Emotion Recognition of Low Resource (Sindhi) Language Using Machine Learning

Tanveer Ahmed*, Sajjad Ali Memon*, Saqib Hussain*, Amer Tanwani, Ahmed Sadat***
tanveermemon92@yahoo.com
*Mehran University of Engineering & Technology, Jamshoro, Pakistan.
**Ghulam Ishaq Khan Institute of Engineering Science & Technology, KPK, Pakistan

Summary: One of the most active areas of research in the field of affective computing and signal processing is emotion recognition. This paper proposes emotion recognition of low-resource (Sindhi) language. This work's uniqueness is that it examines the emotions of languages for which there is currently no publicly accessible dataset. The proposed effort has provided a dataset named MAVDESS (Mehran Audio-Visual Dataset Mehran Audio-Visual Database of Emotional Speech in Sindhi) for the academic community of a significant Sindhi language that is mainly spoken in Pakistan; however, no generic data for such languages is accessible in machine learning except few. Furthermore, the analysis of various emotions of Sindhi language in MAVDESS has been carried out to annotate the emotions using line features such as pitch, volume, and base, as well as toolkits such as OpenSmile, Scikit-Learn, and some important classification schemes such as LR, SVC, DT, and KNN, which will be further classified and computed to the machine via Python language for training a machine. Meanwhile, the dataset can be accessed in future via https://doi.org/10.5281/zenodo.5213073.

Keywords: *Emotion Recognition, Signal Processing, Affective Computing, Machine Learning, Dataset.*

1. INTRODUCTION:

What Exactly is an Emotion?

Emotions are described as powerful sensation such as fear or love as well as the portion of person's personality that is made up of feelings. Although, there is no unanimity on a definition of emotion in the scientific literature in the field of psychology. A mental consciousness experience with high potency and strong voluptuary content (pleasure/displeasure) is defined as an emotion. Meanwhile, some people view emotions as a complex psychological event including a variety of responses, including a physiological response, an expressive response, and a subjective experience. Feelings and, by extension, emotions, are an important aspect of human activity. Emotions have a significant influence in how people think and behave, thus analyzing how people express their emotions can provide insight into their cognitive processes.

In the era of industry 4.0, there's an increasing impulse among researchers to enable more natural contact between robots and humans. When gadgets & machines can comprehend, explicate, & identify human emotion, this becomes conceivable. Analysts in the fields of signal processing & affective computing have investigated the creation of computer approach for emotion identification from a variety of paradigms, including voice, facial expressions, text, and physiological data, in order to achieve this.

Speech is particularly intriguing among these modalities since it is the most natural method for humans to express emotions. Speech emotion recognition can assist emergency services and healthcare professionals in addition to offering social intelligence to machines. Furthermore, the emotion recognition systems that are associated to exigency services for instance, call centers are used to assess the caller's level of anguish, as a result, allocate their information to a nearest health center.

How Important is Multimodal Communication

Emotion that can be conveyed in a single somatic, usually through facial expressions, has become popular in emotion research. Communication that involves emotions in the natural world, on the other hand, is temporal and multimodal. Multisensory integration is important for processing emotional inputs, according to research [1, 2]. Researchers have been inspired to build their own multimodal stimuli due to a lack of verified multimodal sets. [3, 4].

Why we need Facial Expressions

Faces are rarely, if ever, immobile in normal speech, and expressions vary widely. However, the majority of the sets feature solely static facial pictures. There's a lot of evidence today that facial movement helps with affective processing. Dynamic expressions elicit distinct and increased patterns of brain activity than static expressions, according to imaging studies. Dynamic stimuli generate greater imitation responses in viewers' facial muscles than static emotions, according to electromyographic research [5]. As a result, facial expressions which are dynamic could depict more realistic emotion than facial emotions that are static [6].
Emotional responses that are linked to video content recognized to be most confusing piece of work that people may do. In the realm of affective computing, there has been a tendency to construct a stimulus library and identify human emotions elicited by watching video clips. When

Manuscript received August 5, 2021
Manuscript revised August 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.8.46

watching a video, one may feel reliant on his or her intellectual judgement and interpretation of the situation shown in the film. As a result, it's critical to comprehend a human's cognitive understanding of a circumstance and how it relates to their emotions.

