Speech Emotion Recognition for the Urdu Language

Dataset and Evaluation

Nimra Zaheer[1*], Obaid Ullah Ahmad[1], Mudassir Shabbir[2] and Agha Ali Raza[3]

[1]Computer Science Department, Information Technology University, Lahore, Pakistan.
[2]Computer Science Department, Vanderbilt University, Nashville, TN, USA.
[3]Computer Science Department, Lahore University of Management Sciences, Lahore, Pakistan.

*Corresponding author(s). E-mail(s): phdcs17001@itu.edu.pk;
Contributing authors: obaidullah.ahmad@itu.edu.pk;
mudassir.shabbir@vanderbilt.edu; agha.ali.raza@lums.edu.pk;

Abstract

Crafting reliable Speech Emotion Recognition systems is an arduous task that inevitably requires large amounts of data for training purposes. Such voluminous datasets are currently obtainable in only a few languages, including English, German, and Italian. In this work, we present **SEMOUR**$^+$: a *S*cripted *EMO*tional Speech Repository for *Ur*du, the first scripted database of emotion-tagged and diverse-accent speech in the Urdu language, to design an Urdu Speech Emotion Recognition system. Our gender-balanced **14**-hours repository contains **27, 640** unique instances recorded by **24** native speakers eliciting a syntactically complex script. The dataset is phonetically balanced, and reliably exhibits varied emotions, as marked by the high agreement scores among human raters in experiments. We also provide various baseline speech emotion prediction scores on SEMOUR$^+$, which could be utilized for multiple applications like personalized robot assistants, diagnosis of psychological disorders, getting feedback from a low-tech-enabled population,
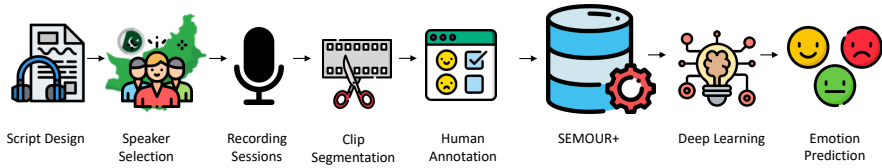
etc. In a speaker-independent experimental setting, our ensemble model accurately predicts an emotion with a state-of-the-art **56%** accuracy.

1 Introduction

Humans are complex beings with an intrinsic ability to express themselves via emotions. Apart from facial expressions and gestures, the tone of verbal communication encapsulates one's sentiments. The field of speech emotion recognition has been sufficiently improved to understand such responses uttered by human beings with the availability of extensive datasets and massive processing power. Such systems can be utilized by robots to educate children (J. Chen et al., 2019), enhance customer's satisfaction in call centers (Li, Dimitriadis, & Stolcke, 2019), monitoring and aiding the elderly (Castillo, Fernández-Caballero, Castro-González, Salichs, & López, 2014a) and introducing smart homes systems to recognize catastrophes and distress (Castillo, Fernández-Caballero, Castro-González, Salichs, & López, 2014b; Kostoulas, Mporas, Ganchev, & Fakotakis, 2008; Scott, Ashby, & Hanna, 2020). Other applications of speech emotion recognition (SER) systems in Human-Computer Interaction (HCI) include enhancing teaching experience in online learning, detecting lies, in-car emotion recognition, and providing aid for psychologists (Ramakrishnan & El Emary, 2013). Other applications include collecting feedback (Han, Jiang, Li, Schuller, & Ruan, 2020) and identifying bias in conversations (Vashistha, Garg, Anderson, & Raza, 2019).

The availability of extensive repositories rich in emotion is a precondition for constructing language-dependent emotion recognizers. Such resources are available for only a handful of languages such as English, German and Italian. Urdu is the national language of Pakistan with 171 million speakers worldwide (Eberhard, Simons, & Fennig, 2020). When paired with its close variant, Hindi, Urdu is the third most widely-spoken language globally. Despite such widespread use, Urdu is scarce in computational resources. Most of the applications above are relevant for the native speakers of the Urdu language in South Asia. Therefore, a reliable speech emotion recognition system for the Urdu language will benefit many people. In this article, we study this research problem and present an extensive speech database: SEMOUR$^+$. To the best of our knowledge, we are the first to develop such a repository. We ensure that our dataset approximates the common tongue in terms of the distribution of phonemes and accents so that models trained on our dataset will be easily generalizable. We collected an extensive resource of more than $27,000$ utterances elicited in eight emotions by 24 speakers with diverse accents. It follows a phonetically balanced script. The utterances are recorded in

Script Design | Speaker Selection | Recording Sessions | Clip Segmentation | Human Annotation | SEMOUR+ | Deep Learning | Emotion Prediction

Figure 1: The figure elaborates the dataset collection process for designing SEMOUR$^+$ along with applying deep learning models for predicting emotions

a sound-proof radio studio by native speakers. Therefore, high-quality audio clips are shared with the research community. We also explore various machine and deep learning models for speech emotion recognition and report an excellent speaker-dependent and independent accuracy of emotion prediction. Figure 1 elaborates the whole process for the acquisition and application of our repository. In summary, the paper has the following contributions:

1. We study the speech emotion recognition problem for the Urdu language and build the first 14-hours dataset: SEMOUR$^+$ that contains $27,640$ high-quality sound instances tagged with eight different emotions and uttered in diverse accents by 24 native speakers belonging to various provinces of Pakistan.
2. We report the human accuracy of detecting emotion in this dataset and other statistics collected during an experiment of 16 human subjects annotating about $5,000$ utterances with an emotion.
3. We train a basic machine learning model on SEMOUR$^+$ on the the first eight actors to recognize emotion in the spoken Urdu language. We report an excellent cross-validation speaker- independent accuracy of 92% which compares favorably with the state-of-the-art.
4. We improve the aforementioned simple speech emotion recognizer for the Urdu language in the speaker-independent setting by extending the dataset in terms of accents and utilizing various deep learning algorithms. We achieve the best result by applying the Ensemble model consisting of CNN and VGG-19 to get a 17% improvement compared to a simple neural network when trained on 24 speakers and prove that diverse accents extension is a valuable addition.

The following section provides a detailed literature overview of existing speech emotional repositories for different languages and Speech Emotion Recognition (SER) systems. In Section 3 , we give details on the design and recording of our dataset. The results of the human evaluation are discussed in Section 4. Section 5 provides details of machine and deep learning frameworks to predict emotions using this dataset for training, followed by a detailed discussion in Section 6. Finally, we conclude our research in Section 7.

Table 1: Summary of available Urdu datasets.

Dataset	No. of Emotions	Size (hr)	#sp.
Raza et al. (2009)	-	3	1
Sarfraz et al. (2010)	-	45	82
Ali et al. (2012)	-	—	50
RATS (Walker et al., 2015)	-	3,000	-
Baang Dataset (Raza et al., 2018)	-	1,207	4,678
Latif et al. (2018)	4	< 1	50
SEMOUR$^+$ (this work)	8	14	24

2 Related Work

In recent years, extensive research has been conducted to produce quality resources for various languages to aid emotion recognition from speech signals. In this section, we discuss numerous datasets available for the Urdu language. Further, we provide details on resources available for other languages. Lastly, we elaborate on the research conducted on speech emotion recognition utilizing various machine and deep learning models.

2.1 Resources available for the Urdu language

Urdu is a resource-scarce language with only one emotionally charged database of spontaneous speech extracted from talk shows. (Latif, Qayyum, Usman, & Qadir, 2018). It is available with 400 instances uttered through four basic human emotions: happiness, sadness, neutral, and anger. Two human evaluators annotate these utterances. There is no sentence-level segmentation, and some instances are void of dialogues. Moreover, such a small repository is unsuitable for data-hungry machine learning algorithms.

Other repositories that exist for the Urdu language are void of emotional speech and can be helpful for speech automation tasks such as text-to-speech systems and automatic speech recognition (Ali, Ahmad, Yahya, & Farooq, 2012; Raza, Hussain, Sarfraz, Ullah, & Sarfraz, 2009; Sarfraz et al., 2010). Huge speech repositories containing spontaneous conversations are useful for speaker identification (Raza et al., 2018) and speech activity detection (Walker et al., 2015). The datasets above will not suit a speaker-independent speech emotion recognition system as the speech has no emotional tags. A considerable dataset diverse in accents is required for recognizing emotions in a speaker-independent system. Table 1 provides a summarized overview of existing Urdu language repositories.

2.2 Resources available for other languages

IEMOCAP is the benchmark database used for English speech emotion recognition systems (Busso et al., 2008). It contains 10,039 dialogues between speakers in 9 emotions rated by three annotators. Ten actors perform this

12-hour repository of scripted and improvised sessions. Another dataset exists for the English language named SAVEE, which is acted by four actors in seven emotions (Jackson & Haq, 2014). MSP-IMPROV contains improvised sessions performed by 12 actors in five emotions eliciting 20 sentences (Busso et al., 2016). RAVDESS is another emotionally charged dataset for the English language performed by 24 actors by eliciting two neutral sentences in seven emotions (Livingstone & Russo, 2018). A recent repository named VESUS has been introduced following a lexically diverse script of 252 instances voiced by ten actors (Sager, Shankar, Reinhold, & Venkataraman, 2019).

For German language, EmoDB contains 10 sentences uttered by 10 speakers in seven emotions (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). Other German-language repositories include VAM (Grimm, Kroschel, & Narayanan, 2008) and FAU-Aibo (Batliner, Steidl, & Nöth, 2008) encapsulating 47 adults and 51 children speakers respectively. CASIA (Zhang & Jia, 2008), CHEAVD (Y. Li, Tao, Chao, Bao, & Liu, 2017) and CASS (Li et al., 2000) resources were developed for the Mandarin language containing six emotions in each database respectively. Emovo was also introduced for the Italian language having seven emotions performed by six actors in 14 sentences (Costantini, Iaderola, Paoloni, & Todisco, 2014). Keio ESD was introduced for the Japanese language containing 940 instances uttered by one male speaker in 47 emotional states (Moriyama, Mori, & Ozawa, 2009). RECOLA for the French language is a 9-hour emotional speech repository containing 46 speakers focusing on five social behaviors (Ringeval, Sonderegger, Sauer, & Lalanne, 2013).

