Performance Evaluation of Learning Classifiers of Children Emotions using Feature Combinations in the Presence of Noise

Abdul Samad
Department of Computer Science
Hamdard University
Karachi, Pakistan
asamad23@gmail.com

Aqeel-ur-Rehman
Department of Computer Science
Hamdard University
Karachi, Pakistan
aqeelrehman1972@gmail.com

Syed Abbas Ali
Department of Computer & Information Systems Engineering, NED University of Engineering & Technology, Pakistan
saaj@neduet.edu.pk

***Abstract*—Recognition of emotion-based utterances from speech has been produced in a number of languages and utilized in various applications. This paper makes use of the spoken utterances corpus recorded in Urdu with different emotions of normal and special children. In this paper, the performance of learning classifiers is evaluated with prosodic and spectral features. At the same time, their combinations considering children with autism spectrum disorder (ASD) as noise in terms of classification accuracy has also been discussed. The experimental results reveal that the prosodic features show significant classification accuracy in comparison with the spectral features for ASD children with different classifiers, whereas combinations of prosodic features show substantial accuracy for ASD children with J48 and rotation forest classifiers. Pitch and formant express considerable classification accuracy with MFCC and LPCC for special (ASD) children with different classifiers.**

Keywords-spoken utterances; special children; learning classifiers; noise; features

I. INTRODUCTION AND RELATED WORK

In the modern era of human-computer interaction, emotional speech recognition is a field of vast concern. SER has a great influence on human behavior and is a key point to build relations. Different emotions have their own characteristics that make it memorable in their own way [1]. Furthermore, "EmoChildRu" has been introduced [2] as the first child emotion database created to recognize speech and voice emotions from children's behavior. Two child emotional speech examination probes are reported in the context of the corpus: grown-up audience members and programmed listeners. Automatic classification results are fundamentally the same as human discernment, despite the fact that the precision is underneath 55 % for both, demonstrating the trouble of child emotion recognition from speech under natural conditions. To improve the state of people with ASD, a few CAL procedures have been executed. Authors in [3] depict the investigation of fluctuated CAL strategies actualized to enhance the everyday life states of such individuals. Furthermore, the authors briefed about the CAL strategies involved in various applications improving correspondence, behavioral and social abilities of such special children. In [4], it was noticed that it is not easy to observe mental and emotional conditions in autism spectrum conditions (ASC) aspects. A technique was proposed to recognize their speech in [5] independently from the speaker. In this technique, MEL & BARK scale, Equivalent Rectangular Bandwidth in filter space along with gamma toned features were utilized at the front end, whereas at the back end, Fuzzy C Means (FCM), Multivariate Hidden Markov Models (MHMM) and Vector Quantization (VQ) approaches were applied. Individual words and short sentences in Tamil language were used to evaluate the performance by three variants. The data of two speakers were tested against the features of eight speakers. A prominent real situational database was organized in [6] to detect fear thorough the feature classifier "interjection" through speech in extreme emotional and real world emergencies. MFCCs along with Support Vector Machine with variant interjections were utilized to categorize speech emotions. In [7], Urdu language has been taken to recognize emotions from primary age children. The authors used 3 different prosodic features with 5 different classifiers and four emotions. They reported that J-48 classifier achieved the highest accuracy. A profound architecture was utilized in [8], which uses a convolutional network for extricating space shared highlights and a long short-term memory network for arranging emotions utilizing space particular features. A complete cross-corpora exploration of different avenues regarding various speech emotions areas uncovers that transferable features give increases extending from 4.3% to 18.4% in speech emotions recognition. The fundamental of Deep Neural Networks (DNNs) is to perceive human emotions from a speech signal. Mel-frequency Cepstral Coefficient (MFCC) is selected as a one of the frequently used speech features from crude sound information. In next step, DNN nourished the extricated speech features to prepare the system. Also, a hand crafted database was presented improving the utilization of the system [9]. The work-related recognition, classification, emotion detection children with ASD is still an open topic of research, while researchers are now more concerned to help these children by making them realize the emotions in the real world. Under this scope, an autism-based game "emotify" has been developed [10]. It comprises two levels of difficulty and attempts to teach children about neutral, anger, sadness and happiness emotions.