2. RELATED WORK:

Although there is a rising interest in identifying emotions using video stimuli, many questions remain, such as how and to what extent videos may trigger sentiments. To begin answering these questions, we will need to put together a collection of video stimuli.
Since, everyone has done these tests on a pre-defined collection of Emotion Datasets from various languages all across the world. The originality of our planned work is not only that we will test this approach on our own Generated Datasets, but that we will also appropriately promote it in the area of research and development, raising Sindhi's standing as a low-resource language in affective computing. The main objective of this study is to create a set of stimuli that communicates emotions clearly and is accessible to everyone.

Moreover, there is plenty of study literature on emotion recognition, the vast amount of is written in Western languages such as French, English & German. Due to the fact that the bulk of datasets are in these languages. Based on our findings, despite the region's population of over 1.8 billion people, there's distinct lack of datasets from the Asian languages including Urdu & Sindhi.

Several academics have lately sought to develop & generate datasets for voice emotion detection of languages spoken in Asia. For instance, Koolagudi et al. [7] presented huge dataset for Telugu language. Telugu language is spoken mostly in south India. A total of 12 thousand pronouncement for eight different emotions are included in the dataset, including happy, anger, neutral, sorrow, sarcasm, disgust, surprise & fear. Furthermore, Syed et al. [8] presented the compilation named emotion-Pak, which comprised four emotions varying five widely used Pakistani languages. For each of the five languages, ten native speakers were used to record the data. Despite the fact that this dataset is extremely relevant to our research, we were unable to receive a response after demanding access to the Emotion-Pak Corpus from Syed.

We have presented a new speech emotion dataset with 1200 video recordings that are used for speech-based emotion identification to train machine learning models in Sindhi language, one of the most widely spoken languages in Asia. Sindhi is spoken by about 32 million people in South Asia, the majority of whom live in Pakistan's Sindh region, as well as one of India's recognized languages.

The remaining paper is laid out as follows: We developed the process for collecting the Sindhi speech emotion dataset in part 3, and we go through the methodology for determining the dataset's baseline classification performance in section 4. Section 5 contains the experimental data and discussion & conclusion.

3. Dataset Development

Dataset Collection:
In a research conducted by Ekman [9], there are six types of fundamental emotions namely anger, happy, sad, surprise, fear, disgust which was placed in this suggested collection of stimuli. To portray these emotions, footage from numerous films and shows of various genres were used. With the help of Fig 1, which depicts the structure of how data has been collected, we will explain the collection of data for the Sindhi Speech Emotion Corpus. We split the data into two halves as listed in Table 1: one half (600 emotions gleaned from the internet and YouTube) and the other half (600 emotions were recorded by random native Sindhi Speakers). Anger, Disgust, Happy, Sad, Surprise, and Fear were among the feelings experienced (100 each emotion).

Project Methodology:

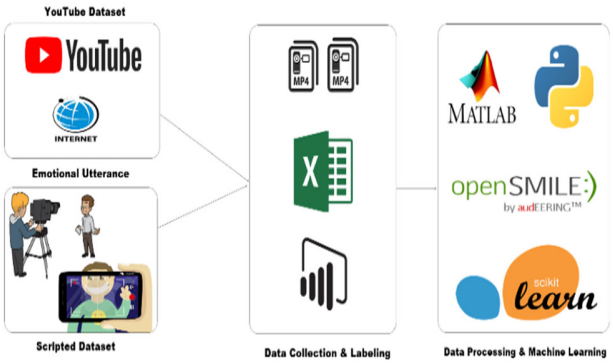


Figure 1. Methodology Diagram

Table 1. MAVDESS PER CLASS EMOTIONS:

Emotion	YouTube	Scripted/Acted
Anger	100	100
Disgust	100	100
Happy	100	100
Surprise	100	100
Sad	100	100
Fear	100	100
Total	600	600

MAVDESS FEATURE HIGHLIGHTS

Two main promising features of the MAVDESS include:

Scope: The MAVDESS comprises 1200 clips, whereas many sets have less than 100 clips [10]. Only three other collections [11] that we are aware of contain over 1000 recordings of dialogue dynamics. MAVDESS is made up of 100 unprofessional persons who individually execute 600 different vocalization's expressing emotions such as sad, happy, anger, fear, surprise, & disgust. Each person's recorded Scripted and YouTube short is accessible in two formats: audio-video & audio-only. Since imaging research has demonstrated the critical regions of mind become used while seeing repeated stimulus over & over again [12]. Machine learning researchers can also benefit from a huge corpus of recordings. This database will especially suit supervised machine learning techniques, such as emotion classifiers, since it provides a huge amount of data for testing, training and evaluating alternative classifiers.

Intensity of Emotion: On other hand, emotions that were performed at certain emotional intensity level: normal level and intense level. Only two other sets that we are aware of allow controlled intensity modification. Intensity is foremost important characteristics of emotion [13], because it plays key part in numerous emotion theories. Furthermore, if one can notice, the terms activation & intensity can be interchangeably used. Meanwhile, intense facial emotions are also recognized faster than lower intense equivalents although they depict greater imitation reviews from viewers. As a result, when researchers are looking for clear, unambiguous emotional exemplars, strong displays may be beneficial. Normal intensity expressions, on the other hand, may be necessary when researching minor variations in perception of emotion also when researchers are looking for representations that are comparable to those seen in ordinary life.

Designing & validating a new Dataset

Validation and reliability statistics supporting the MAVDESS are presented in the sections below. Each of the 100 participants assessed 1200 files subset for the validation task. The same 100 individuals supplied intra-participant test-retest data for the reliability task. Participants were first asked to identify the expressed emotion in order to validate it. These methods were created to help with nonverbal emotional displays involving relatively motionless faces. Vocal production, on the other hand, necessitates a substantial amount of facial movement, on other hand lexical content with movements associated to emotional expressiveness. As a result, standard muscle coding techniques cannot be used to validate the MAVDESS.

For all stimuli, the validation task provides emotional correctness measures, intensity, and sincerity. These results give a detailed look into the MAVDESS corpus. We offer a composite "SAM" score to aid researchers in selecting relevant stimuli. SAM scores are a weighted sum accuracy & intensity measurements that range from 0 to 9. The equation is set up in such a way that stimuli with higher accuracy, intensity, and authenticity get higher goodness ratings.

Development of the MAVDESS stimuli

People: One hundred nonprofessionals (Mean = 27.0 years; SD = 3.76; Average Range = 21–33; all men) consented to create stimuli. People who spoke Sindhi as their native language and spoke with a neutral accent were qualified. For a variety of reasons, professional actors were not chosen over lay performers because, actors' representations of emotion are less easily recognized than those of ordinary expressers, according to studies [14]. This might be uncertain that same thing holds true for facial expressions or in case of expressions linked to audio-visual, implying that acted emotions are more exaggerated than actual emotions.

Selection of emotions: For the speech, six emotions were chosen: joyful, sad, furious, fear, surprise, and disgust. Emotions that are primary have a lengthy history in philosophy [15], and there are current proponents. Despite criticism, the discrete model of emotion is a viable alternative for creating and labelling emotion sets. As a result, most current sets include these six emotion labels. Some theorists have questioned the classification of surprise as a fundamental emotion, on other side many have advocated for its inclusion as a primary emotion [9]. We have included surprise in the MAVDESS because it is present in many current sets [9].

Procedure and design: The MAVDESS was developed using a process in the same fashion they were recorded while speaking, one dialogue of each emotional situation. The first author and two study assistants watched the audition films and assessed the expressions for correctness, intensity, and authenticity. Emotions were used to block trials, with emotions has lower intensity being followed by their high-intensity equivalents. This arrangement in particular category of emotion has allowed individuals to enter and stay in the intended mood. With everyone, a conversational script was utilized. Each emotional state was described in detail. Emotional labels were included in the description to guarantee that the individual understood what feeling was desired. For each degree of intensity, a story depicting an emotional scenario was presented.