An emotional dataset for the Spanish language containing two sessions and four emotions uttered by one actor was developed in 1999 (Montero, Gutiérrez-Arriola, Colás, Enriquez, & Pardo, 1999). Russian repository contains 61 native speakers speaking ten sentences in six emotions (Makarova & Petrushin, 2002). Database for Danish language was also introduced with five emotions enacted by four radio theater actors (Engberg, Hansen, Andersen, & Dalsgaard, 1997). Moreover, for the Turkish language, TURES (Oflazoglu & Yildirim, 2013) and BAUM-1 (Zhalehpour, Onder, Akhtar, & Erdem, 2016) were developed with 582 and 31 speakers containing seven emotions in each repository, respectively. REGIM-TES was designed for Arabic having 12 actors uttering ten sentences in five emotions (Meddeb, Karray, & Alimi, 2017). For the Persian language, SheMo was introduced in which 87 native speakers recorded audio clips in six emotions (Nezami, Lou, & Karami, 2019). Resources for Telugu and Hindi languages are also available containing ten speakers acting on a script based on 15 sentences in the respective language (Koolagudi, Maity, Kumar, Chakrabarti, & Rao, 2009; Koolagudi, Reddy, Yadav, & Rao, 2011). Both repositories contain eight emotions. Table 2 summarizes the specifications for resources available for other languages.

2.3 Speech emotion recognition systems

The ultimate goal of developing repositories is to train data-hungry learning algorithms for emotion recognition from speech signals. Deep learning models produce better results than shallow learning algorithms (Fayek, Lech, & Cavedon, 2017). These architectures are also flexible to various input features that can be extracted from acoustic signals. Neural network models for SER are built around three-block architecture: The convolutional neural network (CNN) block to extract local features and recurrent neural network (RNN) to capture context from all local features or to give attention (Bahdanau, Cho, & Bengio, 2015) to only useful features. The weighted mean is taken of the result using a dense neural network layer (M. Chen, He, Yang, & Zhang, 2018; Zhao, Mao, & Chen, 2019). It has been demonstrated that the CNN module combined with long short-term memory (LSTM) neural networks works better than standalone CNN or LSTM-based models for SER tasks in cross-corpus settings (Parry et al., 2019). Other deep learning-based models like zero-shot learning, which learns using only a few labels (Xu et al., 2019) and Generative Adversarial Networks (GANs) to generate synthetic samples for robust learning, have also been studied (Chatziagapi et al., 2019).

Various acoustic features are utilized for solving speech emotion recognition problems. They vary from using Mel-frequency Cepstral Coefficients (MFCC) (Sahidullah & Saha, 2012) to using Mel-frequency spectrogram (Stevens, Volkmann, & Newman, 1937). Other most commonly used ones are ComParE (Schuller et al., 2016) and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2015). To improve the performance of SER systems, auxiliary data has been incorporated with the speech corpus. A movie script database has been used to generate a personalized profile for each speaker while classifying emotions for individual speakers (J.-L. Li & Lee, 2019). Fusion techniques, that fuse words, and speech features to identify emotion have been studied by Sebastian and Pierucci (2019).

Some recent works focus on multi-modal techniques utilizing textual and audio features using BERT and attention-based GRU (Gated Recurrent Unit) to predict emotions with 75% unweighted accuracy on IEMOCAP (Kumar, Kaushik, & Raman, 2021). Another work focuses on improving the SER system built from ASR by introducing an self-attention mechanism and confidence measure for word-level instances (Santoso, Yamada, Makino, Ishizuka, & Hiramura, 2021). Leem, Fulford, Onnela, Gard, and Busso (2021) proposed a novel architecture consisting of decoupled ladder network that predicts emotion from noisy speech signals. Another work proposed a novel architecture utilizing traditional LSTM with graph isomorphism network exceeding the state-of-the-art graph-based architectures with 65.53% accuracy for IEMOCAP dataset (Liu & Wang, 2021). Other recent works that efficiently recognize emotions from audio signals utilize TDNN architectures (Kumawat & Routray, 2021), Stochastic Process Regression (T, Sanchez, Tzimiropoulos, Giesbrecht, & Valstar, 2021), and Multi-Task Learning (Cai, Yuan, Zheng, Huang, & Church, 2021).

Table 2: Summary of datasets available with number of emotions (No. of Em.) and number of samples or total duration (No. of sam./dur.) for other languages (lang.) English (Eng.), German (Ger.), Mandarin (Man.), Italian (Ital.), Japanese (Jap.), French (Fre.), Spanish (Sp.), Russian (Ru.), Danish (Dan.), Turkish (Tur.), Arabic (Ara.), Persian (Per.), Telugu (Tel.) and Hindi (Hin.).

