Introducing the Urdu-Sindhi Speech Emotion Corpus: A Novel Dataset of Speech Recordings for Emotion Recognition for Two Low-Resource Languages

Zafi Sherhan Syed[1], Sajjad Ali Memon[2]
Mehran University
Pakistan

Muhammad Shehram Shah[3]
RMIT University
Australia

Abbas Shah Syed[4]
University of Louisville
USA

Abstract—**Speech emotion recognition is one of the most active areas of research in the field of affective computing and social signal processing. However, most research is directed towards a select group of languages such as English, German, and French. This is mainly due to a lack of available datasets in other languages. Such languages are called *low-resource languages* given that there is a scarcity of publicly available datasets. In the recent past, there has been a concerted effort within the research community to create and introduce datasets for emotion recognition for low-resource languages. To this end, we introduce in this paper the *Urdu-Sindhi Speech Emotion Corpus*, a novel dataset consisting of 1,435 speech recordings for two widely spoken languages of South Asia, that is Urdu and Sindhi. Furthermore, we also trained machine learning models to establish a baseline for classification performance, with accuracy being measured in terms of unweighted average recall (UAR). We report that the best performing model for Urdu language achieves a UAR = 65.00% on the validation partition and a UAR = 56.96% on the test partition. Meanwhile, the model for Sindhi language achieved UARs of 66.50% and 55.29% on the validation and test partitions, respectively. This classification performance is considerably better than the chance level UAR of 16.67%. The dataset can be accessed via https://zenodo.org/record/3685274**

Keywords*—*Speech emotion recognition; affective computing; social signal processing

I. Introduction

According to the Oxford dictionary [1], the word emotion is defined as *a strong feeling such as love, fear, or anger; the part of a person's character that consists of feelings*. However, in research literature from the field of psychology, one finds that there is no consensus on a definition of emotion. According to [1] an emotion is *any mental experience with high intensity and high hedonic content (pleasure/displeasure)*. Meanwhile, [2] defines emotion as *a complex psychological event that involves a mixture of reactions: 1) a physiological response, 2) an expressive reaction (distinctive facial expression, body posture, or vocalization), and 3) some kind of subjective experience (internal thoughts and feelings)*.

Expression of feelings and by extension emotions is a fundamental part of human behavior. Emotions play an important role in how one thinks and behaves which means that analysis of emotions exhibited by individuals can be used to gain insights into their thought process.

In the age of artificial intelligence, there has been a growing desire amongst the research community to enable interaction between machines (say, robots) and human beings on a more natural level. This is possible when machines can understand, interpret, and recognize human emotion. To achieve this, researchers from the field of affective computing and social signal processing have explored the development of computational methods for emotion recognition from various modalities such as speech [3], [4], facial expressions [5], [6], text [7], [8], and physiological signals [9], [10].

Amongst these modalities, speech is particularly interesting since it is the most natural way for human beings to exhibit emotions [3]. In addition to providing social intelligence to machines, speech emotion recognition can be used to assist emergency services and healthcare professionals. For example, an emotion recognition system linked with emergency services call centers can be useful to gauge the intensity of distress of the caller and subsequently assign their call to a higher priority.

While a great deal of research literature is available on emotion recognition, an overwhelming majority of it caters to western European languages such as English, German, and French – this is mainly because most datasets available are in these languages. Based on our literature survey, we find that there is a particular scarcity of datasets from the South Asian family of languages, even though the region is home to more than 1.891 billion people [2].

We note that recently there have been efforts by several researchers to design and create datasets for speech emotion recognition for South Asian languages. Koolagudi et al. [11] had published a large dataset for speech emotion recognition for Telugu language, a language predominantly spoken in Southern India. The dataset consists of 12,000 utterances in total for eight types of emotions including anger, disgust, fear, happiness, neutral, sadness, sarcasm, and surprise. In [12], Syed et al. introduced the Emotion-Pak Corpus, which included four emotions which include sadness, comfort, anger, and happiness in five languages spoken in Pakistan. These

[1] https://www.oxfordlearnersdictionaries.com/definition/english/emotion

[2] https://en.wikipedia.org/wiki/South_Asia

languages include Urdu, Sindhi, Balochi, Punjabi, and Pashto. The dataset was recorded using ten native speakers for the five languages. While this dataset is most relevant to our work, we could not get a reply from Syed et al. after requesting access to the Emotion-Pak Corpus. Finally, Latif et al. [13] introduced an emotion corpus for Urdu language. The dataset consists of 400 audio recordings for four emotions that were collected from television programs. The dataset is available for academic research on speech emotion recognition [3].