Post Processing & Technical information: Individual recordings were made in a confined room using a Cannon D7200 at 720p, 1920x1080 screen resolution at 30 frames per second and files saved in MP4 format. We positioned the camera 0.3 meters away from the performer. The stimuli

were then visualized on a 15" HP Ultrabook and accessible using MATLAB 2009b and the Open Smile Toolkit [16].

Description of MAVDESS files

Experimental design: There are 1200 recordings in the MAVDESS, with 100 different subjects (all male). Six different vocalizations were created by each participant, comprising of 600 planned spoken utterances and 600 Acted utterances from the internet. Each of the 1200 vocalizations was exported into two different modalities: audio-video (facial and vocal) and only-vocal (no face but audio voice).

Filename convention: A unique filename is assigned to each MAVDESS file. The filename is made up of hyphens that divide the name and numerical identifiers (e.g., Anger 001. mp4). A separate experimental factor's level is defined by each two-digit numerical identification. Actor.mp4 or.wav is a video or audio file. Table 2 shows how levels are numerically coded.

Table 2. MAVDESS filenames Description.

Identifier	Description
Format	1 = Audio-video, 2 = Audio-only
Mode	1 = Scripted Speech, 2 = Acted YouTube Speech
Emotion	1 = Disgust, 2 = Surprise, 3 = Happy, 4 = Angry, 5 = Sad, 6 = Fear.
Emotion Intensity	1 = Normal, 2 = Strong
Dialogues	6! = 36
Person	1 = First Person, . .100 = Hundredth Person

Ethics declaration: Human volunteers were utilized in the MAVDESS and validation experiments. Prior to any experiment or recording, all subjects provided signed informed consent. These people gave their written informed agreement to have their case information published. The database's recording techniques following validation experiment were authorized from Mehran University's council.

Download and Accessibility: The MAVDESS's major objective was to provide a publicly free and accessible verified stimulus set to researchers and other interested parties. The MAVDESS database is currently in processing of improved accuracy benchmarks and will be accessed via reserved DOIT from the open access repository Zenodo (https://doi.org/10.5281/zenodo.5213073) provided under a Creative Commons Attribution-4.0 license for free and without restrictions.

4. Methodology for Baseline Classification Performance

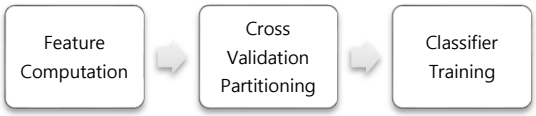


Figure 2. Baseline Process Diagram

When a new dataset is provided for academic study, the standard method in the fields of affecting computing is to offer a baseline classification performance as given in Figure 2. This aids in the dataset's familiarization with the wider research community. As a result, we'll also provide a categorization performance baseline for the Sindhi Speech Emotion Corpus. Our aim is to employ open-access and publicly available technologies (at least for non-commercial research) in order to easily replicate the baseline classification performance.

Step 1: Feature Computation Using Open-Smile Toolkit Feature Sets:

The foremost stage is to create audio attributes that can act relevant audial properties of speech for the job at hand. For this purpose, we've utilized the openSMILE [17] programmed to extract the audial characteristics, later which frequently be used to identify emotion [18]. This research makes use of five openSMILE feature sets (ComParE, eGeMAPS, IS09, IS10, Prosody). The explanation of feature sets utilized in this investigation is given below:

ComParE feature set: Stands for Computational Paralinguistics Challenge (ComParE), referred to as a brute-force because it contains 6,373-dimensional feature sets [19]. Furthermore, ComParE feature set comprises of LLD, MFCC, spectral, & energy.

eGEMAPS feature set: Stands for Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set [20], comprises of 88-dimensional vector based on functionals for various types of voice quality characteristics. Voice quality is determined by voicing probability, jitter, and shimmer information, while pitch and loudness parameters define it. The eGeMAPS also employs harmonic difference characteristics to characterize speech quality in addition to these.

IS09-Emotion feature set: Stands for Inter Speech 2009 (IS09) Emotion feature set [21], creates a 384-dimensional vector from functionals of four distinct types of features: prosodic, voice quality, spectral, and temporal speech characteristics. The IS09-Emotion feature set uses pitch and voicing likelihood as prosody and voice quality

characteristics.

