I.J. Information Technology and Computer Science, **2015, 02, 54-59**
Published Online January 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2015.02.07

Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus

Syed Abbas Ali
Department of Computer & Information Systems Engineering, N.E.D University of Engineering & Technology, Karachi, Pakistan
E-mail: saaj@neduet.edu.pk

Anas Khan
Department of Telecommunications Engineering, N.E.D University of Engineering & Technology, Karachi, Pakistan
E-mail: anaskhan28@hotmail.com

Nazia Bashir
Department of Telecommunications Engineering, N.E.D University of Engineering & Technology, Karachi, Pakistan
E-mail: nazyabashir@outlook.com

***Abstract*—**Emotion plays a significant role in human perception and decision making whereas, prosodic features plays a crucial role in recognizing the emotion from speech utterance. This paper introduces the speech emotion corpus recorded in the provincial languages of Pakistan: Urdu, Balochi, Pashto Sindhi and Punjabi having four different emotions (Anger, Happiness, Neutral and Sad). The objective of this paper is to analyze the impact of prosodic feature (pitch) on learning classifiers (adaboostM1, classification via regression, decision stump, J48) in comparison with other prosodic features (intensity and formant) in term of classification accuracy using speech emotion corpus recorded in the provincial languages of Pakistan. Experimental framework evaluated four different classifiers with the possible combinations of prosodic features with and without pitch. An experimental study shows that the prosodic feature (pitch) plays a vital role in providing the significant classification accuracy as compared to prosodic features excluding pitch. The classification accuracy for formant and intensity either individually or with any combination excluding pitch are found to be approximately 20%. Whereas, pitch gives classification accuracy of around 40%.

***Index Terms*—**Prosodic Features, Learning Classifiers, Speech Emotion, Regional Languages of Pakistan.

I. Introduction

One of the natural and effective ways of communication among human beings is speech. Emotion is one of the speech-oriented application in which mental state of speaker conveys to others using spoken utterances termed as speech emotion recognition (SER) [1]. In speech emotion recognition, emotional state of the speaker can be extracted from his or her spoken utterances. There are few common emotions including sadness, Anger, Happiness, Neutral are used to identify the speech emotion from the spoken utterances using machine learning system with limited computational resources [2]. Automatic speech emotion recognition with the help of learning machine playing a significant role in the field of human-machine interaction for improving the effectiveness of human machine interface [3]. Several applications of speech emotion recognition systems includes: 1) medical diagnosis for psychiatric patients 2) emotion analysis during telephonic conversation 3) mental stress analysis during human conversation 4) E-learning for student emotional state etc. SER is considered as a statistical pattern recognition problem which comprises of three core phases: (1) feature extraction, (2) feature selection and (3) pattern classification [4]. Acoustic features of speech signal such as the intensity, timing, pitch, articulation and voice quality highly associate with the underlying emotion [5]. Most of SER acoustic features can be divided in two main categories: prosodic features and spectral feature. Prosodic features of SER are commonly used to provide important emotional clues of the speaker [6,7]. Prosodic features are usually based on information such as intensity, formant and pitch etc. Whereas, spectral features contain the information acquired from the spectrum of speech. Spectral features provide corresponding information for prosodic features and express the frequency contents of the speech signal. In speech emotion recognition, prospective prosodic features are derived from each spoken utterance of speaker for computational mapping between speech patterns and emotions. Selected prospective prosodic features used for training and testing using different classification methods to recognize the speech emotions. Despite of the extensive efforts, classifying the prospective prosodic feature is still one of the challenging tasks among SER research communities [1]. Classification is the final phase of SER system. During 1990s, most of the SER systems were based on Linear Discriminant Classification (LDC) and Maximum Likelihood Bayes algorithm (MLB) [4]. In 2000, the most

significant classification techniques were based on Neural Network (NN) classification method for speech emotion recognition [8]. Around 2002, Hidden Markov Model (HMM) [10] and Support Vector Machine (SVM) [9] have received significant attention among research communities. Although different classifiers have its own pons and corns, but researchers are still trying to find the optimal solution. This paper is an attempt to analyzing the impact of prosodic feature (pitch) to observe the behavior of learning classifiers in term of classification accuracy on demonstrative speech emotion corpus taken from random, non-actors and daily life peoples.