Corresponding author: Abdul Samad

At the second level, children are helped in expressing their feelings which would be evaluated and examined. Machine learning approaches are exploited to develop a multilingual emotion recognition system. This paper evaluates the performance of learning classifiers when dealing with prosodic and spectral features and their combinations considering special (ASD) children as noise in terms of classification accuracy.

II. LEARNING CLASSIFIERS AND FEATURES

In daily routine conversations the prosodic features play a vital role [11]. The parameters used in expressing the speech to perceive the feelings of users are speech rate, length, pitch, formant, intensity, Mel frequency cepstrum coefficient (MFCC) and LPCC [12]. Two spectral features and three prosodic features (intensity, pitch, and formant) and their potential combinations are utilized in this research.

- Pitch: Pitch and frequency are correlated. The analysis of every speech frame is obtained by their statistical values throughout the sample. These values [13] depict the clear picture of properties of audio parameters.
- Intensity: Intensity demonstrates the prosodic feature encoding and the emotion based spoken utterance expressions [12].
- Formant: Formant is a critical recurrence segment of speech which gives quantifiable results of the consonants and vowels of the speech signal [12].

Four learning classifiers were used in the experimental framework. The evaluation of the performance of these classifiers regards spectral and prosodic features and their combinations. The comprehensive description of the learning classifiers can be found in [14]. The classifiers are:

- J48: A family of decision tree algorithms used to figure the feature vector for different examples. The classes for the recently produced events are being learned on the basis of the training examples. With the support of tree grouping calculation the elementary dispersion of the information is successfully justifiable [15].
- Multi-Layer Perceptron (MLP): A class of deep artificial neural network that contains three layers at minimum. The first is the input layer while the last one is the output layer. The middle is a hidden layer and different MLPs can have various numbers of invisible layers [16].
- Rotation forest: Rotation forest [15] eliminates randomly any subsets of classes, performs Bootstrap on the remaining data and finally performs PCA and establishes free decision trees.
- Logit Boost Classifier: Boosting [16] works on the principal that a set of weak learners can be used to create a strong classifier. Logit Boost provides higher weights for misclassified classifiers.

III. CORPUS COLLECTION AND RECORDING SPECIFICATION

The corpus has been collected from both categories of children (normal and with ASD) in Urdu and it comprises of 200 samples, equally divided for both cases. As per the research methodology, ASD children have been considered as noise in the experimental framework. The recording specification has been considered in standard conditions with Signal-to-Noise Ratio $\geq$ 45dB. Microsoft Windows 7 sound recorder has been utilized to record the emotion based spoken utterances of normal and special (ASD) children. The configuration is 16 bit, Mono, PCM with a testing rate of 48KHz with Microphone hazard and awareness of 54dB±2dB and 2.2W separately, 3.5mm mash stereo and link length of 1.8m. The choice of a spoken utterance incorporated these qualities a) semantically impartial, b) simple to investigate, c) reliable with any circumstance exhibited, and d) having comparable importance for every dialect. The sentence was: — "Mujhe Khelna Hai" or "I have to play".

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of learning classifiers is evaluated in the experimental framework for normal and special children making use of prosodic and spectral features and their combinations on spoken utterances recorded in Urdu. The corpus collection comprises of 200 spoken utterance samples in different emotions equally distributed in normal and special children. The experimental framework further classifies inter and intra feature combinations with four different classifiers (logit boost, MLP, J48 and rotation forest) with the following feature configurations: 1) separate prosodic and spectral features, 2) combinations of all three prosodic features (intensity, formant, and pitch) with two spectral features (LPCC and MFCC). The objective of the proposed framework is to identify the behavior of the four classifiers in a single feature or different combinations of spectral and prosodic features on spoken utterances of special (ASD) and normal children in terms of classification accuracy. The experimental results for both corpuses taken from normal and special (ASD-treated as noise) children in Urdu follow.