IS10-Paralinguistics feature set: Stands for Inter Speech 2010 (IS10) Paralinguistic feature set [22], using functionals for eight different types of speech prosodic, voice quality, and spectral properties, it generates a 1,582-dimensional vector. Prosody is defined by pitch and loudness, but voice quality is defined by voicing likelihood, jitter, and shimmer.

Prosody Feature Set: Based on a combination of four types of auditory low-level descriptors, the prosody feature set generates a 35-dimensional vector [23]. Pitch and loudness are two prosody qualities, as are two types of voice quality attributes, harmonic to noise ratio (HER) and the chance that a speech segment is voice speech.

Step 2: Cross Validation Model:

Cross-validation is an experimenting technique for testing machine learning models on a small amount of data. The single parameter in the method is k, which defines how many groups a given data sample should be divided into. As a result, the approach is also known as k-fold cross-validation. When a specific value for k is provided, it can be substituted for k in the model's reference, for example, k=4 for 4-fold cross-validation [24].
In other words Cross-validation is a machine learning approach that assesses a machine learning model's competency on unknown data. That is, a small sample size is utilized to see how the model performs in general when it is used to forecast data that was not used during training. [24]

To start examining simply divide the dataset into k (typically 4) folds, with one for testing and k-1 for training and validation. Training + Validation and test data from the same iteration should never overlap, and test data from separate folds should never overlap. The error on the left-out test set is calculated at the end of each cycle.
In basic terms, train on the training set, modify parameters on the validation set, and test on the test set.

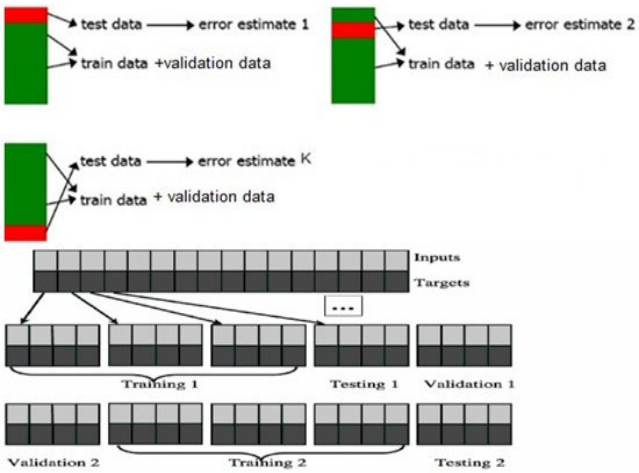


Figure 3. Cross Validation Process [20]

Step 3: CLASSIFIERS (Classification PERFORMANCE Methods):

A classifier for emotion recognition may be trained once all of the audio attributes for all of the recordings in the dataset have been computed. As previously stated, we use cross-validation to ingress the machine learning models preformation prediction. Cross-validation allows you to estimate the execution of machine learning models using data that isn't the same as the data that was used to train them.

We used four distinct classifiers: Logistic Regression (LR), Support Vector Classifiers (SVC) (primarily LSVC and RBFSVC), Decision Tree (DT), and K-Nearest Neighbors (KNN).

1. **Logistic Regression:** Logistic regression (LR) is a technique for estimating distinct values from a set of individualistic variables. Furthermore, it is utilized as a basic categorization that forecasts the likelihood of an event occurring. The LR always produces a categorical variable, such as (1/0, Yes/No).
The LR classifier is implemented using the scikit-learn toolkit. The complication of the logistic regression method is enhanced across a logarithmically spaced grid with values ranging from 10-7 to 107. A l2-penalty is used to train the classifier for up to 10,000 iterations. The data is fitted to Logit functions, which is why LR is also known as Logit regression.

2. **Support Vector Machine:** Support Vector Classification or SVM classify or regress a sequence of patterns and compute performance metrics using a given input model. It categorizes data by presenting each item as a point in n-dimensional space, with the coordinate value of each characteristic.