In this paper, we introduce a novel speech emotion dataset consisting of 1,435 audio recordings which can be used to train machine learning models for speech-based emotion recognition in two South Asian languages, namely Urdu and Sindhi. Urdu [4] is the national language as well as the *lingua franca* of Pakistan and is also widely spoken in India. There are upwards of 68.62 million native speakers of Urdu and more than 101.58 million individuals speak Urdu as a secondary language. Meanwhile, Sindhi [5] has more than 25 million native speakers in South Asia, mostly centered in the Sindh province of Pakistan. It is one of the three official languages of the Sindh province in addition to being one of the recognized languages of India.

The rest of the paper is organized as follows: In section II we introduce the methodology for collection of Urdu-Sindhi Speech Emotion Corpus whereas in section III we detail the methodology for establishing the baseline classification performance for the dataset. Experimental results and discussion is provided in section IV, and conclusion in provided in section V.

II. Dataset Collection

In this section we shall introduce the data collection methodology for the Urdu-Sindhi Speech Emotion Corpus with the aid of Fig. 1 which illustrates data collection framework. We prepared 10 sentence scripts each for seven types of emotional utterances in Urdu and Sindhi languages. These emotions include anger, disgust, happiness, neutral, sarcasm, sadness, and surprise. The scripts were validated by the authors of this paper as well as two post-graduate students before being passed down to volunteer participants.

Participations for this study were recruited from amongst undergraduate students currently studying in the Department of Telecommunication Engineering at Mehran University, Pakistan. These participants were instructed to recording themselves uttering the scripts with the predefined emotions and send audio recordings to the authors via WhatsApp [6]. We specifically chose to utilize a WhatsApp based data collection instead of a bespoke recording studio/room since the former enables us to recruit a larger number of participants, including those who may not be able to come to the recording studio.

The audio recordings sent by participants were collected via Twilio [7], an API that provides connectivity with WhatsApp and a desktop computer. Through this process, we were able to collect 734 speech recordings for Urdu language and 701 recordings for Sindhi language. A summary of the number of recordings for each emotion is provided in Table I Each of these recordings was manually checked to ensure that their content was as desired for this study. Readers who are interested in the dataset can access it via https://zenodo.org/record/3685274.

TABLE I. Summary of the number of examples per class for the Urdu-Sindhi Speech Emotion Corpus

Emotion	Urdu	Sindhi
Anger	115	102
Disgust	111	87
Happiness	94	103
Neutral	70	98
Sadness	114	96
Sarcasm	114	118
Surprise	116	97
Σ	734	701

III. Methodology for Baseline Classification Performance

It is common practice in the field of affecting computing and social signal processing to provide a baseline classification performance for every novel dataset when it is introduced for academic research. This helps the larger research community getting familiarized with the dataset. Therefore, we shall provide a baseline classification performance for the Urdu-Sindhi Speech Emotion Corpus as well. Our motivation is to use open-source and freely available tools (at least for non-commercial research) so that the baseline classification performance can be reproduced with relative ease.

A generic process flow diagram for speech emotion classification is illustrated in fig. 2. The first step is to compute audio features which can represent acoustic characteristics of speech which are relevant for the task at hand. For this purpose, we use five types of feature sets from the OpenSmile toolkit [14], [15] which include the Prosody feature set, the IS09-Emotion feature set, the IS10-Paralinguistics feature set, the ComParE feature set, and the eGeMAPS feature set. As the reader shall see, these feature sets have proven to be useful for quantifying paralinguistic characteristics of speech such as prosody, voice quality, speech spectra etc. In subsequent paragraphs, we shall briefly describe these feature sets.

Prosody feature set: The Prosody feature set produces a 35-dimensional vector based on functionals of four types of acoustic low-level descriptors. These include two prosody features, which include pitch and loudness, and two types of voice quality features, that is harmonic to noise ratio (HER) and the probability with which a speech segment contains voice speech (voicing probability). We refer the reader to [15], [14] for further details about the prosody feature set.