A. Prosodic Features

Pitch demonstrates great precision in portraying the states of children under all four classifiers while classifying special children more precisely than normal children (Table I). The classification accuracy of rotation forest with prosodic feature pitch for ASD (noisy) children spoken utterances was significantly better than the accuracy of the other classifiers. Intensity also shows good classification accuracy for ASD children for all classifiers except from rotation forest. All learning classifiers demonstrate higher classification accuracy with formant for ASD children in comparison with normal children

B. Spectral Features

MFCC has significant accuracy with MLP and logit boost classifier in case of normal children, on the other hand, only rotation forest shows considerable classification accuracy for ASD children. LPCC has very good accuracy only from logit boost classifier in both normal and special children. The outcome demonstrates that in any study of LPCC the logit boost classifier ought to be utilized.

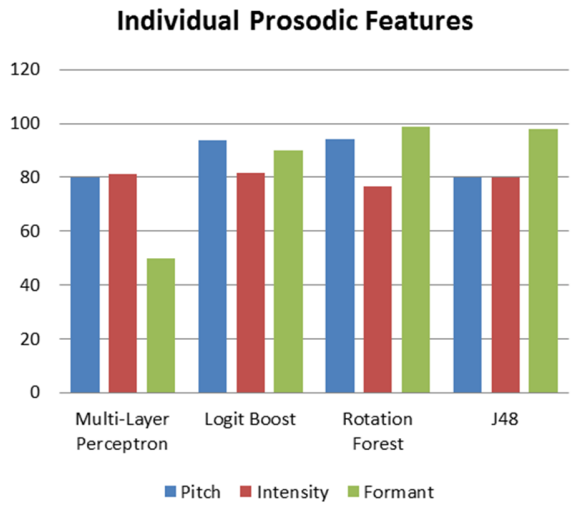

Fig. 1. Prosodic features

TABLE I. CLASSIFICATION ACCURACY FOR PROSODIC FEATURES

Feature	Classifier	ASD children classification accuracy (%)	Normal children classification accuracy (%)
Intensity	MLP	81.3	65.6
	Logit boost	81.5	90.5
	Rotation forest	76.5	64.5
	J48	81.3	65.6
Pitch	MLP	80	71.4
	Logit boost	93.8	72
	Rotation forest	94.4	76.7
	J48	80	71.4
Formants	MLP	50	11
	Logit boost	90	78.6
	Rotation forest	99	63.2
	J48	98	60

TABLE II. CLASSIFICATION ACCURACY FOR SPECTRAL FEATURES

Features	Classifier	ASD children classification accuracy (%)	Normal children classification accuracy (%)
MFCC	MLP	64.7	85.7
	Logit boost	64.7	85.7
	Rotation forest	83.3	61.1
	J48	10	50
LPCC	MLP	50	10
	Logit boost	84.6	62.9
	Rotation forest	9	50
	J48	11	50

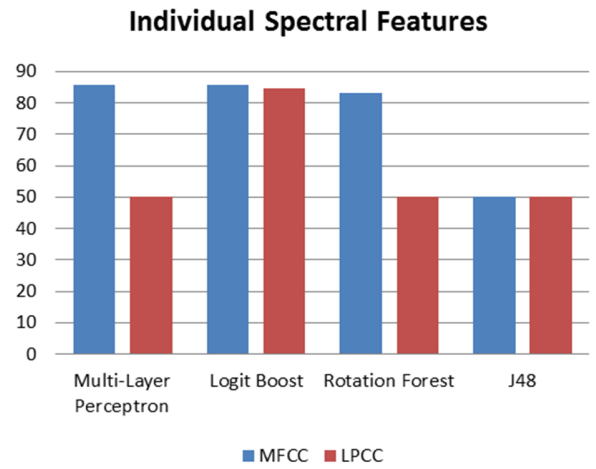

Fig. 2. Spectral features

C. Inter Combination of Prosodic and Spectral Features

The prosodic feature pitch shows significant accuracy with the combination of two other prosodic features. In combination with intensity, pitch illustrates considerable classification accuracy with logit boost and J48 for ASD (special) children. Pitch with formant performs well in classifying special (ASD) children with all classifiers except logit boost. In combination with intensity and formants, MLP and rotation forest show significant classification accuracy in comparison with the other two classifiers for ASD children. J48 has comparable accuracy in classifying special and normal children.