We have used 2 SVC: LSVC & RBFSV

i) **LSVC:** Similar to the SVC parameter "Linear," but using Lib Linear instead of Lib SVM as the implementation. LSVC has a more customizable penalty and loss function. LSVC is also superior for scalability with huge amounts of data.
ii) **RBFSVC:** RBFSVC stands for Radial Based Function SVC; it performs the same duties as LSVC but is nonlinear. RBFSVC uses Gaussian Kernel or RBF from training samples to conduct classification.

3. **Decision Trees**: The supervised learning method Decision Trees (DT) works with both continuous and categorical dependent variables. The DT divides the dataset into two or more homogeneous groups.

4. **K-Nearest Neighbors:** K-Nearest Neighbors (KNN) is known as regression and classification method. KNN keeps track of all available cases and categorizes new ones based on the votes of its K-Neighbors. Following that, KNN selects the instances of its class that are most prevalent among its nearest neighbors, as determined by the distance function, also known as Hamming Distance.

5. Experimentation Results & Conclusion:

As discussed in the preceding section, audio-video characteristics has been calculated. The dataset is split into two partitions as mentioned in cross validation model, with a ratio of 75:25 [20] as mentioned in Figure 1. The training partition is used to train the classifier, the validation partition is used to tune its hyperparameter, and the classification results are compared to the test partition. We have given the findings in terms of accuracy for both the Scripted and Acted partitions for completeness' sake.

Table 4. Scripted Emotions Accuracy Table

Classifiers	LR	LSVC	RBFSVC	DT	KNN
Feature Set	Acc.	Acc.	Acc.	Acc.	Acc.
ComParE	31.8	32.5	29	24	28
IS09Emotion	28.9	31.6	28	22	23.7
IS10Paraling	30.9	32	26	25	28
Prosody	20.1	21	21	21	20
eGeMAPS	24.2	24	20	20	20.4

Here in Table 4 of Scripted Emotion Table, one can notice that we have reported ComParE feature set to be robust in terms of accuracy followed by IS10 Classifier. While on the other hand Prosody feature set performed poorly. One can also notice that these feature sets has given best accuracy when used with LSVC (Linear Support Vector) and LR (Logistic Regression) while KNN gave poor results.

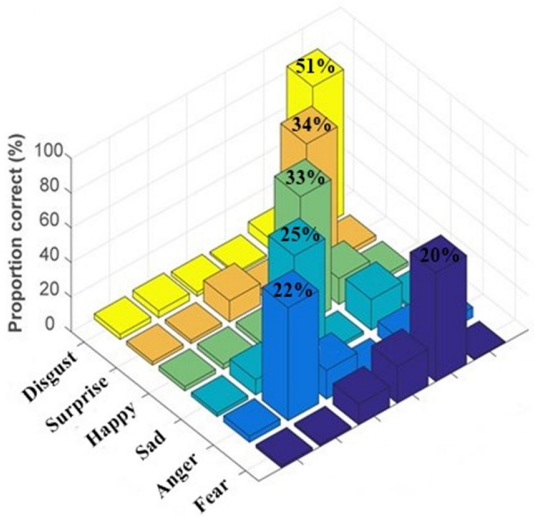


Figure 4. Scripted Emotions Individual Accuracy Chart

Here in Figure 4, we have computed Scripted Individual Accuracy of best classifier which is Logistic Regression, one can notice that the Disgust emotion gave the best accuracy of 51%, followed by emotion Surprise, while the emotions which expressed Fear were barely recognized.

Figure 5. Scripted Emotions Confusion Matrix

Here in Figure 5, Confusion Matrix which helps to visualize algorithm performance and identification between different emotions. Here we have computed confusion matrix of best classifier which is Logistic Regression, one can notice that Disgust emotion has been predicted as true most of the time, followed by emotion Surprise, while the emotions which expressed Sad were barely recognized and usually mislabeled with Fear.

