRODUCTION

The research of spoken language is dramatically increased during the last decade due to the increase of speech application in many fields like virtual reality, video-games, human-computer speech interaction, security, and medical applications. We use emotions to express and communicate our feelings in everyday life. Our experience as speakers as well as listeners tells us that the interpretation of meaning or intention of a spoken utterance can be affected by the emotions that are expressed and felt. With recent advances in man-machine communication technologies, through automatic spoken dialog management systems, the question of how human emotion is encoded by a speaker and decoded by a listener has now attained practical importance [1].

It has long been known that speech prosody, that is patterns in pitch and amplitude modulation and segmental durations carry emotional information in the acoustic speech signal [2]. The investigation of the phonetic characteristics of the emotional speech can be carried out only after making the proper classification of the emotional states.

Emotion classifications of the researchers differ according to the goal of the research and the field. Also the scientist’s opinion about the relevance of dividing different emotions is important. There is no list of basic emotions. However, it is possible to define list of emotions which are usually chosen as basic ones like erotic (love) (shringar), pathetic (sad) (karuna), wrath (anger) (roudra), quietus (santa), normal(neutral).

According to Paninian Pratishakhya emotions can be broadly classified as erotic (love) (shringar), mirth(happiness) (hAsya), pathetic (sad) (karuna), wrath (anger) (roudra), herosima (bIra), terror(fear) (bhayAnaka), disgusting (boredom) (bibhasta), marvellous (adbhuta), quietus (shAnta), motherly affection (bAtslya) and devotional (bhakti) [3,4].

This research paper deals with the effect of vocal characteristics of vowels in two folds. In the first approach we investigated the role of duration, mean pitch, jitter (local) and shimmer (local) of vowels in emotional voices spoken by male as well as female speakers. In second approach we have automated the process of identifying the emotions by using a sophisticated classification technique called back propagation feed forward neural network (BPFFNN). The training and testing of BPFFNN classifier is carried out by feeding the merged input feature vectors called variants of jitter and shimmer. In section 2, we discuss the emotional speech database and acoustic parameters used in this research work. In sect. 3, we present the essentials of our research methodology. In Section 4, we summarize the results. The final section includes a brief discussion and conclusions.

2. EMOTIONAL ODIA SPEECH CORPUS AND ACOUSTIC PARAMETERS

2.1. Emotional Speech Database

The Odia emotional database contains recordings with five emotions: Normal(neutral), Santa(quietus), Karuna(sad), Raoudra (anger), Shringar (love). The speakers are persons aged between 25-35 years, born and educated (higher education) in the eastern part of Odisha, without manifested pathologies. The persons have been previously informed about the objective of the project. The speakers are allowed to read the short sentences or phrases fragments, with five emotional states. The recordings use a sampling frequency of 16000 Hz, 16bits. The emotional states were recorded in a laboratory room with noise cancellation microphone and desktop computer. Manual monitoring is carried out to judge, whether the speaker delivered the speech in intended emotion or not and rerecording is permissible until desired emotion delivery is reached. The developed emotional speech corpus is initially used to identify the emotion by human beings those are not involved in recording phase. The recognition rate was 92% for the Normal tone, 89% for Karuna, 94% for Raoudra, and 87.3% for Shringar and 85% Santa.

2.2 Parameterization of Acoustics Signals

2.2.1 Duration

Utterance durations for vowels are measured from the corresponding label files produced by manual segmentation procedure. On average, utterance durations become longer when speech is emotionally elaborated. For this study, we have taken three male & three female speakers. The words & sentences are recorded at the bit rate of 16000Hz.The Vowels are extracted from the three zones i.e. CV,V,VC of words. CV stands for Consonant to Vowel transition, V for steady state vowel, VC for Vowel to Consonant transition [5,6].

2.2.2 Fundamental frequency (Pitch)

We calculated the pitch contours of each utterance using Praat speech processing software. Global level statistics related to F0 such as minimum, maximum, mean were calculated from smoothed F0 contours [7].

2.2.3 Jitter measurement

The pitch period varies over time, depending on the utterance, intonation, emotional state of the speaker, and other factors. The long term pitch variations are often intentional, and are due to changes in the intonation or transitions between phonemes.

2.2.4 Jitter (absolute)

Jitter (absolute) is the cycle-to-cycle variation of fundamental frequency i.e. average absolute difference between consecutive periods which is expressed as:

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1} |T_I - T_{i+1}| \qquad (1)$$

Where Ti are the extracted f0 period lengths and N is the number of extracted f0 periods.

