Performance Evaluation of Learning Classifiers for Speech Emotions Corpus using Combinations of Prosodic Features

Syed Abbas Ali
Department of Computer and Information Systems Engineering, NED University of Engineering & Technology, Karachi, Pakistan.

Sitwat Zehra
Karachi Institute of Biotechnology and Genetic Engineering, University of Karachi, Pakistan.

Afsheen Arif
Karachi Institute of Biotechnology and Genetic Engineering, University of Karachi, Pakistan.

ABSTRACT

This paper introduces the speech emotion corpus, a multilingual speech emotion database recorded in the provincial languages of Pakistan: Urdu, Punjabi, Pashto and Sindhi for analyzing the speech emotions present in the recorded speech signals with the four different emotions (Anger, Sadness, Comfort and Happiness). The objective of this paper is to evaluate the performance of the learning classifiers (MLP, Naive Bayes, J48, and SMO) for speech emotion corpus recorded in the provincial languages of Pakistan with different combinations of prosodic features in term of classification accuracy and time taken to build models. The experimental results clearly show that the J48 classifier performs far better than all other classifiers in term of both classification accuracy and model building time. SMO indicates slightly better classification accuracy than Naïve Bayes classifiers whereas; Naïve Bayes exhibit minimum model building time as compared to MLP.

Index Terms

Learning Classifier, Prosodic Features, Speech Emotion Corpus, Emotions

1. INTRODUCTION

Emotion is a physiological and physical process initiated by conscious and/or unconscious perception of state and is associated with mood, temperament, personality and motivation of the person. Emotions are possibly the most fascinating of all mental processes and a composite state of feeling that results in physical and psychological variations that impact thinking and behavior of the persons. Marvin Minsky [1] boldly stated that "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions". Emotions carried in speech express the physiological and psychological states of the speaker's mind. In daily life conversation, people may often understand the word spoken by speaker but misunderstand the emotion in his/her conversation and vice versa. Sixty four speech resources reviewed in [2] for a multilingual speech emotion corpus and it is interesting to find out that such resources are very infrequent, particularly when dealing with European languages. There are very limited resources available in two languages (German and English in [3,4], Slovenian and English [5], Spanish and English [6]), and no speech emotion corpus available in the regional languages of Pakistan. Multilingual speech emotion corpus is realized due to lack of consideration. The aim of Speech emotion recognition (SER) technology is to improve the quality of human-computer interaction. The main focus of research in the field of SER is to resolve the technical difficulties of recognition system by developing the recognition algorithms, but some fundamental questions related to use of such SER technology remain neglected: 1) Can it certainly improve human-computer interaction? 2) For which kinds of application is it suitable? 3) How best implementation can be provided? These fundamental questions are not being addressed because of the current state of the technology and the speech emotion recognition systems are not being capable enough to implement it in real applications. This paper introduces a multilingual speech emotion corpus recorded in the provincial languages of Pakistan and attempt to evaluate the performance of machine learning classifiers for speech emotion corpus recorded in the provincial languages of Pakistan with different combinations of prosodic features in term of classification accuracy and model building time. Rest of the paper is organized as follow. The consequent section discusses the proposed multilingual speech emotion corpus. Three prosodic elements of speech are reviewed in section 3. In section 4, four learning classifiers were defined to evaluate the performance of classifiers for speech emotion corpus. The experimental results and discussions are presented in section 5. Finally, conclusions are drawn in section 6.

2. A MULTILIGUAL SPEECH EMOTION CORPUS

Multilingual speech emotion corpus is an effort to experimentally collect the speech emotion in provincial languages of Pakistan (Sindhi, Urdu, Punjabi and Pashto) with the four different emotions (Anger, Sadness, Comfort and Happiness) defined in [9] using 10 (5 female and 5 male) native speakers from four different regions of Pakistan. The male speakers are selected from the age group of above and below 30 years; three of them were commerce students while rests were professionals. Similarly female speakers are selected from the age group of above and below 30 years; three were literature students while rests were professionals. The recording specification of the proposed speech emotion corpus development based on the ITU recommendations. The recording has been performed in standard recording environment having SNR≥45dB.Built-in sound recorder of

Microsoft Windows 7has been used to record the entire speech emotion of native speakers. The recording format is 16 bit, Mono, PCM and sampling rate of 48 KHz with microphone impendence and sensitivity of 2.2W and 54dB±2dB respectively, pulp stereo type of 3.5mm and length of cable is 1.8m. The selection of a carrier sentence was highly exclusive, subsequent some a-priori well-known desiderata. The carefully chosen sentence should have following characteristics:

- Selected sentence for a specific context should be semantically neutral and do not have any emotional value and interpretation in the sentence, which is in the mind of listener.
- Selected sentence should be consistent with any situation present in the speech emotion.
- It should be correct and follow the general guidelines of each language to avoid distraction or confusion in the encoders, and subsequently in the listeners.
- Selected sentence should be easily identified for speech emotion analysis
- Based on the previous studies reported in [7, 8], selected sentence for our speech emotion corpus was:

"Let's go home"

This selected sentence spoken in four different regional languages of Pakistan (Urdu, Sindhi, Punjabi and Pashto) with their native speakers, for example in Urdu language:

"Chalo Ghar Chalo" "چلو گھر چلو" (Angry)