IS09-Emotion feature set: The OpenSmile IS09-Emotion feature set produces a 384-dimensional vector based on functionals of four types of features with one each to describe the prosodic, voice quality, spectral, and temporal characteristics of speech. Similar to the Prosody feature set discussed earlier, the IS09-Emotion feature set uses pitch and voicing probability as prosody and voice quality features, respectively. In addition to these, Mel Frequency Cepstral Coefficients (MFCC) features are used to describe the spectral characteristics of voice,

[3] https://github.com/siddiquelatif/URDU-Dataset

[4] https://en.wikipedia.org/wiki/Urdu

[5] https://en.wikipedia.org/wiki/Sindhi_language

[6] https://www.whatsapp.com

[7] https://www.twilio.com/

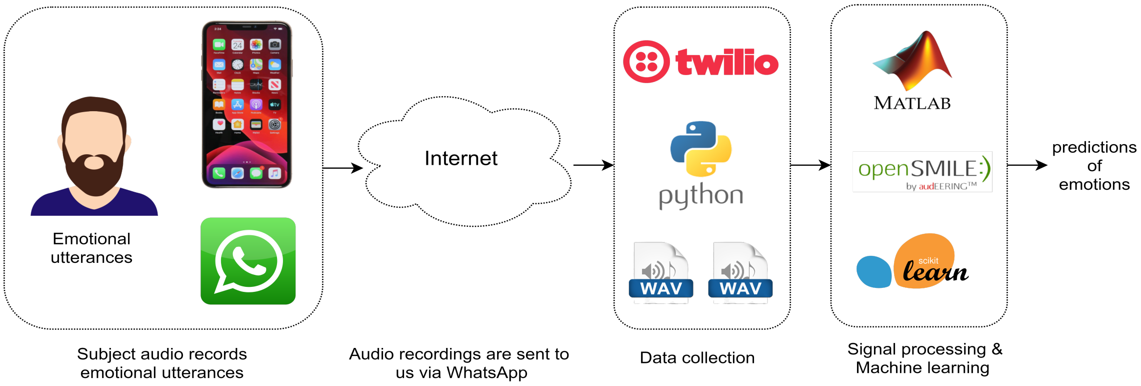

Fig. 1. Illustration of data collection framework

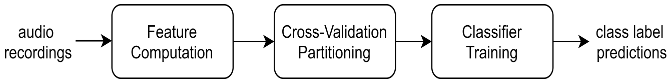

Fig. 2. Illustration of the pipeline for baseline classification

whereas the zero crossing rate of the voice signal is used to describe its temporal characteristics. The IS09-Emotion feature set was introduced for the year 2009 edition of the Interspeech Computational Paralinguistics Challenge [16] and the feature set was shown to be useful for the task of emotion recognition from speech. We refer the reader to [15], [16], [14] for further details about this feature set

IS10-Paralinguistics feature set: The IS10-Paralinguistic feature set produces a 1,582-dimensional vector based on functionals for eight types of features which describe the prosodic, voice quality, and spectral characteristics of speech. Prosody is characterized using pitch and loudness features, whereas voice quality is characterized using voicing probability, jitter, and shimmer features. Spectral characteristics of voice are described using MFCCs, spectral bands filtered by log-Mel filters, and the line spectral pairs of frequencies features which represent linear prediction coefficients. The IS10-Paralinguistic feature set was introduced for the year 2010 edition of the Interspeech Computational Paralinguistics Challenge [17] and these features were shown to be useful for a variety of classification tasks related to speech paralinguistics.