2.2.5 Jitter (relative)

It is the average absolute difference between consecutive periods, divided by the average period and expressed as percentage:

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \qquad (2)$$

2.2.6 Jitter (rap)

It is defined as the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period.

2.2.7 Jitter (ppq5)

It is the five point period perturbation quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period.

2.2.8 Shimmer measurement

The amplitude of the air flow velocity within a glottal cycle may differ across consecutive pitch periods, this characteristics is called shimmer. Hence shimmer is the parameter which represents the amplitudes perturbation of the voice signal. In this research work we have considered various types of shimmer namely

2.2.9 Shimmer (relative)

Shimmer (relative) is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as percentage

$$shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \qquad (3)$$

2.2.10 Shimmer (db)

Shimmer (db) is described as the variability of the peak to peak amplitudes in decibels i.e. the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20:

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1} |20\log(A_{i+1}/A_i)| \qquad (4)$$

Where A_i are the extracted peak-to-peak amplitudes data and N is the number of extracted fundamental frequency periods.

2.2.11 Shimmer (apq3)

Shimmer (apq3**)** is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.

2.2.12 Shimmer (apq5)

It is defined as the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and it four closest neighbors, divided by the average amplitude.

2.2.13 Shimmer (apq11)

It is represented as the 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitudes of a period and the average of the amplitudes of it and its ten closest neighbors, divided by the average amplitude [8].

3. NEURAL NETWORK AS RESEARCH METHODOLOGY

The human brain is an organ of a massive network of about 10^{11} interconnecting neural cells called neurons which perform highly parallel computing. It is seen that human brain can accurately identify an emotion of a person with less effort. The brain is exceedingly robust and fault tolerant which has greatly intrigued scientists, leading to the development of the field of artificial neural networks (NN)[9]. Artificial neural networks borrow ideas from human beings neural network structures to develop better techniques in computing, artificial intelligence, data analysis, modeling and prediction. Researchers use many types of neural network for broad applications. In this research study, we have used feed forward back propagation neural network (FFBPNN) for classifying the emotions because they allow to take into account speech features information and to build complex decision regions [10].

A successful pattern classification methodology depends heavily on the particular choice of the features used by the classifier .The Back-Propagation is the best known and widely used learning algorithm in training multilayer feed forward neural networks. The feed forward neural net refer to the network consisting of a set of sensory units (source nodes) that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes. The input signal propagates through the network in a forward direction and on a layer-by-layer basis. Back propagation is a multi-layer feed forward, supervised learning network based on gradient descent learning rule. This FFBPNN provides a computationally efficient method for changing the weights in feed forward network, with differentiable activation function units, to learn a training set of input-output data. Being a gradient descent method it minimizes the total squared error of the output computed by the network [11].

Back propagation training begins with setting node weights and offsets to small random values. Then a training example and desired outputs are presented to the net. The actual outputs of the net are calculated, error signals are formed, and an iterative algorithm is used to adapt the node weights - starting at the output nodes and working back to the first hidden layer. An example and desired outputs are again presented to the net and the entire process is repeated until weights converge [12]. The aim is to train the network to achieve a balance between the ability to respond correctly to the input patterns that are used for training and the ability to provide good response to the input that are similar, called test set. FFBPNN architecture is show in the Fig 1.

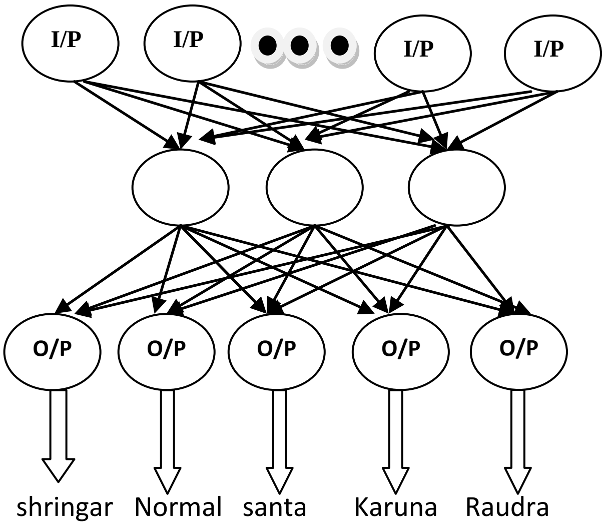

Fig 1. Architecture of FFBPNN

4. RESULTS

The experimental results are carried out by looking into multiple angles. Initially, we examined the role of individual acoustic parameters in determination of emotions from Odia speech in vowel domain and later we automated emotion identification process by adopting FFBPNN as classifier. The results are described in the following experimental sections.