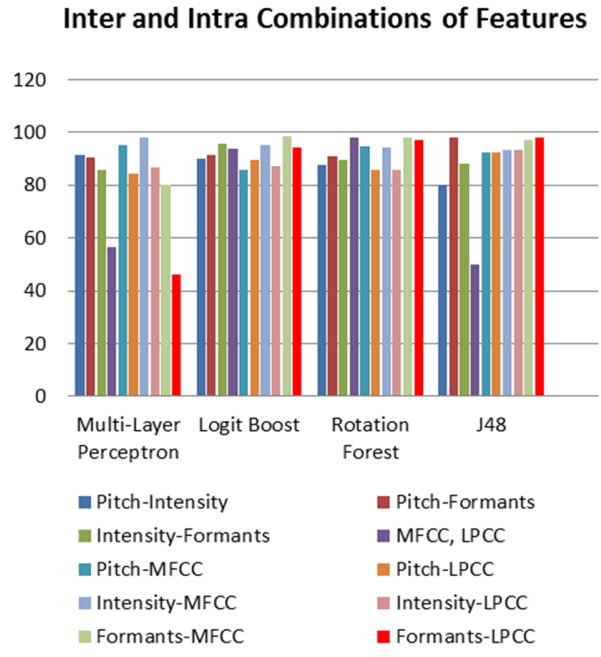

Fig. 3. Inter and intra combinations of features

D. Intra Combination of Prosodic and Spectral Features

In Intensity-MFCC, MLP performs better in classifying the normal children, while logit boost is considerably good in

classifying noisy spoken utterances for special children. In this case, rotation forest and J48 are performing averagely. Similarly, in Intensity-LPCC combination, the MLP is good and accurate. Nevertheless, logit boost more accurately classifies the normal children's spoken utterances. Rotation forest and J48 perform averagely.

TABLE III. CLASSIFICATION ACCURACY FOR PROSODIC AND SPECTRAL FEATURES

Feature combination	Classifier	ASD children classification accuracy (%)	Normal children classification accuracy (%)
Pitch, intensity	MLP	88	91.3
	Logit boost	90	88
	Rotation forest	87.5	87.5
	J48	80	71.4
Pitch, formant	MLP	90.5	81.5
	Logit boost	88	91.3
	Rotation forest	90.9	84.6
	J48	98	60
Intensity, formant	MLP	85.7	77.8
	Logit boost	92	95.7
	Rotation forest	89.5	76
	J48	88.3	88.3
MFCC, LPCC	MLP	56.5	56
	Logit boost	93.8	71.9
	Rotation forest	98	57.1
	J48	11	50
Pitch, MFCC	MLP	95.2	85.2
	Logit boost	77.8	85.7
	Rotation forest	94.7	79.3
	J48	92.3	65.7
Pitch, LPCC	MLP	84.2	72.4
	Logit boost	89.5	75.9
	Rotation forest	85.7	77.8
	J48	92.3	65.7
Intensity, MFCC	MLP	98	80
	Logit boost	82.1	95
	Rotation forest	94.4	76.7
	J48	69.7	93.3
Intensity, LPCC	MLP	80.8	86.4
	Logit boost	87	84
	Rotation forest	85.7	77.8
	J48	69.7	93.3
Formant, MFCC	MLP	53.5	80
	Logit boost	98.4	75
	Rotation forest	98	57.1
	J48	97	60
Formants, LPCC	MLP	46.4	45
	Logit boost	94.1	74.2
	Rotation forest	97	61.5
	J48	98	60

Table III provides the results of the learning classifiers with combinations of prosodic and spectral features for classifying the noisy spoken utterances in terms of classification accuracy. The most significant results can be observed of the combination of LPCC with pitch and formant in classifying the noisy (special children) utterances, whereas, pitch and formant with MFCC also show substantial classification accuracy for special (noisy) children. Other combinations such as intensity and LPCC and intensity with MFCC also provide considerable results in classification.