Rest of the paper is organized as follow. The consequent section discusses the three prosodic features of speech emotion including the four learning classifiers. Demonstrative speech emotion corpus collection and specifications are defined in section III. The experimental results and discussions are presented in section IV. Finally, conclusions are drawn in section V.

II. Prosodic Features and Learning Classifiers

Prosody is the study of the entire elements of language that contribute toward rhythmic and acoustic effects. Prosody is the combination of pitch, energy variation and duration of speech segment, which added sense to the spoken utterances to provide speaker emotion such as sadness, anger, happiness, neutral etc. Prosodic features are treated as major correlates of vocal emotion for discriminating and identifying the emotion from spoken utterances and emotion present in daily life conversation respectively [11]. Speaker emotional state can be indicated by prosodic features, some of the significant prosodic features used to recognize emotion from speech utterances are energy, speech rate, pitch, duration, intensity, formant, Mel frequency cepstrum coefficient (MFCC) and linear prediction cepstrum coefficient (LPCC) [12,13,14] Three prosodic features (intensity, pitch and formant) with the possible combinations of features with and without pitch were used in this experimental framework to analyzing the impact of pitch on learning classifiers.

- Pitch: One of the important perceptual property based prosody features used to detect emotion from spoken utterances are called pitch or glottal wave form. Vibration rate of vocal cord produces pitch signal and depends on the sub glottal air pressure and tension of vocal cord [15]. .Pitch has psycho acoustical sound attribute rather than objective physical property and can be measured as frequency.
- Intensity: Intensity is used to encode prosodic information and shows emotion of spoken utterance in term of energy of speech signal which depend on short term energy and short term average amplitude [16]. Energy of speech signal affected by stimulation level of emotions, due to which intensity can be used in the field of emotion recognition [17].
- Formants: Formant is one of the important prosodic feature and significant frequency component of speech which provides quantifiable frequency content of the characteristic and vowel of speech signal. Formant is defined as resonant frequency and unique frequency component of vocal track filter and human speech respectively [18].

Classification is a machine learning based data mining techniques used to classify each item in a data set into one of the predefined set of classes or groups. The most commonly used speech emotion learning classifiers are Naïve Bayes (NB), Support vector machine (SVM), C.45 decision stump, artificial neural network (ANN) and K-nearest-neighbor methods (k-NN). These learning classifiers have been compared on speech emotion assets in [19]. Experimental framework in this study shows that the classification accuracy for J48, Decision stump, adaboostM1 and Classification via regression found to give significant results in the presence of prosodic feature (pitch) as compared to other classifiers using WEKA data mining software.

J48: A Weka's implementation of C4.5 decision tree algorithm. A greedy approach is implemented using C4.5 method which built decision tree in top-down recursive divide and conquer fashion. In Top down approach, training set is recursively divided into smaller subsets as the tree is being constructed with a set of labeled training sample and their associated class labels [20].

AdaboostM1: Class for boosting a nominal class classifier using the Adaboost M1 method. Only nominal class problems can be tackled and dramatically improves performance [21].

Classification via regression: Class for doing classification using regression methods. Class is binarized and one regression model is built for each class value [22].

Decision stump: Class for building and using a decision stump and used in conjunction with a boosting algorithm. Decision stump performs classification (based on entropy) or regression (based on mean-squared error) and treating missing data as a separate value [23].

III. Corpus Collection and Specifications

The emotion corpus for this research work has been collected in five provincial languages of Pakistan: Urdu, Pashto, Punjabi, Sindhi and Balochi with four different emotions (Anger, Happiness, Neutral and Sad). In this initial research studies, the speech samples were taken from random, non-actors, daily life peoples with an aim of evaluating real time speech emotion samples for practical implementation. 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 based on well-known desiderata, according

to which sentence should be 1) semantically neutral, 2) easy to analyze, 3) consistent with any situation presented and 4) having similar meaning for each languages. Based on the previous studies [24], the carrier sentence was: "It will happen in seven hours"

Urdu	یہ سات گھنٹے میں ہو گا
Pashto	کیږي ساعتو اوو دابهپه
Punjabi	نگاوِچگھنٹےستایہ ہو
Sindhi	ستهي .ٽيندو ۾ ڪن ڪلا
Balochi	بئيتہاکلاکہفتهي