Dataset	Lang.	No. of Em.	No. of sam./dur.
IEMOCAP (Busso et al., 2008)	Eng.	9	10, 039
SAVEE (Jackson & Haq, 2014)	Eng.	7	480
MSP-IMPROV (Busso et al., 2016)	Eng.	5	9*hr*
RAVDESS (Livingstone & Russo, 2018)	Eng.	7	336
VESUS (Sager et al., 2019)	Eng.	5	12, 600
EmoDB (Burkhardt et al., 2005)	Ger.	7	535
VAM (Grimm et al., 2008)	Ger.	3	947
FAU-Aibo (Batliner et al., 2008)	Ger.	6	9h
CASIA (Zhang & Jia, 2008)	Man.	6	500
CASS (Li et al., 2000)	Man.	6	6h
CHEAVD (Y. Li et al., 2017)	Man.	6	2hr 20min
Emovo (Costantini et al., 2014)	Ital.	7	588
Keio ESD (Moriyama et al., 2009)	Jap.	47	940
RECOLA (Ringeval et al., 2013)	Fre.	5	9h30min
Montero et al. (1999)	Sp.	4	−
Makarova and Petrushin (2002)	Ru.	6	3, 660
Engberg et al. (1997)	Dan.	5	10min
TURES (Oflazoglu & Yildirim, 2013)	Tur.	7	5100
BAUM-1 (Zhalehpour et al., 2016)	Tur.	13	1, 510
REGIM-TES (Meddeb et al., 2017)	Ara.	5	600
SheMo (Nezami et al., 2019)	Per.	7	3h 25min
Koolagudi et al. (2009)	Tel.	8	1, 200
Koolagudi et al. (2011)	Hin.	8	1, 200

3 Script Design

Resource construction encapsulating acoustic signals is a tiresome endeavor with numerous inherent challenges. Our target is to build a vast repository containing emotional clips recorded in different accents for a reliable speech emotion recognizer targeting the Urdu language. Formally, let U be a vector of utterances where each instance $u_i \in U$ is an audio clip of length t_i. The goal is to acquire a sample U representing Urdu language that is phonetically balanced, lexically complex, manifesting diverse accents, and contains a uniform distribution of emotions. This section provides details for script design, the recording process of SEMOUR$^+$ and the obstacles faced during the process.

Urdu is an Indo-Aryan language with 171 million speakers worldwide (Eberhard et al., 2020). It is the national language of Pakistan and an official language in many states of India. The language is known for its rich inventory of 67 phonemes compared to 36 phonemes in English and 35 phonemes in the Chinese language (Kabir & Saleem, 2002). It contains a huge list of consonants, short, long nasal, and long non-nasal vowels. It follows a Perso-Arabic written

**Table 3**: The top five frequently used Urdu language words.

Sr. no.	Word	Frequency
1	کے	743949
2	میں	582882
3	کی	575545
4	ہے	466908
5	اور	413788

style named *Nastaliq* (Ijaz & Hussain, 2007). There are four dialects: Urdu, Dakhini, Hyderabadi Urdu and Rekhta (Ghulam & Soomro, 2018). There exist six diverse types of accents namely Urdu, Punjabi, Pashto, Saraiki, Balochi, and Sindhi for Urdu dialect (Qasim, Nawaz, Hussain, & Habib, 2016). Punjabi is spoken in Punjab, whereas Sindhi in Sindh, Pashto and Saraiki in Khyber Pukhton Khaw (KPK) along with Balochi accent spoken in Baluchistan respectively. In contrast, the Urdu accent predominates in all provinces of Pakistan (Atta, van de Weijer, & Zhu, 2020). A language spoken by such a large community poses diverse challenges while procuring a speech dataset. Diversity in dialect and accent is one of the major concerns and selecting speakers to cover such diversity is an uphill task. Moreover, coverage of all phonemes plays a crucial role in designing a rich acoustic dataset.

As mentioned above, our target while constructing this dataset is to mimic daily conversations among native speakers. Therefore, a phonetically balanced repository ensures that frequencies of all sound utterances closely approximate their densities in the set of spoken words. To achieve this goal, we utilize two sources. Firstly, we considered the group of top 5000 most frequently used Urdu words that appear in the list collected here (Ijaz & Hussain, 2007). These words are extracted from news, finance, entertainment, sports, personal communications and consumer information. A sample of the first five most frequently used words extracted from that list is shown in Table 3.