"Chalo Ghar Chalain" "چلو گھر چلیں" (Sad, Happy, Comfort)

3. PROSODIC FEATURES

Prosody is defined as rhythm, stress and inflection of speech in linguistic. Prosody features of the speaker or utterance are as follow: speaker emotion state; the kind of the expression (question, command and statement); the presence of irony or pessimism; variation, anxiety and enthusiasm; or other features of language which cannot be defined by choice of grammar or vocabulary. Prosody has been studied as vital source of knowledge in speech emotion research community for examining and understanding the emotional expressions present in acoustic signal [10]. Prosodic features are physically apprehended in the speech as a set of acoustic parameters variation and provide the combination of energy variation and period of speech segment during pitch of speech and speech. Prosody features are used to deliver added sense to the spoken word in natural speech to provide emotion of speaker such as; happiness, comfort, sadness and anger. Different combinations of three prosodic features are used in this experiment for statistical analysis of emotions in speech signal.

- **Intensity:** Intensity is defined as energy of the speech signal which is used to encode prosodic information. Physiologically, variations of intensity partly correlate with fundamental frequency variations in case of domestic reputations [11]. Recent research studies identified the role of intensity variations into information focus [13] and prosodic group [12] independent of fundamental frequency variations.
- **Pitch:** One of the important prosody features that have perceptual property which is used to ordering the sound by using frequency scale. Pitch can be measured as a frequency and it is not a purely objective physical property but it is a subjective psycho acoustical attribute of sound [14].
- **Formants:** Formant is defined as human vocal tract acoustic resonance and significant or unique frequency components of human speech. Formants are used to differentiate between vowels by quantifying the frequency components of vowels sounds. Few whistle tones of formants are come from periodic collapse of Venturi effect low pressure zones but chamber resonance produced the values of formants [15].

4. LEARNING CLASSIFIERS

Classification is a classic data mining techniques based on machine learning and it is used to classify each item in a data set into one of the predefined set of groups or classes. The most commonly used learning classifiers for speech emotion recognition are K-nearest-neighbor methods (k-NN), C.45 decision tree, Support vector machine (SVM), Artificial neural network (ANN) and Naïve Bayes (NB). These learning classifiers have been compared on speech emotion assets in [16, 17, 18, 19, 20, 21]. Experimental framework made use of J48, Naïve Bayes, Multilayer Perceptron, and Sequential Minimal Optimization (SMO) classifiers to evaluate the classification accuracy using WEKA Data Mining software.

- **Naïve Bayes:** The NB classifier predicts class membership probabilities using Bayes theorem. It employs all variable that are included in the data samples, and inspect all of them individually and give equal significance and independence to each other. This implementation is called class conditional independence [22] .The Naive Bayes classifier has higher error rate as compared to Neural Network classifier.
- **J48:** The J48 classifier is the Weka's implementation of the C4.5 decision tree. C4.5 implements a greedy approach in which decision tree is built in a top-down recursive divide and conquer manner. Top down approach means that an algorithm starts with a set of labeled training data and their associated class labels whereas, training set is recursively divided into smaller subsets as the tree is being built [22].
- **Multilayer Perceptron:** A multilayer perceptron (MLP) is a multi-layer, feed forward network that utilizes a supervised learning technique called error back propagation as the learning method [24]. Supervised learning technique consists of two steps for training MLP network: a forward step and a backward step. In forward training step, the computational units are responsible to propagate the signal until it reaches to the output layer; whereas, in backward training all the synaptic weights are adjusted with respect to an error correction rule[23].
- **Sequential Minimal Optimization (SMO):** Support vector machine (SVM) training makes use of SMO algorithm for solving optimization problem. SMO is an improved training algorithm for SVMs developed by J.C.Platt in 1998[25]. The working principle of SMO algorithm based on QP problems, the large QP problems were divided into sequence of small QP problems and contrasting with other algorithms, SMO uses the least possible QP problems to perform quick analysis and improve scaling and computation time.

5. EXPERIMENTAL RESULTS AND DISCUSSION

Experimental frame work is divided into two phases to evaluate the performance of learning classifiers for speech emotion with different combinations of prosodic features (Pitch-Intensity, Pitch-Formant, Intensity-Formant and Pitch-Intensity-Formant). In the first phase, experiments were done using PRAAT (statistical analysis) software to analyze the speech emotion present in the spoken utterances with the four different emotions (Anger, Sadness, Comfort and Happiness). The proposed speech emotion corpus used in the experiments here consists of 404 speech samples taken from the recording of female and male speakers in provincial languages of Pakistan (Urdu, Sindhi, Punjabi, Pashto) and each person spoke the sentence in four different emotions to observe the dependency of emotions on prosodic features.