Table 5. YouTube Emotions Accuracy Table

Classifiers	LR	LSVC	RBFSVC	DT	KNN
Feature Set	Acc.	Acc.	Acc.	Acc.	Acc.
ComParE	32.3	31.4	29	24	28
IS09Emotion	22.6	32	28	22	23.7
IS10Paraling	30.4	31	26	25	28
Prosody	18.2	21	21	21	20
eGeMAPS	22.1	23	20	20	20.4

Here in Table 5 of YouTube Emotion Table, one can notice that we have reported ComParE feature set to be robust in terms of accuracy followed by IS09 Classifier. While on the other hand Prosody feature set again performed poorly. One can also notice that these feature sets has given best accuracy when used with LR (Logistic Regression) standalone while DT (decision tree) gave poor results.

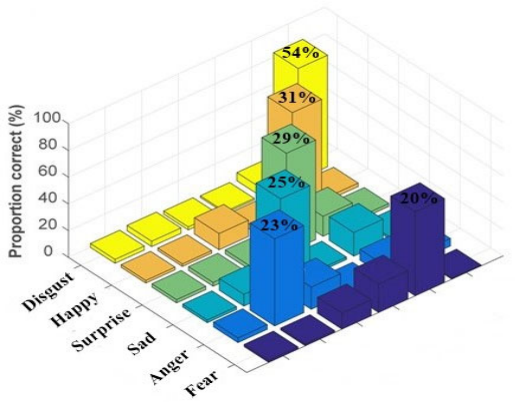


Figure 6. YouTube Emotions Individual Accuracy Chart

Here in Figure 6, we have computed Individual Accuracy of Emotions taken from YouTube, we have only visualized best classifier which is Logistic Regression, one can notice that the Disgust emotion again gave the best increased accuracy of 54%, followed by emotion Happy, while the emotions which expressed Fear again were barely recognized.

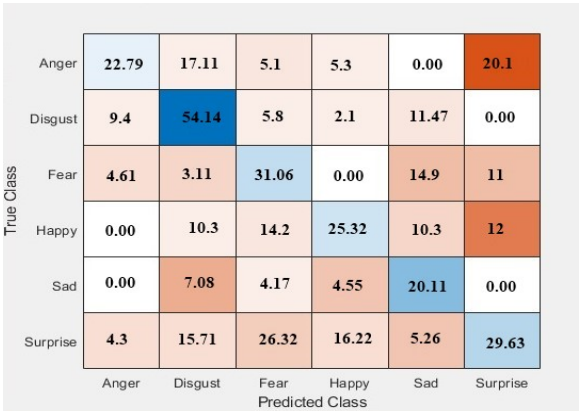


Figure 7. YouTube Emotions Confusion Matrix

Here in Figure 7, Here we have computed confusion matrix of best classifier which is Logistic Regression, one can notice that Disgust emotion has been predicted as true most of the time, followed by emotion Fear, while the emotions which expressed Sad again were barely recognized and usually mislabeled with Anger.

CONCLUSION: In this paper, we offer MAVDESS, a new dataset for training machine learning models for Sindhi language speech emotion recognition. The dataset is now available on the Zenodo platform for academic study. We also ran tests to establish baseline classification performance in terms of Accuracy, utilizing feature sets from the OpenSmile toolbox. We found that LSVC and LR models trained on the ComParE feature set perform the best in terms of classification.

REFERENCES:

[1] de Gelder B, Vroomen J. The perception of emotions by ear and by eye. Cognition & Emotion. 2000; 14(3):289–311.

[2] Dolan RJ, Morris JS, de Gelder B. Crossmodal binding of fear in voice and face. Proceedings of the National Academy of Sciences. 2001; 98(17):10006–10

[3] Kreifelts B, Ethofer T, Grodd W, Erb M, Wildgruber D. Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. NeuroImage. 2007; 37(4):1445–56.

[4] Collignon O, Girard S, Gosselin F, Roy S, Saint-Amour D, Lassonde M, et al. Audio-visual integration of emotion expression. Brain Research. 2008; 1242:126–35.

[5] Sato W, Yoshikawa S. Spontaneous facial mimicry in response to dynamic facial expressions. Cogni- tion. 2007; 104(1):1–18.

[6] Weyers P, Mu¨ hlberger A, Hefele C, Pauli P. Electromyographic responses to static and dynamic ava- tar emotional facial expressions. Psychophysiology. 2006; 43(5):450–3.