Secondly, we use a complete Urdu lexicon of about 46, 000 words collected in Haris Bin Zia and Athar (2018). A uniformly random sample of words from any one of these sources will have, in expectation, the characteristic of being phonetically proportioned. However, we also wanted our dataset to have phrases and sentences with sufficient lexical complexity. The words present in the script are composed of names of colors e.g. [جامنی (Violet), سرخ (Red), سلیٹی (Gray)] , number names [گیاره (Eleven), نو (Nine), پانچ (Five)] and topographical words [ندیاں (Rivers), چٹانوں (Rocks), ریگستانوں (Deserts)]. The two-word phrases are composed of a list of adjectives associated with the nouns e.g. پاک سرزمین (Clean land). The rest of the script consists of simple structured Urdu sentences e.g. بس کا سفر کریں (Take the bus) . So, overall our script consists of 235 instances composed of 43 common words, 66 two-word phrases, and 126 Urdu sentences, as shown in Table 4. Frequently used vocabulary, and a small subset of emotionally enabled sentences against each emotion have also been incorporated in the preparation of the script. It contains both neutral and
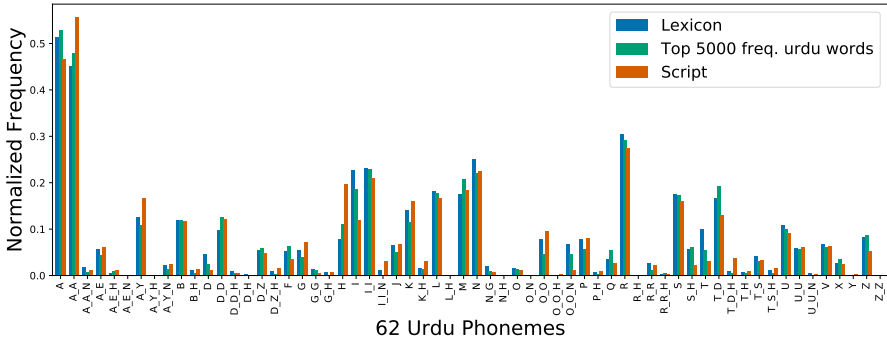
Figure 2: Phoneme comparison of our designed script with the Urdu lexicon of 46,000 words, and 5000 most frequently used words (Haris Bin Zia & Athar, 2018). On the *x*-axis, we have a list of 62 phonemes, and on the *y*-axis we have normalized phoneme occurrence frequency. The histogram shows that SEMOUR$^+$ is phonetically balanced compared to the other two standard datasets. The normalized frequency scale ranges from 0 to 1.

Table 4: Utterance and script level statistics for SEMOUR$^+$.

Category	Instances	Category	Statistics
Words	43	Avg. time per instance	1.755 sec
Phrases	66	Avg. time per Actor	38 min, 38 sec
Sentences	126	Avg. time per Emotion	4 min, 41 sec
Total instances	**235**	**Total instances**	**27,640**
Unique words	**559**	**Total duration**	**13 hr, 55 min, 29 sec**

emotionally charged sentences for each emotion. Table 5 shows a sample of such sentences that are phonetically proportioned.

Haris Bin Zia and Athar (2018) presented a grapheme to phoneme conversion algorithm that produces a phoneme level transformation of Urdu words e.g. for the word جهاؤ we get the following conversion: **B, A, H, A_A, O_O**. This model consists of two LSTM layers with 512 units in each layer and trained on Lexicon of 36,000 words with 64% accuracy. We have utilized this model to extract phonemes against our lists of words for script and the other two sources mentioned above. We plot the computed normalized frequency as described in Figure 2. The script was modified and the procedure mentioned above was repeated until a reasonable balance was achieved. It can be observed that our script approximately covers all phonemes as frequently as they appear in the two standard datasets. Also, note that the model (Haris Bin Zia & Athar, 2018) is trained for 62 phonemes. Hence Figure 2 shows the phoneme normalized frequencies against 62 phonemes.

Once the script was finalized, eight speakers from Punjab province were recruited to elicit 235 instances of eight emotions. The selected speakers have

Table 5: A sample of sentences selected from our short script against each perceived emotion. The left column is the English translation of the Urdu sentences taken from the script.

English Translation	Urdu Sentences	Emotion
Control your anger	اپنے غصے کو قابو میں رکھو	Anger: غصہ
This article will be learned again	یہ مضمون پھر سے یاد کرنا پڑے گا	Boredom: بیزاری
What a mess	کتنی گندی صفائی کی تھی	Disgust: نفرت
The gardens are empty	باغات خالی ہیں	Neutral : غیر جانبدار
Are you really telling the truth?	تم کیا واقعی سچ کہہ رہے ہو	Surprise: حیرت
Ahmed went crazy with joy	احمد خوشی سے پاگل ہوگیا	Happiness: خوشی
He was crying	وہ رو رہا تھا	Sadness: اداسی
I was scared in the dark night	مجھے اندھیری رات میں ڈر لگا تھا	Fear: خوف

Urdu as their first and English as their second language with some acting experience in the past. All speakers have a minimum of intermediate school education aged from 20 to 50 years old. All speakers belong to Lahore city except one, which belongs to Gujranwala.

