Analysis of Children's Prosodic Features Using Emotion Based Utterances in Urdu Language

Sallar Khan
Department of Computer Science
Sir Syed University of Engineering and Technology
Karachi, Pakistan
Sallarkhan_92@yahoo.com

Syed Abbas Ali
Department of Computer Information System, NED University of Engineering and Technology
Karachi, Pakistan
Saaj.research@yahoo.com

Jawaria Sallar
Department of Computer Science
Sir Syed University of Engineering and Technology
Karachi, Pakistan
H.jawaria@yahoo.com

***Abstract*—Emotion plays a significant role in identifying the states of a speaker using spoken utterances. Prosodic features add sense in spoken utterances providing speaker emotions. The objective of this research is to analyze the behavior of prosodic features (individual and in combination with others' prosodic features) with different learning classifiers on emotion based utterances of children in the Urdu language. In this paper, three different prosodic features (intensity, pitch, formant and their combinations) with five different learning classifiers(ANN, J-48, K-star, Naïve Bayes, decision stump) and four basic emotions (happy, sad, angry, and neutral) were used to develop the experimental framework. Demonstrative experiments expressed that, in terms of classification accuracy, artificial neural networks show significant results with both individual and combination of prosodic features in comparison with other learning classifiers.**

Keywords-speech; emotion; recognition; learning; classifiers; prosodic; features; language; Urdu; Pakistan

I. INTRODUCTION AND RELATED WORK

Speech is commonly known as an effective way of communication between human beings [1]. Emotion is an application which is very much concerned with the speech in which we can perceive the speaker's mental state through his/her spoken utterances and this term is known as speech emotion recognition (SER). Using machine learning systems with partial computational resources we can identify speech emotion by the usage of a few emotions (happy, sad, angry, and neutral) [2]. The efficiency of human machine interface in the field of human machine interaction can significantly be improved by the help of automatic SER with the assistance of learning machines [3]. Speech signal and its acoustic features such as timing, intensity, voice quality, pitch and articulation are highly associated with fundamental emotion [5]. Speech emotion recognition systems have numerous applications which include: 1) telephonic conversations and their emotion analysis 2) psychiatric patient's medical diagnosis 3) student emotional condition e-learning system 4) analysis of mental stress level during an exchange of conversations. There are 3 core phases leading SER as a pattern recognition statistical problem: 1) feature extraction from speech, 2) feature selection, 3) pattern classification [5]. For the evaluation and impact of spectral and prosodic features of emotional speech on classification, Gaussian mixture models (GMM) were used in [6]. Three fundamental parts which could enhance the design of a speech emotion recognition system are discussed in [7]: 1) Selection of suitable features for speech, 2) appropriate classification scheme's design and 3) system performance has been evaluated through the presentation of a database designed for emotional speech. Speech emotion recognition review and analysis are presented in [8] by the use of different learning classifiers. Extraction of both local and global prosodic features from sentences are addressed in [9], furthermore, different words and syllables are suggested for analyzing the speech recognition or affect recognition. The technology of emotion as a crucial component of artificial intelligence is advised in [10]. Furthermore, distinct context must be considered by artificial intelligence for emotion recognition. Five different emotions have been investigated in [11] which are associated with acoustic properties of the prosody of speech. This investigation comes to a result that those speeches which are associated with emotion "love" and "sad" are identified by higher pitch and utterances with lengthier duration. Similar to [11], prosody is recognized as the most fundamental feature of emotional expression in any specific speech in [12].

II. RESEARCH METHODOLOGY

A. Corpus Collection and Specification of Recording

The structure for this research is to conduct interviews from primary level school going children, while the medium chosen will be the regional language of Pakistan Urdu. The same sentence will be asked from all of the participated children. ITU recommendations which were chosen for the recording of the corpus with following specifications are a) 24120 bps and b) SNR>=45 dB [13]. For the experiment analysis, a noise free room will be utilized for taking recording samples by using a microphone, and the entire speech emotion utterances are recorded in the recording format of 2.4 Ohms; sensitivity and 48 kHhz; sampling rate. With four different emotions (happy, sad, angry, and neutral) the Urdu language sentence which is used for experiment is: "I want to play" ("میں کھیلنا چاہتا ہوں")

B. Population

Children were chosen randomly from a primary level school with the age group of the population of this research being 5-10 years.