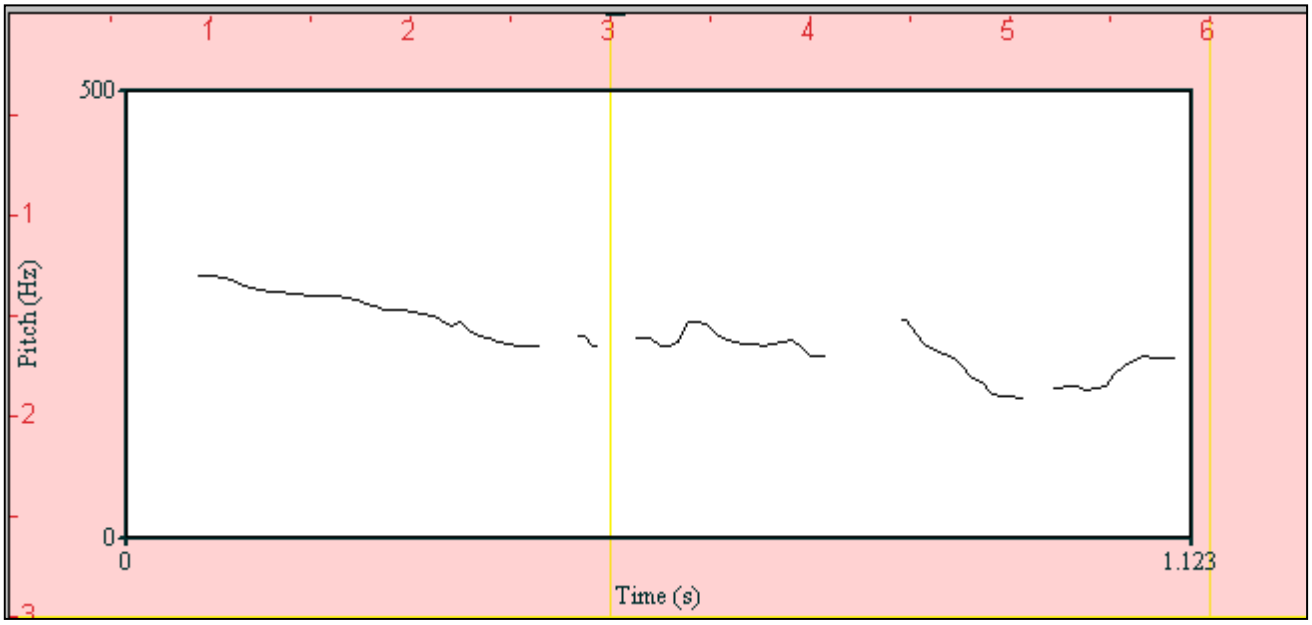

Fig 1: Comfort

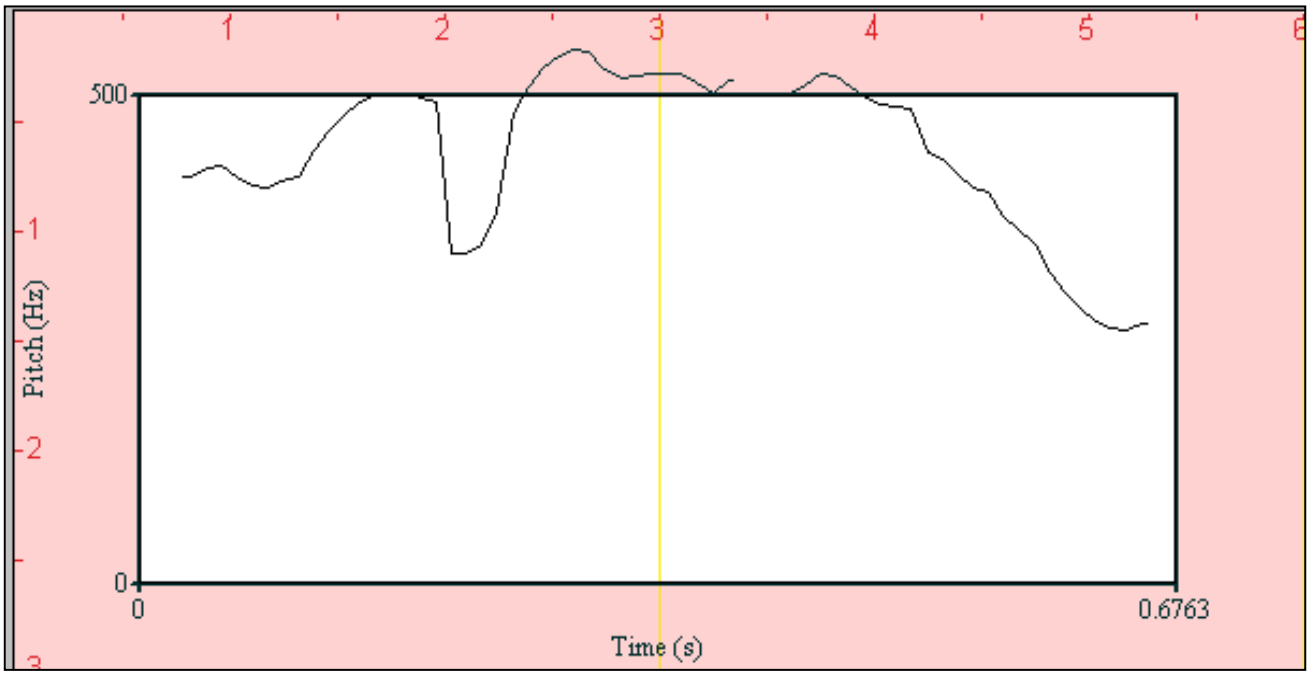

Fig 2: Anger

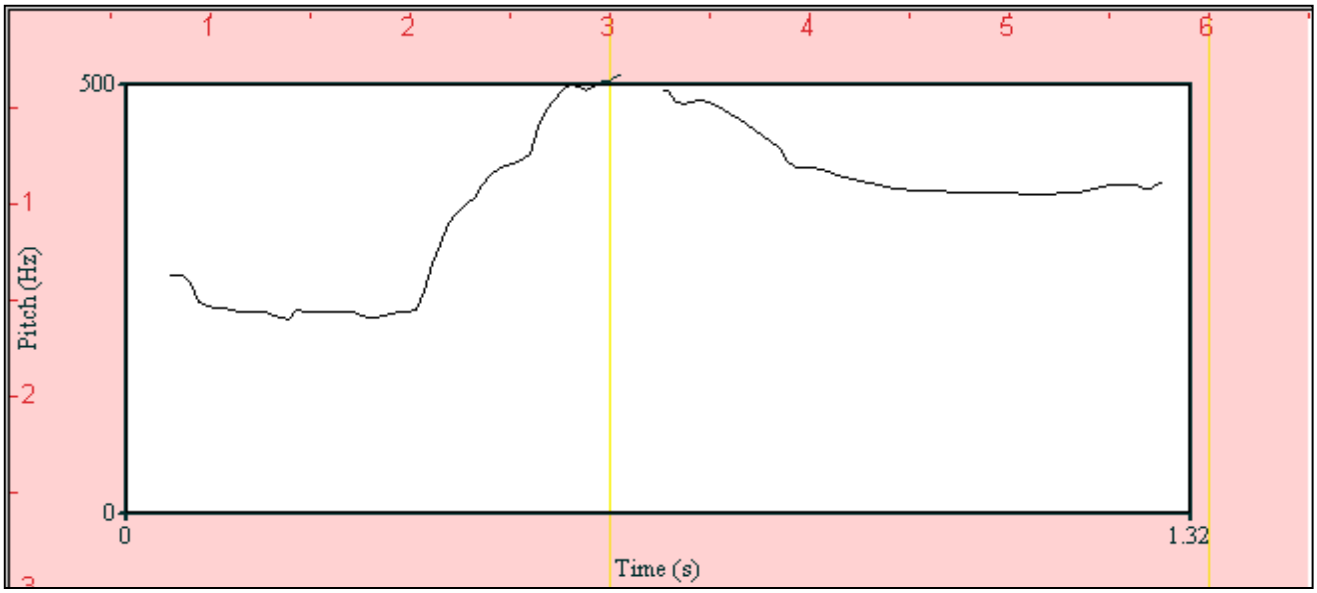

Fig 3: Happy

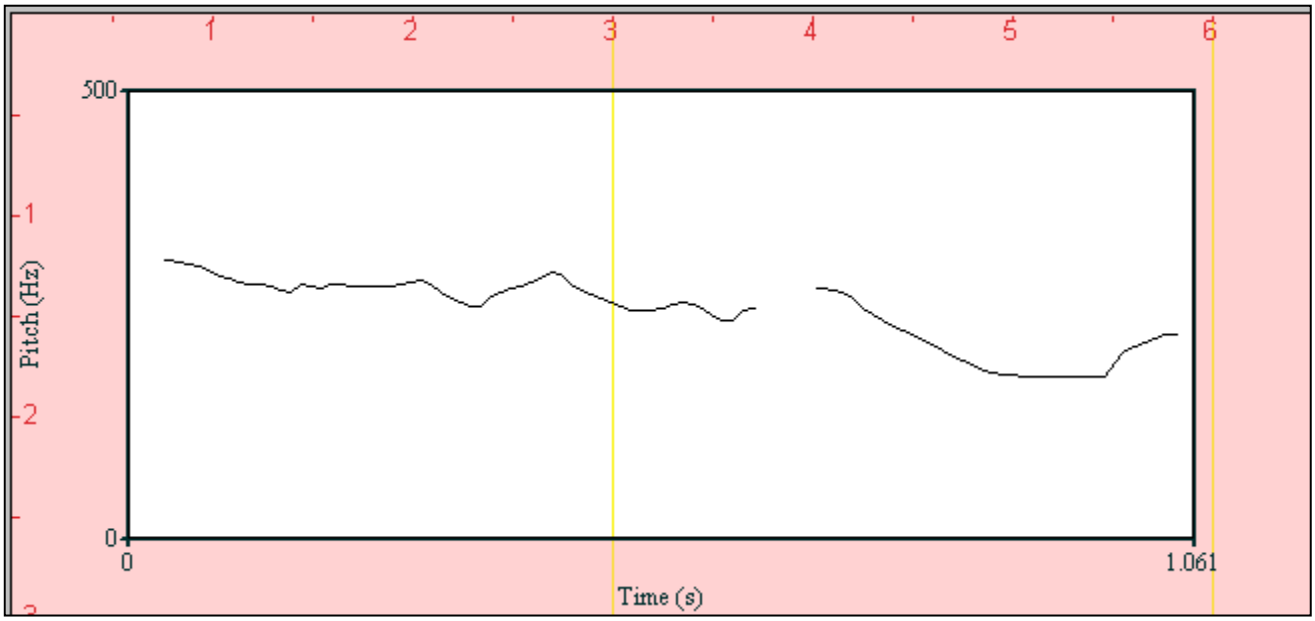

Fig 4: Sadness

Fig 1 to Fig 4 provides the pictorial description of the pitch of the sentence "Let's go home" in urdu language with four different emotions. The PRAAT software was used to observe the mean and standard deviation values of all three prosodic features in four regional languages of Pakistan with four different emotions to perform the statistical analysis of four different speech emotions taken from proposed speech emotion corpus on three prosodic features.