IV. Experimental Results and Discussions

To evaluate the impact of prosodic feature (pitch) on learning classifiers, the experimental framework is divided in to two portions with the possible combinations of prosodic features with and without pitch. In the first portion, we made use of PRAAT software [25] to observe the emotion present in the speaker utterances with four emotions (Anger, Happiness, Neutral and Sad). Demonstrative speech emotion corpus used in this experiment initially consists of 40 samples taken from recording of male and female speakers in the provincial languages of Pakistan (Urdu, Balochi, Pashto Sindhi and Punjabi). Two speaker's one male and one female (ages of 22- 26 years) of five provincial languages were spoken in four different emotions to analyze the dependence of emotions on prosodic features (pitch, Intensity and Formant).

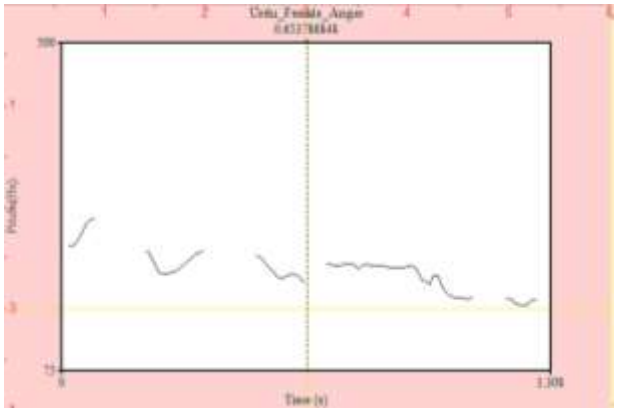

Fig. 1. Prosodic Features in Urdu Female (Anger)

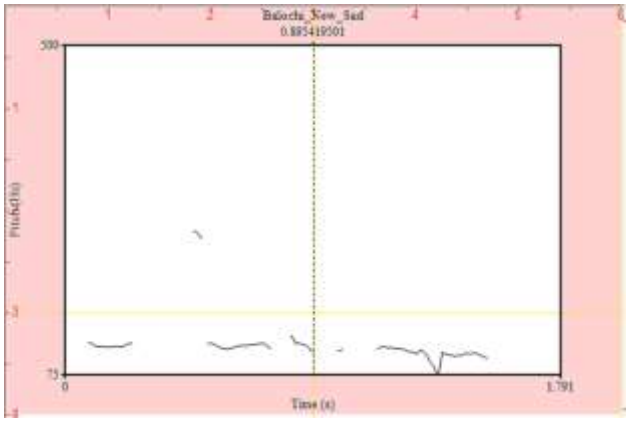

Fig. 2. Prosodic Features in Balochi Male (Sad)

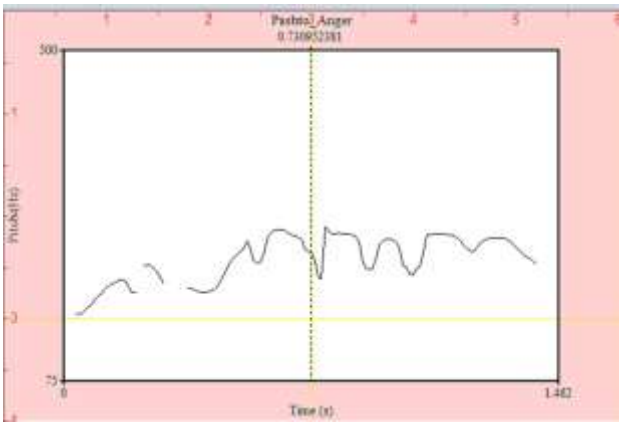

Fig. 3. Prosodic Features in Pashto Female (Anger)

Fig.1 to Fig. 3 provide the pictorial view of the prosodic feature (pitch) of the spoken utterances in Urdu, Balochi and Pashto with following emotions Anger, Sad and Anger respectively. Following observations have been collected using PRAAT software to extract prosodic features in order to show the behavior of prosodic feature (pitch). These observations are not completely demonstrated the total 40 samples for entire emotions and languages but just to show a flavor of how these observations were extracted using PRAAT.