C. Sampling Frame

Sufficient ratio of sampling frame is chosen which can produce the desired result. It is selected from past studies from the realm of SER. For analysis purpose in this research, the data will consist of approximately 10 participants with 40 speech samples (Consisting all four emotions).

D. Procedure

1) Feature Extraction

The extraction process of prosodic features (pitch, intensity and formant) is done through testing of these samples on PRAAT, while only mean values of prosodic features were extracted in the experimental section.

2) Learning Classification

We have used 5 different machine learning classifiers for classification purpose which are Artificial Neural Network (ANN), Naïve Bayes (NB), Decision Stump (DS), J-48, and K-star.

3) Pilot Study of the Proposed Research

For getting familiarized with the software/tools which we have utilized for this study, we have collected the speech samples. The .wav file format was used for the recorded samples, while PRAAT will generate the desired result accordingly. Whereas, WEKA classifier will process these values to generate the classification result.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Extracting Prosodic Features

To analyze the overall impact of prosodic features (intensity, pitch, and formant) on learning classifiers the experimental phase of this research is divided into two portions. In the first portion, the four emotions (happy, sad, angry, and neutral) which are present in the utterances of the speaker have been observed by using the PRAAT software [14]. Speech emotion corpuses which are used in the experimental study consist of initially 40 samples appropriated from a children recording in the age group of 5-10 years in regional language Urdu with four different emotions. To evaluate the dependency of emotions on each prosodic feature (pitch, intensity, and formant) is the primary motivation through these experiments. The extraction process of prosodic features from spoken speech emotion utterances in regional langue Urdu are shown in Figures 1 to 3. These observations are demonstrating the behavior of all three prosodic features which are extracted by using PRAAT software [1]. Not all the total 40 samples are being demonstrated in this section but just a portion to show an aroma that how these features were extracted with the use of PRAAT.

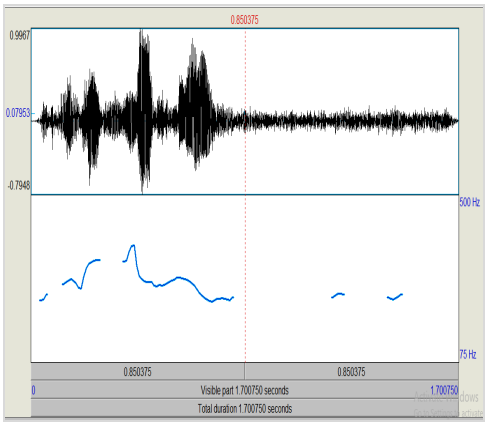

Fig. 1. Extracting prosodic feature (pitch) using PRAAT.

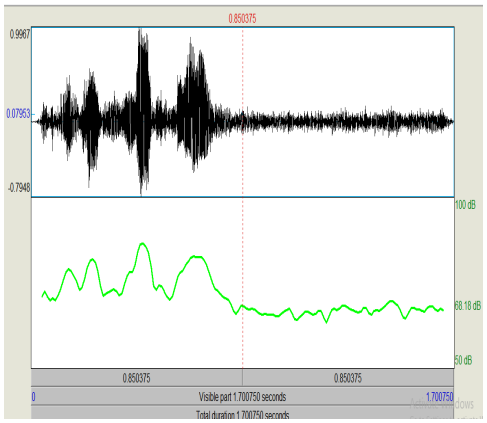

Fig. 2. Extracting prosodic feature (intensity) using PRAAT.

B. Classification and precision accuracy

The second portion is further divided into two parts, in first part we will discuss the overall calculation of classification accuracy of each learning classifier (ANN, NB, DS, J-48, and K-star) against each prosodic feature individually as shown in Table I and in combination as shown in Table II. In the second part, we will discuss the precision accuracy against each emotion (happy, sad, angry, and neutral) which are produced by different learning classifiers.