Table 1: Comparative Analysis of Prosodic Features with Sadness

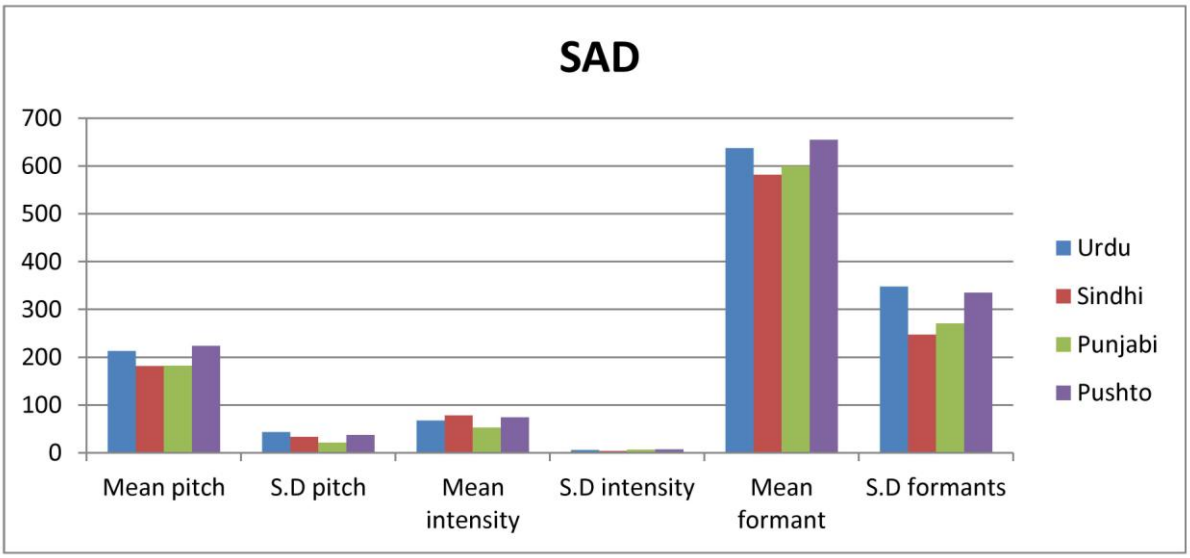

Table 2: Comparative Analysis of Prosodic Features with Happiness

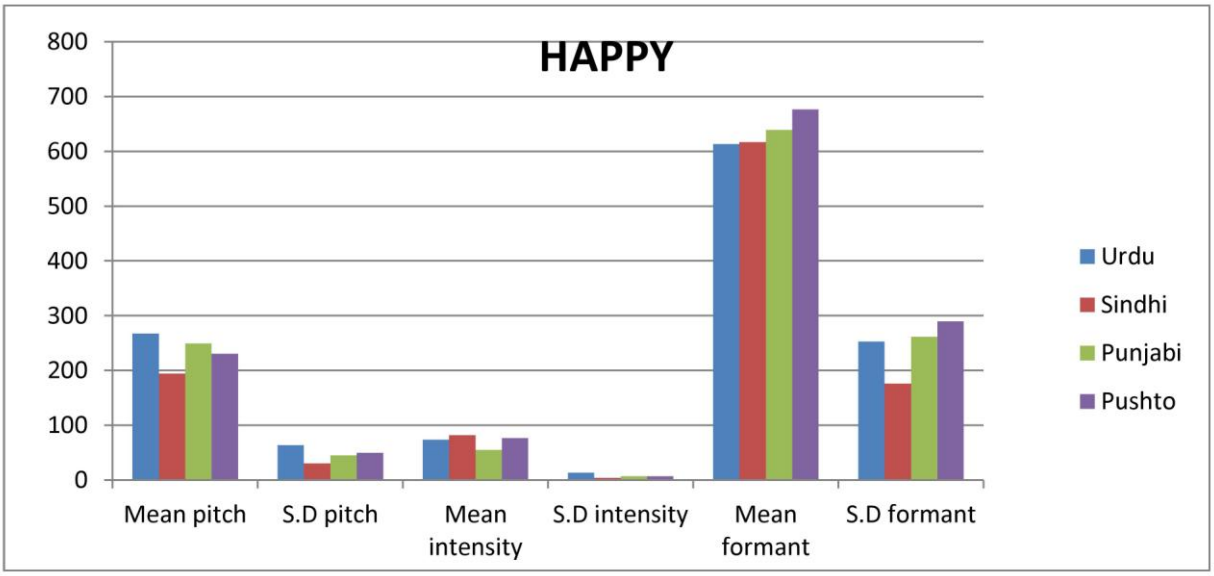

Table 3: Comparative Analysis of Prosodic Features with Comfort

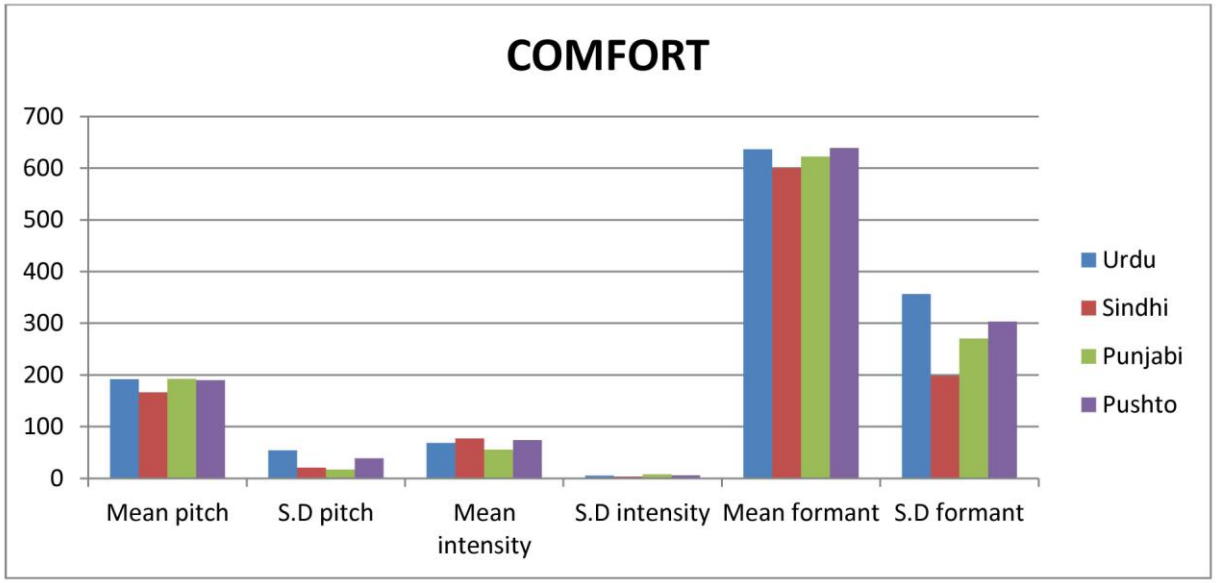

Table 4: Comparative Analysis of Prosodic Features with Anger

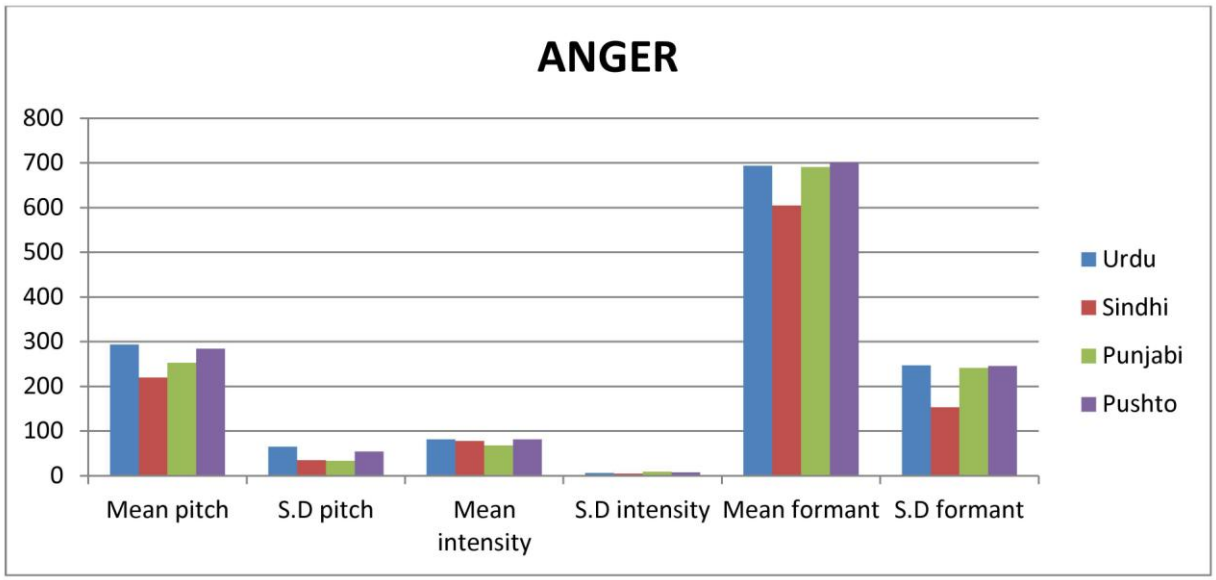

Table 1 to Table 4 provide the comparative analysis of three prosodic features (Pitch, Intensity, Formant) in provincial languages of Pakistan with four different emotions: Anger, Happiness, Sadness and Comfort. Proposed speech emotion corpus was used to develop this comparison among four different speech emotions consist of male and female speakers in provincial languages of Pakistan. The tables describe the mean values of pitch, intensity, formant frequency and the variation among them in term of speech emotion samples. The comparison has been made on the basis of the mean values of speech emotion samples. Experimental results demonstrate the following observations based on three prosodic features with four speech emotions in term of mean value and standard deviation (S.D) that for all of the four emotions, it is observed that each emotion has high pitch, lower intensity and higher formant. The graphical analysis was performed to detect the most suitable prosodic feature in order to determine all the four emotions through this analysis. Graphical analysis shows that Intensity appears to be the best feature for emotion detection as it has the minimum variations in all the four emotions (happy, sad, anger, comfort) whereas the formants and pitch has more variations in comparison with intensity.