Figure 4 to Figure 7 show the comparative analysis of prosodic features (pitch, intensity and formant) in five provincial languages of Pakistan with four different emotions (Anger, Happiness, Neutral and Sad). The tables describe the mean values of intensity, pitch and formant with the deviation among them in term of spoken emotion utterances. The comparison has been made on the basis of the mean values of spoken emotion utterances. It can be observe from the demonstrative experiments that intensity shows quite similar value for all provincial languages but also the value is very close to four emotions and for our experimental work to recognize emotion, threshold values based on intensity will not help accurately. However pitch in these graphs can be clearly observe that perform a very significant role as it shows substantial variation for four different emotions although these values are a little bit varies for five provincial languages but it can be used for detecting emotions independent of region and gender. From the above comparative analysis it can be conclude that "pitch" show considerable variation in mean values of four different emotions which may lead to better classification accuracy using learning classifiers.

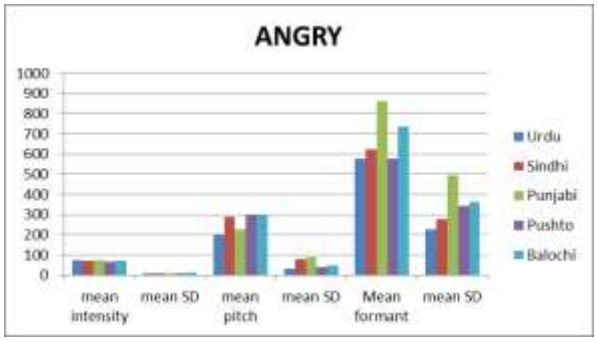

Fig. 4. Comparative analysis of prosodic feature for the emotion Anger

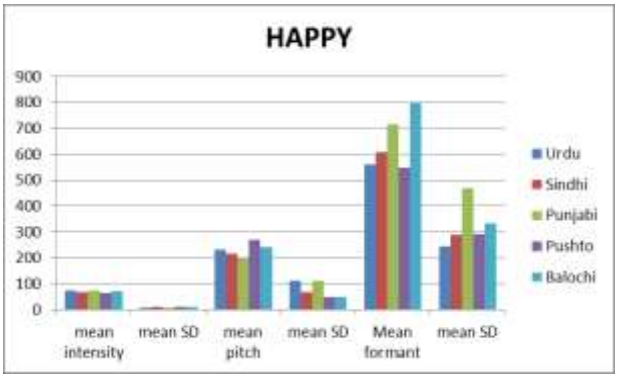

Fig. 5. Comparative analysis of prosodic feature for the emotion Happy

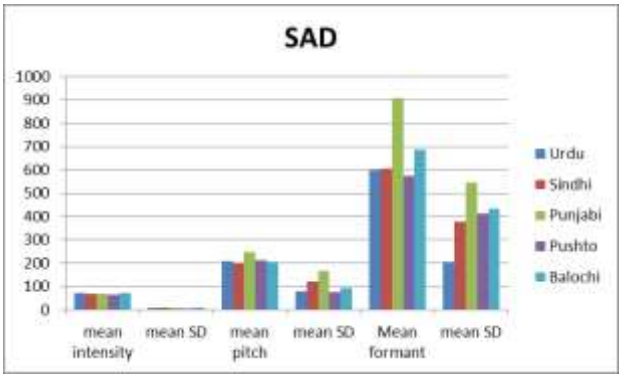

Fig.7. Comparative analysis of prosodic feature for the emotion Sad

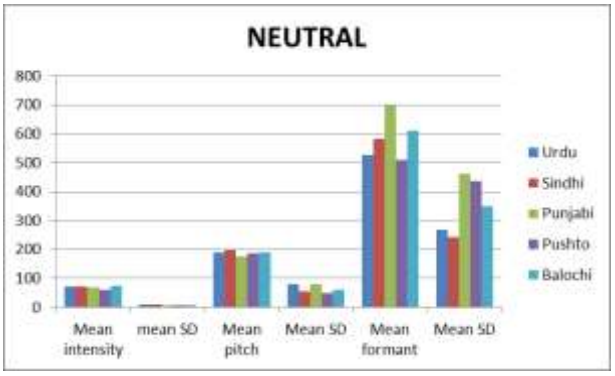

Fig. 6. Comparative analysis of prosodic feature for the emotion Neutral

In the second portion of experiment, we made use of WEKA (Data Mining) software tool [26] to analyze the classification accuracy of four learning classifiers (adaboostM1, classification via regression, decision stump, J48) with and without the prosodic feature (pitch). After extracting prosodic features (pitch, intensity and formant), our experimental flow was directed towards WEKA software to evaluate the impact of prosodic feature (pitch) on learning classifiers. During experiments different learning classifiers were tested in order to investigate which of them gives the best possible results and four learning classifiers (adaboostM1, classification via regression, decision stump and J48) were found the best.