1) Overall classification accuracy

The overall result is satisfactory in terms of classification accuracy. It has been observed (Figure 5) that learning classifiers performed well when we jointly analyzed prosodic features and can classify correctly up to the accuracy of 45% meanwhile separately they produce accuracy of 40% which is shown in Figure 4. Furthermore in terms of learning classifiers, J-48 gave best classification accuracy of 35% for the prosodic feature pitch, while on the other hand ANN and NB both held their classification accuracy higher up to 40% for the prosodic feature intensity and for the third prosodic feature, formant, ANN and DS both have produced classification accuracy around 33%. During the process of analyzing prosodic features jointly, it has been recognized that the combination of intensity

and formant produced the finest classification accuracy of 45% with ANN, while combinations of pitch and formant against J-48, as well as pitch, intensity, and formant with ANN came second, both with accuracy around 43%. Lastly the combination of pitch and intensity produced classification accuracy around 38% with ANN classifier which comes third.

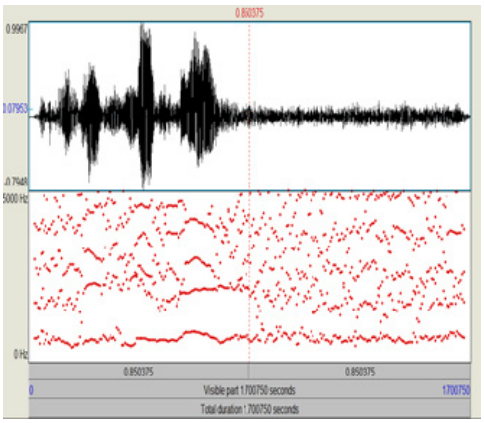

Fig. 3. Extacting prosodic feature (formant) using PRAAT.

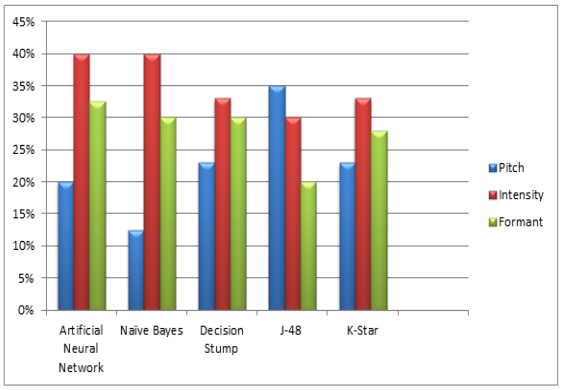

Fig. 4. Classification accuracy of emotions on pitch, intensity and formant.

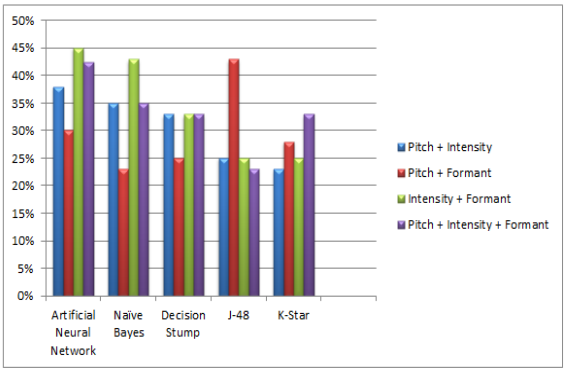

Fig. 5. Classification accuracy of emotions on combination of pitch, intensity and formant.

TABLE I. COMPARATIVE ANALYSIS OF INDIVIDUAL PROSODIC FEATURES USING MACHINE LEARNING CLASSIFIERS

	Classifier	Precision				Accuracy
		Happy	*Sad*	*Angry*	*Neutral*	
Pitch	ANN	0.333	0	0	0.316	20%
	NB	0.125	0	0.222	0.417	12.50%
	DS	0.217	0	0.25	0.273	22.50%
	J-48	0.231	0	0.455	0.429	35%
	K-Star	0.2	0	0.222	0.417	22.50%
Intensity	ANN	0.286	0	0.545	0.4	40%
	NB	0.25	0	0.5	0.375	40%
	DS	0.143	0	0.545	0.273	32.50%
	J-48	0.125	0	0.462	0.263	30%
	K-Star	0.111	0.2	0.545	0.333	32.50%
Formant	ANN	0.4	0.75	0.182	0.2	32.50%
	NB	0.385	0.385	0.267	0	30%
	DS	0	1	0.27	0	32.50%
	J-48	0	0	0.242	0	20%
	K-Star	0.231	0.5	0.308	0.125	27.50%