Experiment 1 (based on Duration)

In this experiment, we focused on durational effect of vowels in emotional speech. Table 1 to Table 6 contains three male and female speakers vowel duration in Millie seconds.

Table 1. Vowel duration of male speaker1 in different emotions

Emotions	**Duration in Millie Seconds (male1)**				
	/a/	**/i/**	**/u/**	**/e/**	**/o/**
Normal	59	81	52	39	51
Santa	57	77	53	43	43
Karuna	42	87	53	36	49
Raoudra	54	56	38	44	39
Shringar	62	83	68	53	44

Table 2. Vowel duration of male speaker2 in different emotions

Emotions	**Duration in Millie Seconds (male-2)**				
	/a/	**/i/**	**/u/**	**/e/**	**/o/**
Normal	64	79	51	44	41
Santa	48	83	54	54	37
Karuna	105	103	64	61	46
Raoudra	79	61	42	44	59
Shringar	103	142	107	51	47

Table 3. Vowel duration of male speaker3 in different emotions

Emotions	**Duration in Millie Seconds (male-3)**				
	/a/	**/i/**	**/u/**	**/e/**	**/o/**
Normal	77	50	77	35	67
Santa	68	64	74	76	48
Karuna	171	129	67	71	67
Raoudra	59	33	49	59	69
Shringar	136	114	94	59	77

Table 4. Vowel duration of female speaker1 in different emotions

Emotions	Duration in Millie Seconds (female-1)				
	/a/	/i/	/u/	/e/	/o/
Normal	49	26	87	44	31
Santa	56	92	93	53	70
Karuna	150	117	133	150	89
Raoudra	50	28	56	39	49
Shringar	114	139	138	54	68

Table 5. Vowel duration of female2 speaker1 in different emotions

Emotions	Duration in Millie Seconds (female2)				
	/a/	/i/	/u/	/e/	/o/
Normal	49	96	52	38	45
Santa	56	135	87	37	47
Karuna	63	110	95	54	48
Raoudra	47	60	57	30	32
Shringar	68	206	118	58	59

Table 6: Vowel duration of female speaker3 in different emotions

Emotions	Duration in Millie Seconds (female-3)				
	/a/	/i/	/u/	/e/	/o/
Normal	51	69	60	38	29
Santa	47	71	58	40	34
Karuna	37	89	76	53	53
Raoudra	53	41	62	46	42
Shringar	57	139	99	53	34

It is observed from the above Table 1 to Table 6 that speech associated with the emotion "love(**Shringar**)" has high duration for female speakers where as the emotion "sad(**karuna**)" has high duration for male speakers.

Experiment 2(Based on Fundamental frequency)

Here we have emphasized over the mean pitch of different vowels in various emotions uttered by male and female speakers. Fig 2 demonstrates the mean pitch plot of five vowels uttered by two categories speakers.

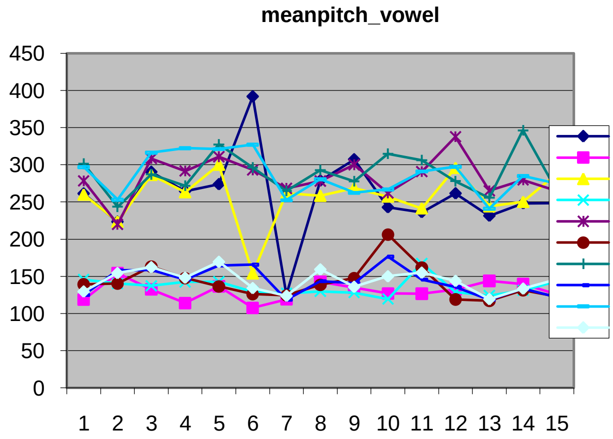

Fig 2: Mean pitch of vowels of speakers in different emotions

From the above Fig 2, the mean pitch for male speaker associated with emotion **karuna** for vowel /i/ has dominance where as in female the speech associated with emotion **love** for vowel /a/ plays dominant role.