V. CONCLUSION

In this paper, the performance of learning classifiers has been evaluated considering prosodic and spectral features and their combinations for children with ASD in terms of classification accuracy. The experimental frame comprises four different classifiers with different inter and intra combinations of prosodic and spectral features. The experiments were conducted on a sample of 200 individuals equally taken from normal and children with ASD which were considered as noise. The conclusions of the experimental results are:

- The spectral features show significant classification accuracy with prosodic features (pitch & formant) with rotation forest and J48 classifiers.
- Separate analysis of the spectral and prosodic features reveals that the classification accuracy of prosodic features is considerably better than of spectral features.
- The intra feature combinations of spectral features with pitch and formant demonstrate better classification accuracy for different classifiers.

The authors are now focusing on developing an experimental framework to perform the same methodology with different emotions for evaluating the performance of classifiers to special children (ASD) in terms of classification accuracy.

REFERENCES

[1] S. Ramakrishnan, "Recognition of emotion from speech: A review", in: Speech enhancement, modeling and recognition algorithms and applications, pp. 121-137, InTech, 2012

[2] E. Lyakso, O. Frolova, E. Dmitrieva, A. Grigorev, H. Kaya, A. A. Salah, A. Karpov, "EmoChildRu: Emotional child russian speech corpus", Lecture Notes in Computer Science, Vol. 9319, Springer, Cham, 2015

[3] S. Dewan, A. Singh, L. Singh, S. Gautam, "Role of emotion recognition in computive assistive learning for autistic person", Indian Journal of Science and Technology, Vol. 9, No. 48, 2016

[4] O. Golan, Y. Sinai-Gavrilov, S. Baron-Cohen, "The Cambridge mindreading face-voice battery for children (CAM-C) complex emotion recognition in children with and without autism spectrum conditions", Molecular Autism, Vol. 6, No. 1, Article ID 22, 2015

[5] R. Arunachalam, Revathi, "A strategic approach to recognize the speech of the children with hearing impairment different sets of features and models", in: Multimedia Tools and Applications, Springer, 2019

[6] S. A. Yoon, G. Son, S. Kwon, "Fear emotion classification in speech by acoustic and behavioral cues", Multimedia Tools and Applications, Vol. 78, No. 2, pp. 2345-2366, 2019

[7] S. Khan, S. A. Ali, J. Sallar, "Analysis of children's prosodic features using emotion based utterances in Urdu language", Engineering, Technology & Applied Science Research, Vol. 8, No. 3, pp. 2954-2957, 2018

[8] A. Marczewski, A. Veloso, N. Ziviani, "Learning transferable features for speech emotion recognition", Thematic Workshops of ACM Multimedia, Mountain View, USA, October 23-27, 2017

[9] M. F. Alghifari, T. S. Gunawan, M. Kartiwi, "Speech emotion recognition using deep feedforward neural network", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 10, No. 2, pp. 554-561, 2018

[10] A. Rouhi, M. Spitale, F. Catania, G. Cosentino, M. Gelsomini, F. Garzotto, "Emotify: emotional game for children with autism spectrum disorder based-on machine learning", 24th International Conference on Intelligent User Interfaces Companion, New York, USA, March 16-20, 2019

[11] K. S. Rao, S. G. Koolagudi, Emotion recognition using speech features, Springer, 2013

[12] P. Shen, C. Zhou, X. Chen, "Automatic speech emotion recognition using support vector machine", International Conference on Electronic, Mechanical Engineering and Information Technology, Harbin, China, August 12-14, IEEE 2011

[13] A. S. Utane, S. L. Nalbalwar, "Emotion recognition through speech", 2nd National Conference On Innovative Paradigms in Engineering & Technology, Nagpur, Maharashtra, India, February 17, 2013

[14] S. A. Ali, S. Zehra, M. Khan, F. Wahab, "Development and analysis of speech emotion corpus using prosodic features for cross linguistic", International Journal of Scientific & Engineering Research, Vol. 4, No. 1, pp. 1-8, 2013

[15] 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

[16] M. Swain, A. Routray, P. Kabisatpathy, "Databases features and classifiers for speech emotion recognition: a review", International Journal of Speech Technology, Vol. 21, No. 1, pp. 93-120, 2018