[7] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "IITKGP-SEHSC: Hindi speech corpus for emotion analysis," in International Conference on Devices and Communications, 2011, pp. 1–5.

[8] S. A. Ali, S. Zehra, M. Khan, and F. Wahab, "Development and Analysis of Speech Emotion Corpus Using Prosodic Features for Cross Linguistics," International Journal of Scientific & Engineering Research, vol. 4, no. 1, 2013.

[9] Ekman P. An argument for basic emotions. Cognition and Emotion. 1992; 6(3–4):169–200.

[10] Belin P, Fillion-Bilodeau S, Gosselin F. The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods. 2008; 40 (2):531–9

[11] Banziger T, Grandjean D, Scherer KR. Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). Emotion. 2009; 9(5):691.

[12] Breiter HC, Etcoff NL, Whalen PJ, Kennedy WA, Rauch SL, Buckner RL, et al. Response and habitua- tion of the human amygdala during visual processing of facial expression. Neuron. 1996; 17(5):875– 87.

[13] Sonnemans J, Frijda NH. The structure of subjective emotional intensity. Cognition & Emotion. 1994; 8(4):329–50

[14] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B, editors. A database of German emo- tional speech. Ninth European Conference on Speech

Communication and Technology (INTER-SPEECH 2005); 2005; Lisbon, Portugal.

[15] Descartes R. The passions of the soul. In: Cottingham J, Stoothoff R, Murdoch D, editors. The philo- sophical works of Descartes. Cambridge: Cambridge University Press (Original work published 1649); 1984.

[16] Brainard DH. The psychophysics toolbox. Spatial Vision. 1997; 10:433–6.

[17] Florian Eyben, Martin W¨ollmer, and Bj¨orn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010, pp. 1459–1462.

[18] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in Proceedings of the 16th International Conference on Multimodal Interaction. ACM, 2014, pp. 494–501.

[19] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: De- ception, Sincerity & Native language," in INTERSPEECH, 2016, pp. 2001–2005.

[20] Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[21] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *INTERSPEECH*, 2009, pp. 312–315.

[22] F S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, S. Lan- guage, P. Group, D. Telekom, and A. G. Laboratories, "The INTER-SPEECH 2010 Paralinguistic Challenge," in INTERSPEECH, 2010, pp. 2794–2797.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in ACM international conference on Multimedia, 2013, pp. 835–838.

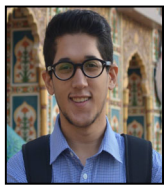
[24] Stavros Petridis, "Imperial College London, Machine Learning Course 395, 2018" https://ibug.doc.ic.ac.uk/media/uploads/documents/ml-lecture3-2018.pdf

Tanveer Ahmed works as a Research Assistant in the Department of Telecommunication Engineering at Mehran University, Jamshoro, Pakistan. He completed his Bachelors and Masters' degree in Communication Engineering from Mehran University, Jamshoro Pakistan. He has also worked previously as Project Implementation Manager at Zong CMPAK and RF Engineer at Huawei Technologies Pakistan.

Sajjad Ali Memon works as an Associate Professor in Department of Telecommunication Engineering at Mehran University, Jamshoro, Pakistan. He completed his Bachelors & Master's degree in Communication Engineering form Mehran University of Engineering and Technology, Jamshoro, Pakistan. He has done his Ph.D. from Dalian University of Technology, Dalian, China.

Saqib Hussain works as a Teaching Assistant in the Department of Telecommunication Engineering at Mehran University, Jamshoro, Pakistan. He completed his Bachelors and Masters' degree in Communication Engineering from Mehran University, Jamshoro Pakistan.

Amer Tanwani He is currently working as Software Engineer, he is founder of startup MapIn app that is used for indoor navigation. He completed his Bachelor's degree in Computer Engineering from Ghulam Ishaq Khan Institute of Engineering Science & Technology, KPK, Pakistan.



Ahmed Sadat He is currently working as IT Support Engineer. He completed his Bachelor's degree in Communication Engineering from Mehran University, Jamshoro, Pakistan, and Masters' Degree from Monash University, Melbourne, Australia.