We have also focused on accents' diversification and covered three major provinces: Punjab, KPK and Baluchistan. We chose a subset of 100 sentences from our script for elicitation in eight emotions. Volunteer speakers from various hometowns belonging to the aforementioned selected provinces were recruited to cover various accents. There are five females and 11 male speakers ranging from 24 to 33 years. All speakers are well-educated with a minimum of a bachelor's degree. Moreover, most of them have different first and second languages such as Urdu, Pushto, Persian, Punjabi and English, with varying Urdu speaking and listening skills. All of the speakers' current residing city is Lahore, with little or no experience in acting which is advantageous for eliciting natural emotions closer to speech in reality. We have three speakers from Faisalabad, two from Sargodha, Sialkot, Swat and Swabi, one from Lahore, Rawalpindi, Bahawalnagar and Kohat making 16 speakers in total. Also, note that Lahore, Rawalpindi, Faisalabad, Bahalwanagar, Sargodha and Sialkot belong to Punjab province, whereas Quetta is located in Balochistan and Swabi, Swat and Kohat belong to KPK province respectively. The speaker IDs. and their demographic information have been made publicly available for further study at: https://tinyurl.com/yc499z7k.

We used the services of a local radio studio to conduct soundproof recording sessions. After the recording, basic noise reduction, voice normalization, and amplification techniques are applied to the audio clips. Each recording is manually segmented based on script sequence, and emotion instance. The final dataset consists of $27,640$ utterances with a cumulative duration of **13 hr, 59 min, 33 sec**. We have uploaded our dataset[1] and are making it publicly available for the research community. Each audio clip of the dataset has 2 channels (stereo) with sample rate, sample size, and bit rate of 44.100 kHz,

---

[1]https://tinyurl.com/yc499z7k

16 bit, and 1411 kbps, respectively. Table 4 showcases the details of utterance and script-level statistics.

4 Human Evaluation

The human annotation on the subset of SEMOUR$^+$ was the first experiment performed for retrieving the human accuracy against the audio clips. We chose 18% of the audio clips uniformly at random from our dataset containing the first eight speakers and presented them to the evaluators one at a time with a list of questions through an application. The evaluators were given a short training on the use of the survey application and the context of each term used in the questions. We chose sixteen annotators to evaluate about 5,000 instances, with each clip receiving at least two evaluations. All of the evaluators have proficient Urdu speaking and Listening skills. All but two have Urdu as their first language; the other two have Urdu as a second language. All the evaluators have a minimum of secondary school education aged from 17 to 32 years old.

After listening to each clip, annotators were asked to input two evaluations, namely *discrete categorical* (a vote among neutral, happiness, surprise, sadness, boredom, fear, anger, and disgust) and *continuous attribute* (a numeric value for each of valence, dominance, and activation). Valence encapsulates degrees of pleasantness elicited in an audio clip varying from negative to positive, activation depicts the levels of intensity such as calmness or excitement, and dominance portrays the speaker's control over the utterance i.e. weak or strong (Russell & Mehrabian, 1977).

We performed extensive analyses on the feedback received from annotators. As shown in Table 7, we see an average accuracy of 78%, i.e., on average, an evaluator correctly identified the emotion in a clip with 78% accuracy. We observe an average accuracy of 79% as shown in Table 8. This is a very high accuracy compared to a random classification resulting in an accuracy of 12.5%. One can conclude, that most of the audio files are labeled with the correct emotion.

Secondly, this also shows that humans can correctly perceive labeled emotions in most of these clips. However, we observe some variance in the accurate recognition of different emotions. For example, two emotions, disgust and fearful, were identified with the lowest accuracy as compared to other emotions. It turns out that some instances of disgust were confused with anger and neutral emotions. We believe that this emotion is very hard to discriminate from other emotions in general, and probably, not as commonly used as other emotions. Similarly, fear was incorrectly classified as sadness in some instances because shivers while crying were perceived as hiccups.

Furthermore, these ratings also give us a yardstick to measure the performance of individual actors. For each actor, accuracy, precision, recall, f1-score, and Cohen's kappa score for measuring rater-s' agreement, are presented in Table 6. It can be observed that actor number 8 performed well, and actor

Table 6: Individual performance of actors with majority voting

Actors	Accuracy	Precision	Recall	F1-Measure	Inter-evaluator score
1	0.66	0.68	0.66	0.65	0.54
2	0.87	0.88	0.87	0.87	0.71
3	0.79	0.8	0.79	0.79	**0.89**
4	0.86	0.86	0.86	0.86	0.67
5	0.82	0.83	0.82	0.81	0.73
6	0.8	0.81	0.8	0.79	0.71
7	0.66	0.72	0.66	0.64	0.42
8	**0.91**	**0.92**	**0.91**	**0.91**	0.88

number 3 had the highest rater-s' agreement. Since, all scores are greater than 0.4, (i.e., fair agreement) therefore, we conclude that our actors performed well, and the general audience could distinguish among the emotions uttered.

We have also performed experiments based on the syntactical complexity of our script, as single words and phrases perceived naturally encapsulate fewer emotions than sentences. Intuitively, it should be harder to elicit emotion in a word or phrase. Our results confirm this intuition in terms of accuracy, precision, recall, and f1-measure as shown in Figure 3b. Raters were asked to rank activation and dominance for each utterance, along with discrete labels. With majority voting, our dataset conforms with generally perceived notions of these values in the discrete emotions, e.g., sadness has *low* activation and *weak* dominance as compared to anger. The results are as shown in Figure 3c.