TABLE II. COMPARATIVE ANALYSIS OF PROSODIC FEATURE COMBINATIONS USING MACHINE LEARNING CLASSIFIERS

	Classifier	Precision				Accuracy
		Happy	*Sad*	*Angry*	*Neutral*	
P+I	ANN	0.385	0	0.455	0.333	37.50%
	NB	0.25	0	0.467	0.333	35%
	DS	0.143	0	0.545	0.273	32.50%
	J-48	0.133	0	0.556	0.188	25%
	K-Star	0.083	0.2	0.5	0.2	22.50%
P+F	ANN	0.25	0.429	0	0.412	30%
	NB	0.182	0.429	0.154	0.222	22.50%
	DS	0.25	0	0.25	0.25	25%
	J-48	0.417	0.333	0.444	0.462	42.50%
	K-Star	0.375	0.429	0.143	0.273	27.50%
I+F	ANN	0.375	0.667	0.545	0.417	45%
	NB	0.375	0.333	0.5	0.429	42.50%
	DS	0.143	0	0.545	0.273	32.50%
	J-48	0.083	0	0.462	0.2	25%
	K-Star	0.154	0.2	0.5	0.167	25%
P+I+F	ANN	0.385	0.5	0.444	0.429	42.50%
	NB	0.25	0.286	0.455	0.357	35%
	DS	0.143	0	0.545	0.273	32.50%
	J-48	0.182	0	0.5	0.125	22.50%
	K-Star	0.222	0.222	0.5	0.333	32.50%

P: Pitch, I: Intensity, F: Formant

2) Precision Accuracy

The plot of precision against each emotion (happy, sad, angry, and neutral) is shown in Figures 6 to 9. Analyzing process shows that J-48 kept the highest precision rate of around 0.45 regarding the happy emotion with the combination of pitch and formant features, while classifiers didn't produce well enough results for the emotion sad except ANN which produced precision rate of around 0.8. For the angry emotion, every classifier produced their best possible results against almost every feature (individually or jointly) but J-48 produced a slightly better precision rate of 0.6 for the combination of pitch and intensity. In the end, for neutral emotion, again J-48 achieved the highest precision rate of around 0.5 with the combination of pitch and formant.

IV. CONCLUSION

Speech emotion corpuses were recorded in the Urdu language with separate regard of four basic emotions. The main motivation behind this research is to analyze the impact of prosodic features (pitch, intensity and formant) on five different learning classifiers. The PRAAT software WEKA tool was used in the experimental framework for emotion

observation in spoken utterances and as learning classifier for the classification purpose. Experimental results made evident that by combining prosodic features (intensity and formant) we can achieve a classification rate up to 45% while 40% classification accuracy can be achieved from an individual prosodic feature (intensity). In terms of classification accuracy, the ANN has been proved overall to perform better than others with a classification accuracy of 45%.

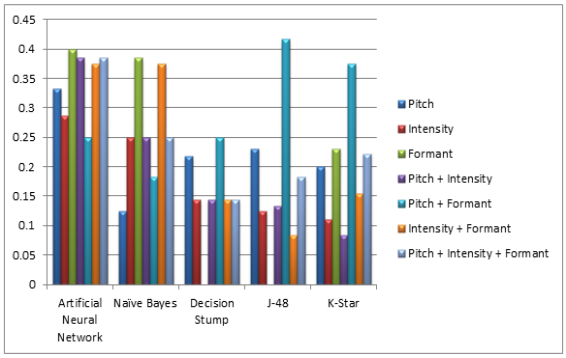

Fig. 6. Precision of emotion 'happy' on all three prosodic features.