From the above comparative analysis of three prosodic features with four speech emotions it can be conclude that for happy, sad, anger and comfort, "Intensity" seems to be the most suitable prosodic feature in order to determine these speech emotion as its mean values (for four regional languages) appear to be closer unlike the other two features and deviation is small.

```
@relation emotiondata

@attribute name {Azhar, Sofia, Nusrat, Nuzat, Akram, Qamar, Annie, Halima,
Hamza, Umer, female1, female2, female4, male1, male2, male6, male3, male4,
male5, female3, lubna, sadia, Akram2, sohail, rida, maham, zahra, rizwan,
adeel, talib, khaliq, afzalkhan, nazia, kk, adeel2, bakar, mubarakali,
shugufta, rida2, maham2}
@attribute age_30 {above,below}
@attribute gender {male,female}
@attribute language {punjabi,sindhi,urdu,pushto}
@attribute meanIntensity  numeric
@attribute emotion {happy,sad,anger,comfort}

@data
Azhar,above,male,punjabi,64.08650396,anger
Sofia,above,female,punjabi,70.52540293,anger
Nusrat,above,female,punjabi,64.9607165,anger
Nuzat,above,female,punjabi,70.12636972,anger
Akram,above,male,punjabi,72.57133472,anger
Qamar,above,male,punjabi,73.25790954,anger
Azhar,above,male,punjabi,51.18455618,comfort
Sofia,above,female,punjabi,52.26995343,comfort
Nusrat,above,female,punjabi,54.21946173,comfort
Nuzat,above,female,punjabi,59.65218688,comfort
Akram,above,male,punjabi,64.56387544,comfort
Qamar,above,male,punjabi,50.68916911,comfort
Azhar,above,male,punjabi,48.20366413,happy
Sofia,above,female,punjabi,67.94989882,happy
Nusrat,above,female,punjabi,57.74034635,happy
Nuzat,above,female,punjabi,50.06525591,happy
Akram,above,male,punjabi,55.93836538,happy
Qamar,above,male,punjabi,48.24582294,happy
Azhar,above,male,punjabi,39.97449943,sad
Sofia,above,female,punjabi,52.81314503,sad
Nusrat,above,female,punjabi,52.95780859,sad
Nuzat,above,female,punjabi,55.2935267,sad
Akram,above,male,punjabi,59.90636216,sad
```

Fig 5: ARFF data file format for Speech Emotion Dataset

In the second phase of the experiments, WEKA Data Mining tool [26] were used to evaluate the performance of learning classifiers with different combinations of prosodic features. Different experiments have been performed on the proposed speech emotion corpus with four classification algorithms: Naïve Bayes, J48, MLP and SMO using different combinations of features: Pitch-Intensity, Pitch-Formant, Intensity-Formant and Pitch-Intensity-Formant for each speech emotion using 10-fold cross validation to prevent over fitting [27]. In WEKA Data Mining tool, data samples should be formatted to the ARFF format. ARFF file format of proposed speech emotions corpus are shown in Fig 5. The WEKA Data Mining Explorer make use of these file format automatically if it does not recognize a given file as an ARFF file, the Preprocess section has an option for importing data set from a database and filtering algorithm use to preprocess this data set. These filters are not only used to transform data but make it possible to remove attributes and instances with respect to particular conditions [28].

Table 5: Classification Accuracy of Learning Classifiers

Learning classifiers	Combinations of Prosodic features	Total No. of Instances	No. of correct instances	No. of incorrect instances	Time to build models (seconds)	Classification Accuracy
Naïve Bayes	**Pitch +Intensity +Formant**	132	52	80	0.02	**39.3939%**
	Intensity + Formant	132	42	90	0	**31.8182%**
	Pitch + Formant	132	48	84	0	**36.3636%**
	Pitch + Intensity	132	49	83	0	**37.1212%**
J48	**Pitch +Intensity +Formant**	132	99	33	0.06	**75%**
	Intensity + Formant	132	98	34	0.01	**74.2424%**
	Pitch + Formant	132	94	38	0.01	**71.2121%**
	Pitch + Intensity	132	98	34	0.01	**74.2424%**
MLP	**Pitch +Intensity +Formant**	132	83	49	0.91	**62.8788%**
	Intensity + Formant	132	63	69	0.7	**47.7273%**
	Pitch + Formant	132	72	60	0.72	**54.5455%**
	Pitch + Intensity	132	82	50	0.73	**62.1212%**
SMO	**Pitch +Intensity +Formant**	132	66	66	0.36	**50%**
	Intensity + Formant	132	60	72	0.35	**45.4545%**
	Pitch + Formant	132	55	77	0.07	**41.6667%**
	Pitch + Intensity	132	55	77	0.07	41.6667%

Table 5 provide the comprehensive table for performance evaluation of learning classifiers for speech emotion corpus recorded in the provincial languages of Pakistan with different combinations of prosodic features in term of classification accuracy and model building time. Experimental results show that the J48 classifier performs far better than all other classifiers in term of both classification accuracy and model building time. Naïve Bayes classifier shows less classification accuracy as compared to SMO whereas; Naïve Bayes exhibit minimum model building time as compared to MLP.