Lastly, we applied *t-Distributed Stochastic Neighbor Embedding (t-SNE)*, a dimensionality reduction technique to our ratings for detailed analysis of emotions. Figure 4a shows the t-SNE relation between ground truth and the independent variables the evaluator had control on, i.e., Emotion-Tag, Valence, Activation, Dominance, and Genuineness (TVADG). The distribution for various emotions in the figure shows the relative impacts of all of the evaluated properties, and an overlap between two color classes indicates a possibly incorrect annotation. For example, it seems that point clouds for happiness and surprise emotions are close to each other, implying that they may have been misclassified for each other. Similarly, sadness and disgust classes have some overlap too. On the other hand, classes of anger, neutral, boredom, and disgust emotion seem quite distinguishable. Figure 4b is the t-SNE relation between ground truth and the independent variables annotator had excluding the emotion-tag property, i.e., VADG. Since the attributes have a very small distribution, the clusters of different classes are subsumed into one another. Although, the boredom class remains the most distinguishable one.

(a)

(b)

(c)
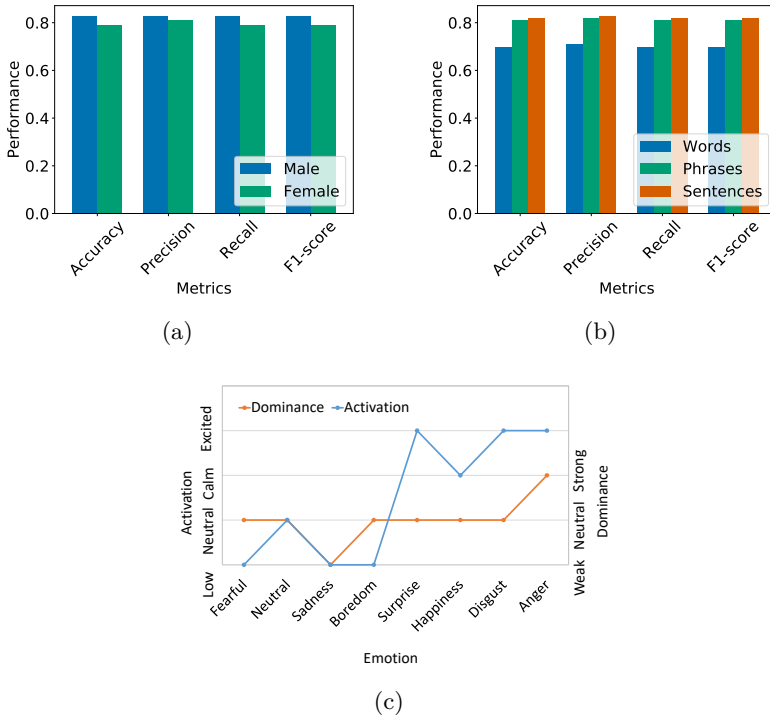
Figure 3: Results for experimentation on human annotation : (a) performance analysis against evaluation metrics for male and female actors. Male actors performed better than female actors when their clips were annotated. The values against males for accuracy, recall, precision, and F1 score are $0.83, 0.83, 0.83$ and 0.83 respectively. The values against females for accuracy, recall, precision, and F1 score are $0.79, 0.81, 0.79$ and 0.79 respectively. (b) performance analysis against evaluation metrics for the script's lexical complexity. Sentences cover rich emotion than phrases and so on. The values against words for accuracy, recall, precision, F1 score are $0.70, 0.71, 0.70$ and 0.70 respectively. The values against phrases for accuracy, recall, precision, F1 score are $0.81, 0.82, 0.81$ and 0.81 respectively. The values against sentences for accuracy, recall, precision, F1 score are $0.82, 0.83, 0.82$ and 0.82 respectively. Performance scale for (a) and (b) ranges from 0 to 1. (c) Continuous attributes categorization for each emotion. The left and right y-axis depict activation (low, natural, calm, and excited) and dominance (weak, neutral, and strong) respectively. Surprise has excited activation and neutral dominance.

Table 7: Confusion matrix for intended vs. perceived labels with all ratings and average accuracy (A.A). Emotions are labelled as: Anger as Ang., Boredom as Bor., Disgust as Dis., Fearful as Fer., Happiness as Hap., Neutral as Neu., Sadness as Sad., Suprise as Sur. and Other as Oth.