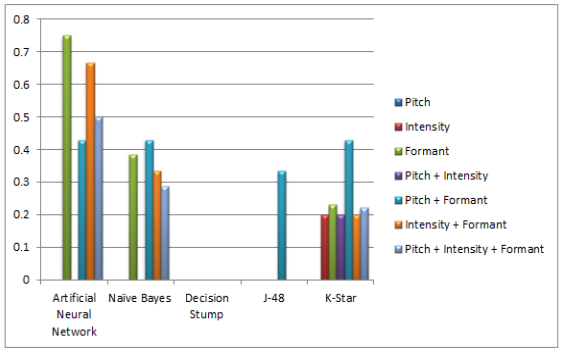

Fig. 7. Precision of emotion 'sad' on all three prosodic features

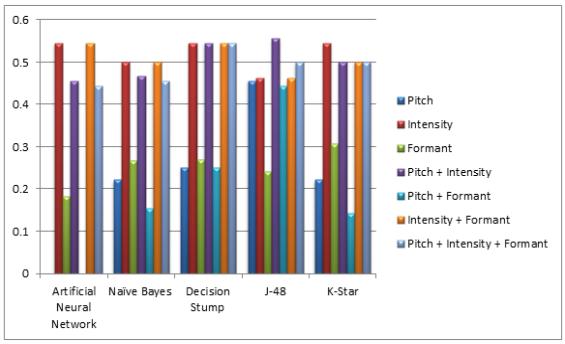

Fig. 8. Precision of emotion 'angry' on all three prosodic features.

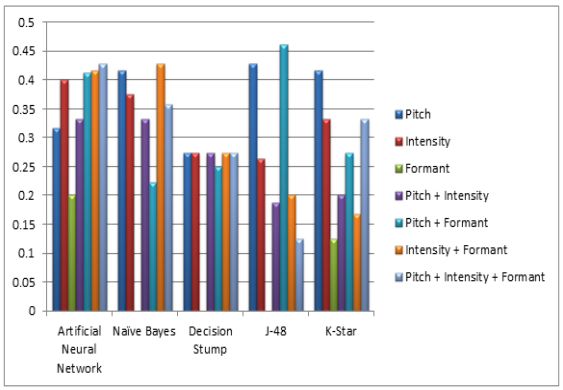

Fig. 9. Precision of emotion 'neutral' on all three prosodic features.

REFERENCES

[1] S. A. Ali, A. Khan, N. Bashir, "Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus", International Journal of Information Technology and Computer Science, Vol. 2, pp. 54-59, 2015

[2] P. Ekman, "An argument for basic emotions", Cognition & Emotion, Vol. 6, No. 3. pp. 169–200, 1992

[3] I. Chiriacescu, Automatic Emotion Analysis Based on Speech, MSc Thesis, Delft University of Technology, 2009

[4] M. B. Mustafa, R. N. Ainon, R. Zainuddin, Z. M. Don, G. Knowles, S. Mokhtar, "Prosodic Analysis and Modelling for Malay", Malaysian Journal of Computer Science, Vol. 23, No. 2, pp. 102–110, 2010

[5] J. Rong, G. Li, Y. P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech", Information Processing & Management, Vol. 45, No. 3, pp. 315–328, 2009

[6] J. Pribil, A. Pribilova, "Determination of formant features in Czech and Slovak for GMM emotional speech classifier", Radioengineering, Vol. 22, No. 1, pp. 52–59, 2013

[7] M. El Ayadi, M. S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, Vol. 44, No. 3, pp. 572–587, 2011

[8] A. Utane, S. Nalbalwar, "Emotion recognition through Speech", 2nd National Conference on Innovative Paradigms in Engineering & Technology, International Journal of Applied Information Systems, pp. 5-8, 2013

[9] K. S. Rao, S. G. Koolagudi, R. R. Vempada, "Emotion recognition from speech using global and local prosodic features", International Journal of Speech Technology, Vol. 16, No. 2, pp. 143–160, 2013

[10] P. Olivier, J. Wallace, "Digital technologies and the emotional family", International Journal of Human-Computer Studies, Vol. 67, No. 2, pp. 204–214, 2009

[11] P. Pattnaik, "Impact of Emotion on Prosody Analysis", IOSR Journal of Computer Engineering, Vol. 5, No. 4, pp. 10–15, 2012

[12] W. L. Jarrold, Towards a theory of affective mind: Computationally modeling the generativity of goal appraisal, PhD Thesis, University of Texas at Austin, 2004

[13] S. A. Ali, M. Andleeb, N. G. Haider, D. R. Khan, "Evaluating the Performance of Learning Classifiers and Effect of Emotions and Spectral Features on Speech Utterances", International Journal of Computer Science and Information Security, Vol. 14, No. 10, pp. 406–412, 2016

[14] P. Boersma, D. Weenink, Praat: doing phonetics by computer, available at: http://www.fon.hum.uva.nl/praat/