6. CONCLUSIONS

This paper introduced speech emotion corpus recorded in provincial languages of Pakistan: Urdu, Punjabi, Pashto and Sindhi with four different emotions (Anger, sadness, Happiness and Comfort). In this initial study, the performance of learning classifiers (MLP, Nave Bayes, J48 and SMO) is evaluated with different combinations of Prosodic features (Pitch-Intensity, Pitch-Formant, Intensity-Formant and Pitch-Intensity-Formant) to classify the speech emotions and identify the strength of learning classifiers in term of classification accuracy and model building time. Experiments have been performed using WEKA Data Mining tool to evaluate the performance of learning classifiers. Experimental results evident that the classification accuracy and model building time of J48 classifier is better than all other classifiers. Naïve Bayes exhibit minimum model building time as compared to MLP whereas; Naïve Bayes classifier shows less classification accuracy as compared to SMO. In future research work, authors are focusing on enhancing the developed speech emotion corpus, considering other prosodic features (shimmers, tonal and non-tonal etc.) to investigate the performance of learning classifiers and investigating the performance of learning classifiers for speech emotions of abnormal and mentally retarded peoples.

7. REFERENCES

[1] M.L.Minsky. The society of mind. New York, N.Y.: Simon and Schuster, 1986.

[2] D. Ververidis and C. Kotropulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication Elsevier, Vol. 48, pp. 1162-1181, 2006.

[3] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russel and M. Wong, "'you stupid tin box' - children interacting with the AIBO robot: a cross linguistic emotional speech corpus," in Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC '04), pp. 171-174, 2004.

[4] K.R. Scherer, D. Grandjean, L.T. Johnstone, G. Klasmeyer, " Acoustic correlates of task load and stress," in Proceedings of the International Conference on Spoken Language Processing (ICSLP '02), pp.2017-2020, 2002.

[5] D.C. Ambrus, "Collecting and recording of an emotional speech database," Technical Report, Faculty of Electrical Engineering, Institute of Electronics, University of Maribor., 2000.

[6] G.M. Gonzalez, "Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in Chicanos/Latinos," Technical Report TR-0039, University of Michigan., 1999.

[7] H.G. Wallbott and K.R. Scherer, "Cues and channels in emotion recognition," Journal of personality and social psychology, Vol. 51, pp. 690-699, 1986.

[8] L. Anolli, L. Wang, F. Mantovani and A. De Toni, "The Voice of Emotion in Chinese and Italian Young Adults," Journal of Cross-Cultural Psychology, Vol. 39, pp. 565-598, 2008.

[9] P. Ekman, "An argument for basic emotions, "Cognition and Emotion, Vol. 6, pp. 169-200, 1992.

[10] M. Kuremastsu et al, "An extraction of emotion in human speech using speech synthesize and classifiers for each emotion," WSEAS Transaction on Information Science and Applications, Vol .5(3), pp.246-251, 2008.

[11] J. Atkinson, "Correlation analysis of the physiological factors controlling fundamental voice frequency," Journal of the Acoustic Society of America, Vol. 63(1), pp.211-222, 1978.

[12] C. Tseng, and Y. Lee, "Intensity in relation to prosody organization," In International Symposium on Chinese Spoken Language Processing, pp.217-220, Hong-Kong, China.2004

[13] D. Beaver, B. Zack Clarck, E. Flemming, T.F. Jaeger and M. Wolters, "When semantics meets phonetics: Acoustical studies of second-occurrence focus," Journal of the Linguistic Society of America, Vol. 83(2), pp. 245-276, 2008.

[14] Formant-Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Formant.

[15] Pitch-Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Pitch_(music).

[16] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining Efforts for Improving Automatic Classification of Emotional User States," In Proc. of IS-LTC, pages 240—245, 2006.

[17] I. Iriondo, S. Planet, J.C. Socoro, and F. Alias, "Objective and Subjective Evaluation of an Expressive Speech Corpus," Advances in Nonlinear Speech Processing, LNCS, 4885:86, 2007.

[18] D. Morrison and L.C. De Silva, "Voting Ensembles for Spoken Affect Classification," Journal of Network and Computer Applications, 30(4):1356—1365, 2007.

[19] M. Shami and W. Verhelst, "Automatic Classification of Expressiveness in Speech: A Multi-Corpus Study," Speaker Classification II, LNCS, 4441:43—56, 2007.

[20] L. Vidrascu and L. Devillers, "Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center," In Proc. of 2005 IEEE International Conference on Multimedia and Expo, pages 944—947, 2005.

[21] S. Yilmazyildiz, W. Mattheyses, Y. Patsi, and W. Verhelst, "Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication," PCM 2006, LNCS, 4261:1—8, 2006.

[22] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Elsevier, 2nd edition, 2006.

[23] S. Haykin, R. Neurais – Principios e Pratica, Bookman, 2 ed., Porto Alegre, 2001.

[24] C.M Bishop, Neural Networks for Pattern Recognition, Oxford, New York, 1995.

[25] J.C.Platt, "Sequential Minimal Optimization: A fast algorithm for training Support Vector Machine," Technical Report MSR-TR-98-14, Microsoft Research, 1998.

[26] R. R. Bouckaert, E. Frank, M. H. R. Kirkby, P. Reutemann, S. D. Scuse, WEKA Manual for Version 3-7-5, October 28, 2011.

[27] I. H.Witten and E.Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2005.

[28] R. Dimov, WEKA: Practical Machine Learning Tools and Techniques in Java, 2006/07.