Experiment 3 (Based on Jitter)

This experiment is carried out by extracting local jitter features using Praat speech processing tool. We plot the local jitter values of five vowels in different emotions which are shown in Fig 3.

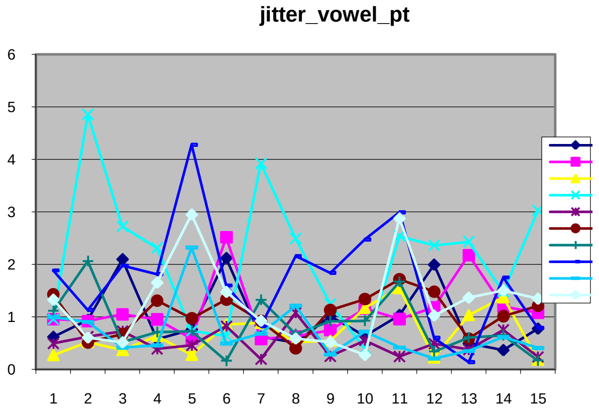

Fig 3: Jitter (local) for vowels in different emotions

According to Fig 3, the emotion "anger" of vowel /i/ is having a dominant role in jitter for male speaker. The jitter value for female speaker has dominant vowel /u/ for the emotion "love".

Experiment 4 (Based on Shimmer)

This experiment signifies the role of local shimmer acoustic parameter of vowels in various emotional Odia speech. The local shimmer values are extracted using Praat speech processing tool. Fig 4 visualizes the local Shimmer effects on different emotional vowels.

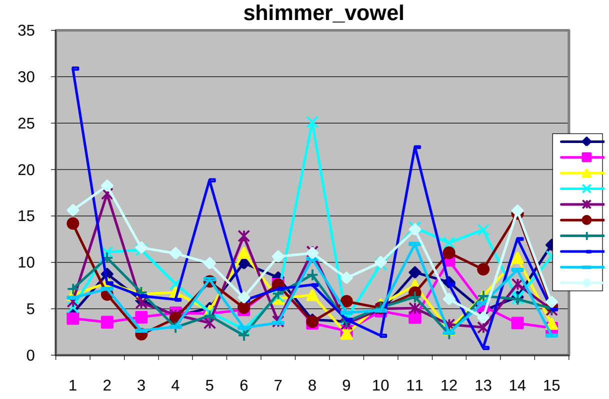

Fig 4: Shimmer (local) for vowels in different emotions

We observed from Figure4 that the Shimmer (local) in the emotion "anger" of vowel /o/ has dominant role & for female in anger of vowel /i/ has dominant role.

Experiment 5

In this research, we used two-hidden-layers feed-forward back propagation neural networks. We experimented with many different sizes of the hidden layer, i.e., 5 to 20 neurons. In most cases, the training accuracy flattened out with hidden-layer count of 10. The training iterations (*epochs*), however, did reduce with higher number of hidden layer neurons. The FFBPNN used 4 jitter variants and 5 shimmer variants as feature vectors for individual vowel (/a/,/i/,/u/,/e/,/o/) which leads to generate a feature vector magnitudes of length 65 of each emotion category. This feature vector of length 65 is fed as inputs, and produced 5 separate predictions (one for each emotion). That means, we had 65 neurons in the input layer, and 5 neurons in the output layer. From the overall dataset of 1600 emotional vowel speech samples (spoken by 12 male and 8 female speakers), we used 90% as the training set, and set aside the remaining 10% as validation/testing set. The learning rate of 0.3, and the training error tolerance of 0.01 were used. Epoch count was limited to 3000. Fig 5, Fig 6 demonstrate the trained network, comparison between actual target and predicted output respectively.