Table 1. Classification accuracy of learning classifier excluding Pitch

Prosodic feature(s)	Learning Classifier	Total no. of Instances	No. of correct instances	No. of incorrect instances	Time to build model (sec)	Classification accuracy
Intensity	adaboostM1	40	8	32	0	20%
	Classification via regression	40	8	32	0.08	20.00%
	Decision stump	40	8	32	0	20%
	J48	40	9	31	0	22%
Formant	adaboostM1	40	8	32	0	20%
	Classification via regression	40	10	30	0.06	25.00%
	Decision stump	40	8	32	0	20%
	J48	40	7	33	0	18%
Intensity & Formant	adaboostM1	40	7	33	0	18%
	Classification via regression	40	8	32	0.16	20.00%
	Decision stump	40	7	33	0	18%
	J48	40	4	36	0	10%

Table 1 and Table 2 provide comprehensive table for performance evaluation of learning classifiers in term of classification accuracy with possible combination of prosodic feature with and without pitch. Experimental results clearly evident that the classification accuracy for formant and intensity either individually or with any combination excluding pitch are found to be approximately 20%. whereas, pitch gives classification accuracy of around 40%.

Figure 8 provides the comprehensive statistical analysis of prosodic features with and without pitch and it can be clearly observed that pitch is indeed performing a significant role in performance of learning classifiers.

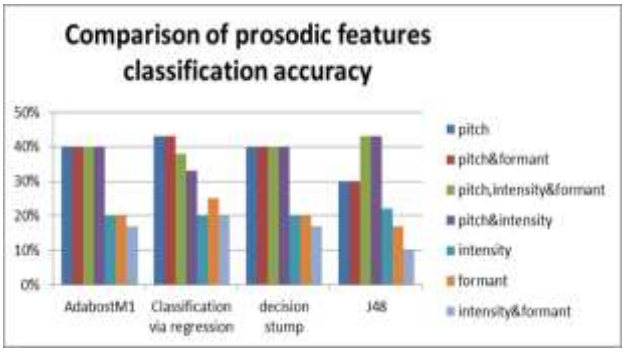

Fig. 8. Comparison of PITCH with other prosodic features on learning classifiers

Table 2. Classification accuracy of learning classifier including PITCH

Prosodic feature(s)	Learning Classifier	Total no. of Instances	No. of correct instances	No. of incorrect instances	Time to build model (sec)	Classification accuracy
Pitch	adaboostM1	40	16	24	0.02	40%
	Classification via regression	40	17	23	0.08	43%
	Decision stump	40	16	24	0	40%
	J48	40	12	28	0	30%
Pitch & Intensity	adaboostM1	40	16	24	0	40%
	Classification via regression	40	13	27	0.19	33%
	Decision stump	40	16	24	0	40%
	J48	40	17	23	0	43%
Pitch & Formant	adaboostM1	40	16	24	0.02	40%
	Classification via regression	40	17	23	0.05	43%
	Decision stump	40	16	24	0	40%
	J48	40	12	28	0	30%
Pitch, Intensity & formant	adaboostM1	40	16	24	0	40%
	Classification via regression	40	15	25	0.07	38%
	Decision stump	40	16	24	0	40%
	J48	40	17	23	0	43%

V. Conclusion

In this paper, demonstrative speech emotion corpus recorded in five regional languages of Pakistan: Urdu, Balochi, Pashto Sindhi and Punjabi having four different emotions (Anger, Happiness, Neutral and Sad). This study is an attempt to analyzing the impact of prosodic feature (pitch) on four learning classifiers (adaboostM1, classification via regression, decision stump, J48) in term of classification accuracy. Demonstrative experiments have been performed using PRAAT software and WEKA tools to observe the emotion present in the speaker utterances and evaluate the performance of learning classifiers with and without pitch respectively. Experimental results clearly show that the pitch gives classification accuracy of around 40%. Whereas, formant and intensity either individually or with any combination excluding pitch are found to be approximately 20%. For future research, authors are considering other prosodic features (shimmers, tonal and non-tonal etc.) to analyzing the performance of learning classifiers.