ComParE feature set: The Computational Paralinguistics Challenge (ComParE) is a 6,373-dimensional feature set which was introduced for the year 2016 edition of the Interspeech Computational Paralinguistics Challenges [18]. The ComParE feature set is often referred to as a brute-force feature set since it includes features which describe a wide range of acoustic characteristics. It has been shown to work well for a variety of tasks related to speech paralinguistics and has been used to establish strong baselines for classification and regression tasks for Interspeech Computational Paralinguistics Challenges [18], [19], [20], [21]. We refer the reader to [18], [14] for further details about this feature set.

eGeMAPS feature set: The Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set was designed by some of the leading researchers in the field of social signal processing in order to facilitate a common framework for research into speech paralinguistics. It was also intended to serve a more efficient and lower dimensional feature set than the ComParE feature set. The eGeMAPS feature set produces an 88-dimensional vector based on functionals for various types of prosody, voice quality, and spectral features. Similar to IS10-Paralinguistics feature set, prosody is characterized through pitch and loudness features, and voice quality is characterized by voicing probability, jitter, and shimmer features. In addition to these, the eGeMAPS also uses harmonic difference features to describe voice quality. These include H1-H2 and H1-H3, which quantify differences in the amplitude of second and third harmonics with respect to the amplitude of the first harmonic. The eGeMAPS feature set uses eight types of features to describe the spectral characteristics of speech. Spectral features used in eGeMAPS include alpha ratio, the Hammarberg index, spectral slopes, spectral flux, formant frequencies, relative energies for each formant frequency with respect to the first formant, and the bandwidth for the first formant frequency. We refer the reader to [22], [14] for further details about the eGeMAPS feature set.

Once audio features have been computed for all audio recordings in the dataset, a classifier can be trained for emotion recognition. We choose the logistic regression classifier for this purpose although any other classification algorithm could have also been used. We make use of cross-validation in order to assess the predictive performance of these machine learning models. Cross-validation makes it possible to infer the performance of machine learning models outside of the samples which were used to train those models.

IV. Experimentation, Results and Discussion

We use the implementation of logistic regression classifier which is available in the scikit-learn toolkit [8]. The complexity value of the logistic regression algorithm is optimized over a logarithmically spaced grid between 10^{-7} to 10^{7}. The classifier is trained with an $l2$-penalty for up to 10,000 iterations.

Audio features are computed as per the discussion in the previous section. The dataset is divided into three partitions, that is training, validation, and test with a 60:20:20 ratio. The classifier is trained using the training partition, its hyperparameter is optimized using the validation partition, and the

[8] https://scikit-learn.org

classification results being compared against the test partition. For the sake of completeness, we report the results for both validation and test partitions.

A. Classification Performance for Urdu Language

In table II, results for the classification performance of five audio feature sets is summarized for Urdu language. Here, one can note that for the validation partition, the ComParE feature set provides the highest UAR i.e. 65.49%, which is a considerably strong performance given that chance level UAR is only 14.28%. Amongst other features, one finds that the IS10-Paralinguistics feature set provides the second-best performance, achieving a UAR of 59.46%. Interestingly, the IS09-Emotion and eGeMAPS feature sets which were explicitly designed for tasks related to emotion recognition do not yield good classification results as compared to ComParE or IS10-Paralinguistics feature sets. On the test partition, the ComParE feature set achieves a UAR = 56.96% whereas the IS10-Paralinguistics achieves a UAR = 59.40%.

TABLE II. Summary of Classification Performance of the Five OpenSmile Features for *Urdu* Language

Feature Set	**Comp.**	**Validation**		**Test**	
		UAR	***Acc.***	***UAR***	***Acc***
Prosody	10^{2}	32.61%	32.65%	25.24%	26.53%
IS09Emotion	10^{-1}	42.38%	41.50%	46.72%	46.94%
IS10Paraling	10^{0}	59.46%	59.86%	59.40%	59.86%
ComParE	10^{2}	65.49%	65.31%	56.96%	57.14%
eGeMAPS	10^{0}	38.65%	38.78%	33.89%	35.37%

In fig. 3, the confusion matrix of the best performing model (based on ComParE features) for speech emotion recognition in Urdu language has been shown. Here, one can note that the class with the most accurate prediction of its labels is *Surprise*, which is followed by *Sadness* and *Neutral*. Meanwhile, it is apparent that the classifier had most difficulty in classifying *Disgust* emotion, often mistaking it for *Happiness* and *Sadness* emotions.