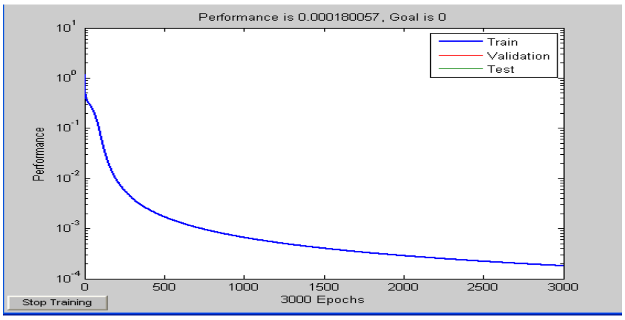

Fig 5: Performance of FFBPNN after 3000 epochs

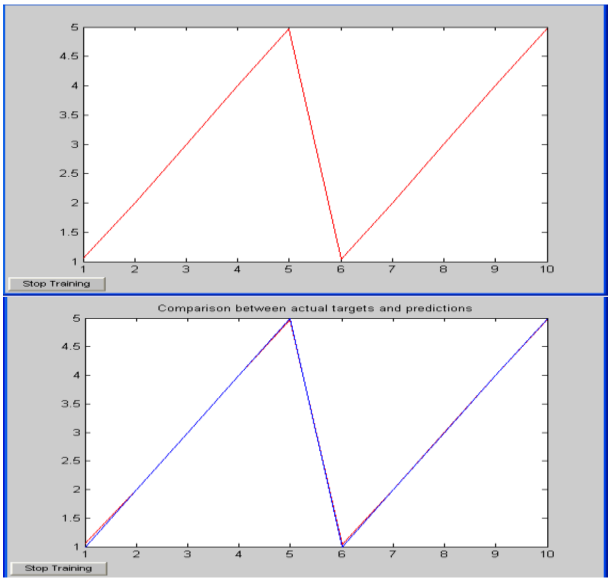

Fig 6. Comparison between target output and predicted output.

Table 7 indicates the FFBPNN classifier accuracy rate for five emotions in vowel domain. The average accuracy rate of the classifier is 86.88%.

Table 7. FFBPNN classifier accuracy rate in %.

Emotions	Normal	Santa	Karuna	Raoudra	Shringar
Normal	84.6	7.2	6	0	2.2
Santa	11	86	2	0	1
Karuna	8	6	82.8	0	3.2
Raoudra	2	0	0	94	4
Shringar	6	2	3	2	87

5. CONCLUSION

In this study, we investigate acoustic properties of speech associated with five different emotions (love (shringar), pathetic (sad) (karuna), wrath (anger) (roudra), quietus (shAnta), normal (neutral)) intentionally expressed in speech by 12 semi-professional actors and 8 semi-professional actresses in Odia vowel domain. Results show speech associated with love (shringar) and sad (karuna) are characterized by longer utterance duration, higher pitch. However we observed that for jitter, anger/love has dominance over others, whereas for shimmer the emotion anger plays a vital role. We also automated the process of identifying the expression of emotions in Odia spoken language by adopting FFBPNN as classifier with jitter and shimmer variants as feature vectors. FFBPNN achieved accuracy rate of 86.88% in vowel domain. In future we are going to incorporate genetic algorithm in training phase of FFBPNN.

6. REFERENCES

[1] Quatieri,T.F., “Discrete-Time Speech Signal Processing Principles and Practice”,Pearson Education,Third Impression 2007

[2] Mohanty S., Bhattacharya S., Bose S., Bose S., Swain S., “A Technique to Prosody analysis in the Synthesis of Natural Speech in the context of Indian Languages” Recent Advances in Computing and Communications, Proceedings of the 12th International Conference on advanced Computing and communications. (2004).

[3] Oh-Work Kwon et. al. “Emotion Recognition by speech signal” EUROSPEECH- GENEVA., 2003.

[4] Mohanty S., Bhattacharya S., Bose S., Bose S., Swain S., “ An Approach to Parametric Base Mood analysis in Oriya Speech Processing” proceedings of frontier of research in Speech and Music (FRSM)’, ITCSRA Kolkatta, India, (2005).

[5] Rabinier L. and Juang B. H. “fundamental of Speech Recognition”, Prentice Hall, (1993).

[6]] Martland, P., whiteside, S. P., Beet, S. W., Baghai-Ravary, L., “Analysis of Ten Vowl sounds across gender and Regional/Cultural Accent”.

[7] Praat software website: http://www.fon.hum.uva.nl/praat.

[8] Mohanty, S, Swain,B.K, “Language Identification using Support Vector Machine” http://desceco.org/O-COCOSDA2010/proceedings/paper_43.pdf.

[9] Bhatti M W, Wang Y & Guan L, A neural network approach for human emotion recognition in speech, 0-7803-8251-X/04/2004 IEEE.

[10] Marques de Sa, J.P, “Pattern Recognition Concepts, Methods and Applications”, Springer 2001.

[11] Beale M H, Hagan M T, Demuth H B, Neural Network Toolbox User’s Guide.

[12] Juang B. H, “Pattern Recognition in Speech and Language Processing”, CRC Press, 2003.