References

[1] M. E. Ayadi, M. S. Kamel and F. Karray, 'Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition, 44(16), 572-587, 2011.

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

[3] I. Chiriacescu, "Automatic Emotion Analysis Based On Speech", M.Sc. THESIS Delft University of Technology, 2009.

[4] J. Rong, G. Li, Y.P. Phoebe Chen, "Acoustic feature selection for automatic emotion recognition from speech", Information Processing and Management, 45, pp. 315–328, 2009.

[5] M.B. Mustafa, R.N. Ainon1, R. Zainuddin, Z.M. Don, G. Knowles, S. Mokhtar, "Prosodic Analysis And Modelling For Malay Emotional Speech Synthesis" , Malaysian Journal of Computer Science, pp. 102-110. 2010.

[6] S. Wu, T.H. Falk, W.Y. Chan, "Automatic speech emotion recognition using modulation spectral features", Speech communication,vol. 53,pp. 768–785,2011.

[7] 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.

[8] J. Nicholson, K. Takahashi, R. Nakatsu, "Emotion recognition in speech using neural networks", Neural Computation. Appl. Vol. 9, pp. 290–296, 2000.

[9] Z.-J. Chuang and C.-H. Wu, "Emotion recognition using acoustic features and textual content", In Proc of IEEE international conference on multimedia and expo (ICME'04), Vol. 1, pp. 53–56, IEEE Computer Society, 2004.

[10] M. Song, C. Chen, and M. You, "Audio-visual based emotion recognition using tripled hidden markov model", In Proceedings of IEEE international conference on acoustic, speech and signal processing (ICASSP'04), Vol. 5, pp. 877–880, IEEE Computer Society, 2004.

[11] S.R.Karathapalli and S.G.Koolagudi, "Emotion recognition using speech features", Springer Science+ Business Media New York, 2013.

[12] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "Speech Emotion Recognition Using Hidden Markov Model", Euro speech, 2001.

[13] P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference on Electronic And Mechanical Engineering And Information Technology, 2011.

[14] Z. Ciota, "Feature Extraction of Spoken Dialogs for Emotion Detection", ICSP, 2006.

[15] A.S. Utane and S.L. Nalbalwar, "Emotion recognition through Speech" International Journal of Applied Information Systems (IJAIS), pp.5-8, 2013.

[16] P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine",

International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.
[17] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features and Methods", Elsevier Speech communication, vol. 48, no. 9, pp. 1162-1181, September, 2006.
[18] E. Bozkurt, E, Erzin, C. E. Erdem, A. Tanju Erdem, "Formant Position Based Weighted Spectral Features for Emotion Recognition", Science Direct Speech Communication, 2011.
[19] 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.
[20] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Elsevier, 2nd edition, 2006.
[21] Yoav Freund, Robert E. Schapire: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, San Francisco, 148-156, 1996.
[22] E. Frank, Y. Wang, S. Inglis, G. Holmes, I.H. Witten (1998). Using model trees for classification. Machine Learning. 32(1):63-76.
[23]]http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/DecisionStump.html
[24] S.A. Ali., S Zehra., et.al. "Development and Analysis of Speech Emotion Corpus Using Prosodic Features for Cross Linguistic", International Journal of Scientific & Engineering Research, Vol. 4, Issue 1, January 2013.
[25] [http://www.fon.hum.uva.nl/praat/
[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.

Authors' Profiles

Syed Abbas Ali: Male, Karachi, Pakistan, Department of Computer Science & Information Technology, his research directions include Machine Learning Algorithms and Speech Emotion Recognition.

Anas Khan: Male, Karachi, Pakistan, Department of Telecommunications Engineering, N.E.D University of Engineering & Technology, and his research interest include digital signal processing.

Nazia Bashir: Female, Karachi, Pakistan, Department of Telecommunications Engineering, N.E.D University of Engineering & Technology, and her research interest include digital signal processing.

How to cite this paper: Syed Abbas Ali, Anas Khan, Nazia Bashir,"Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus", International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.2, pp.54-59, 2015. DOI: 10.5815/ijitcs.2015.02.07