B. Classification Performance for Sindhi Language

In table III, the results for classification performance of speech emotion recognition for Sindhi language is summarized. Here, one can note that the ComParE feature set again provides the best classification performance on the validation partition. It achieves a UAR = 66.54%, which is comparable to the UAR achieved by the same feature set for Urdu language. Similarly, we find that the IS10-Paralinguistics feature set achieves the second-best performance with a UAR = 62.17%. On the test partition, these features achieve a UAR = 55.29% and UAR = 46.82%, respectively.

The confusion matrix for the best preforming model (based on ComParE features) for Sindhi language is shown in fig. 4. Here, one can note that the classifier performs best for *Happiness*. It performs worst for the *Neutral* class, often mistaking it for emotions of *Anger*, *Sadness*, and *Sarcasm*.

Overall, we report that the ComParE feature set is suitable for emotion recognition in the two South Asian languages considered, that is Urdu and Sindhi. We hypothesize that this

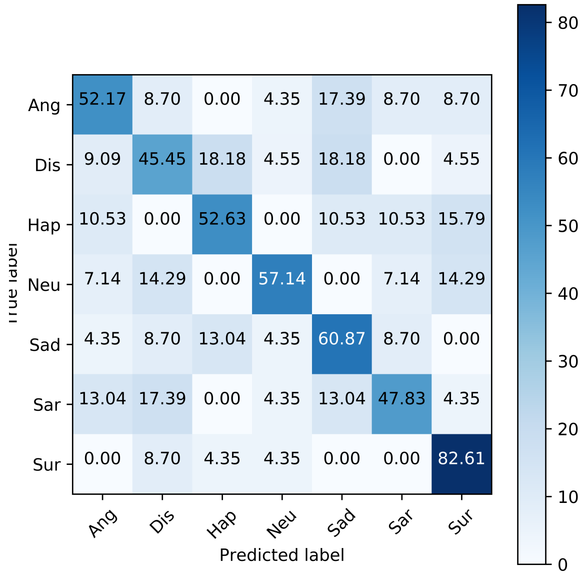

Fig. 3. Confusion matrix for the best performing model (based on ComParE features) for Urdu language

TABLE III. Summary of Classification Performance of the Five OpenSmile Features for *Sindhi* Language

Feature Set	**Comp.**	**Validation**		**Test**	
		UAR	***Acc.***	***UAR***	***Acc***
Prosody	10^{1}	32.76%	33.57%	31.22%	31.21%
IS09Emotion	10^{1}	55.51%	55.00%	43.22%	43.26%
IS10Paraling	10^{1}	62.17%	62.14%	46.82%	46.81%
ComParE	10^{-2}	66.54%	66.43%	55.29%	56.03%
eGeMAPS	10^{7}	48.41%	47.14%	32.89%	32.62%

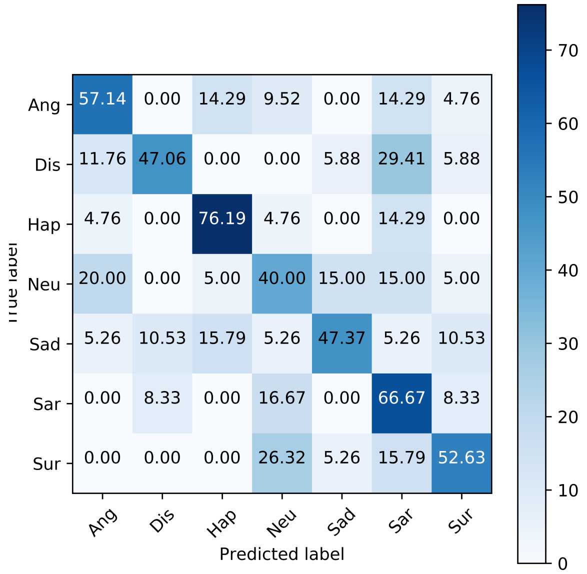

Fig. 4. Confusion matrix for the best performing model (based on ComParE features) for Sindhi language

is due to the *brute force* nature of the ComParE feature set as it includes a large number of features (i.e. 6,373 in total!) which can capture various characteristics of speech.

C. Cross-language Classification Performance

Finally, we seek to quantify how well machine learning models perform when they are optimized for speech emotion recognition in one language, say Urdu, and are tested for the other language, say Sindhi, and vice versa. One would assume that given the two languages are widely spoken in the same region, emotional intonation between the two languages may be similar and as a result, some degree of transferability between models may exist.

To this end, we summarize in table IV the results of cross-language classification performance of the top-two performing feature sets, that is IS10-Paralinguistics and the ComParE feature set. Contrary to our surmisal, one finds that there is little transferability of information between the two languages. When the logistic regression model is trained on Urdu language, the highest UAR it achieves on the test partition of the Sindhi language is 19.15% which is rather poor. Similarly, a model trained on Sindhi language only achieves a maximum UAR of 17.69% on the test partition of Urdu language.

We believe that the results in table IV are particularly interesting because they show that the transferability of machine learning models for emotion recognition does not always hold even when the two languages belong to the same language group and are spoken in the same region. However, one can argue that the more powerful machine learning models, such as those based on deep learning [23] are likely to perform better than logistic regression.

TABLE IV. SUMMARY OF CROSS-LANGUAGE CLASSIFICATION PERFORMANCE

Trained on Urdu				
Feature Set	**Test (Urdu)**		**Test (Sindhi)**	
	UAR	***Acc***	***UAR***	***Acc***
IS10Paraling	59.40%	59.86%	16.40%	16.31
ComParE	56.96%	57.14%	19.39%	19.15
Trained on Sindhi				
Feature Set	**Test (Sindhi)**		**Test (Urdu)**	
	UAR	***Acc***	***UAR***	***Acc***
IS10Paraling	43.22%	43.26%	17.10%	17.01
ComParE	46.82%	46.81%	17.32%	17.69

V. CONCLUSION

In this paper, we introduced a novel dataset, called the Urdu-Sindhi Speech Emotion Corpus, which can be used to train machine learning models for speech emotion recognition for two low-resource languages. We have made the dataset available for academic research on the Zenodo platform. Furthermore, we also conducted experiments to establish baseline classification performance in terms of UAR using feature sets from the OpenSmile toolkit – a toolkit used by researchers in the field to set empirical baselines for classification performance. Based on our experiments, we reported that logistic regression models trained on the ComParE feature set are the best performing in terms of classification performance for speech emotion recognition for both Urdu and Sindhi languages.

ACKNOWLEDGMENT

The authors would like to thank Noor-ul-ain Suhail and Tanveer Memon for their help in data collection and organization.

REFERENCES

[1] M. Cabanac, "What is emotion?" *Behavioural processes*, vol. 60, no. 2, pp. 69–83, 2002.

[2] J. S. Nairne, *Psychology*, 6th ed., 2013.

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

[4] M. B. Akcay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[5] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Computer Speech & Language*, vol. 1, no. 1, pp. 1–15, 2018.

[6] N. Samadiani, G. Huang, B. Cai, W. Luo, C.-H. Chi, Y. Xiang, and J. He, "A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data," *MDPI Sensors*, vol. 19, no. 8, pp. 1–27, 2019.

[7] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: A survey," *Social Network Analysis and Mining*, vol. 28, pp. 1–8, 2018.

[8] E. Kim and R. Klinger, "A Survey on Sentiment and Emotion Analysis for Computational Literary Studies," *arXiv:1808.03137*, pp. 1–26, 2018.

[9] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *MDPI Sensors*, vol. 18, no. 7, pp. 1–41, 2018.

[10] M. Egger, M. Ley, and S. Hanke, "Emotion Recognition from Physiological Signal Analysis: A Review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[11] 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.

[12] 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.

[13] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages," in *International Conference on Frontiers of Information Technology*, 2018, pp. 88–93.

[14] AudEERING, "OpenSMILE - audEERING," 2013. [Online]. Available: https://www.audeering.com/opensmile/

[15] 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.

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

[17] S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, S. Language, P. Group, D. Telekom, and A. G. Laboratories, "The INTERSPEECH 2010 Paralinguistic Challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

[18] 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: Deception, Sincerity & Native language," in *INTERSPEECH*, 2016, pp. 2001–2005.

[19] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold and Snoring," in *INTERSPEECH*, 2017, pp. 1–5.

[20] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats," in *INTERSPEECH*, 2018, pp. 1–5.

[21] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Noth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *INTERSPEECH*, 2019, pp. 1–5.

[22] F. 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.

[23] Y. A. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.