A. A. =78%		Perceived Labels								
		Ang.	Bor.	Dis.	Fea.	Hap.	Neu.	Sad.	Sur.	Oth.
Ground truth	**Ang.**	**972**	3	46	5	13	52	0	142	14
	Bor.	1	**1101**	21	5	2	65	45	0	7
	Dis.	167	30	**770**	29	34	127	15	55	25
	Fea.	1	22	26	**781**	12	70	246	73	15
	Hap.	1	5	5	7	**1067**	64	8	78	13
	Neu.	10	84	30	5	7	**1076**	12	8	16
	Sad.	0	49	7	149	12	25	**995**	6	1
	Sur.	34	2	28	15	42	85	2	**1021**	16

(a)   (b)
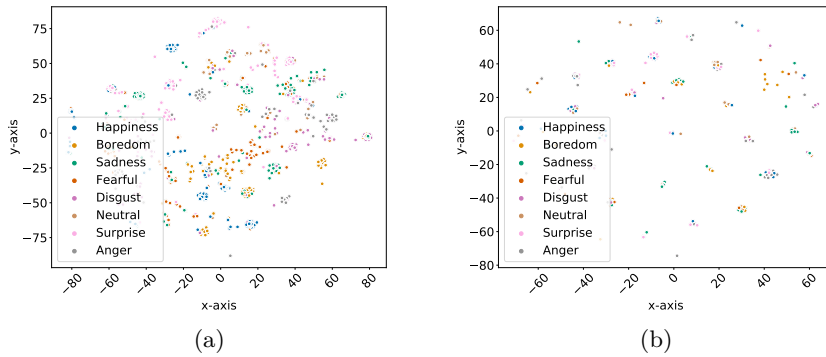
Figure 4: t-SNE plots against human annotation. (a) Ground truth plotted against Tag, Valence, Activation, Dominance, Genuineness (TVADG). (b) Ground Truth plotted against Valence, Activation, Dominance, Genuineness (VADG).

5 Speech Emotion Recognition

The ultimate goal of building a sizeable balanced dataset is to be able to train machine learning models that can predict emotions in an unseen sound clip. This section discusses the performance and evaluation of the first machine learning model to use SEMOUR$^+$ as a dataset. We aim to solve a multi-label classification problem that predicts emotion based on features extracted from an acoustic clip. Once the dataset U is constructed, let L be a vector of labels such that $l_i \in L$ is the emotion associated with the input instance u_i. Let the training set be V, and the testing set be W such that $V, W \subseteq U$, $V \cap W = \emptyset$

Table 8: Confusion matrix for intended vs. perceived labels with majority voting and average accuracy (A.A). Emotions are labelled as: Anger as Ang., Boredom as Bor., Disgust as Dis., Fearful as Fer., Happiness as Hap., Neutral as Neu., Sadness as Sad., Suprise as Sur. and Other as Oth.

A. A. =79%		Perceived Labels								
		Ang.	Bor.	Dis.	Fea.	Hap.	Neu.	Sad.	Sur.	Oth.
	Ang.	**542**	1	11	2	6	18	0	42	1
	Bor.	1	**605**	2	2	0	11	2	0	0
	Dis.	141	25	**388**	6	6	45	1	9	4
Ground truth	**Fea.**	1	21	22	**473**	3	27	56	15	4
	Hap.	1	5	3	6	**572**	15	1	18	3
	Neu.	10	78	23	5	7	**498**	2	0	1
	Sad.	0	45	7	120	11	18	**418**	0	1
	Sur.	31	2	19	11	30	63	1	**451**	13

and $V \cup W = U$. Our goal is to learn a function $\mathcal{F} : U \rightarrow L$ so that, (1). function $\mathcal{F}$ outputs correct emotion l_j for an input occurrence $v_j \in V$ for the maximum number of instances in V, (training accuracy), (2) $\mathcal{F}$ outputs a correct emotion for an input instance $w_j \in W$ that corresponds to an unseen audio clip, assuming unseen clip is drawn randomly from the distribution of W, (test accuracy).

This section elaborates the evaluation of our composed resources using machine and deep learning techniques. The machine learning is applied to the first eight native speakers of the Punjab province. Then, we extend our dataset to diversify the accents to improve the accuracy by applying deep learning models to the whole repository.

5.1 Machine Learning

Starting from the basic architecture, we aim to solve speech emotion recognition problem using neural networks and report performance metrics in various experimental settings using SEMOUR$^+$. The following experiments under machine learning were performed on the first eight actors selected from Punjab province. Later, we extend our speakers in terms of accents in the deep learning section to improve the speaker-independent accuracy. The features used for training were Mel-frequency Cepstral Coefficients (MFCCs), chromagram, and Mel-spectrogram. Collectively, for each sample, an array of 40 coefficients of MFCC, 12 pitch classes values, and 128 mean values of Mel-spectrogram form a feature vector of size, $d = 180$. This feature vector is used for the task of classification. Visualization of features for neutral emotion audio sample are shown in Figure 5.

A five-layered neural architecture was trained for classification, as shown in Figure 5. Four hidden layers with 1024, 512, 64, and 48 neurons were used respectively along with 8-dimensional fully connected output layer. The parameters for epochs, L1-regularization, learning rate, and batch size were set

Table 9: Result comparison with classical machine learning algorithms (Algo.) and our proposed deep neural network for 80% training and 20% testing random split.

Algo./Eval. Metrics	Accuracy	Precision	Recall	F1- score
Gaussian Naive Bayes	0.44	0.45	0.44	0.41
Logistic Regression	0.64	0.64	0.64	0.64
SVM	0.64	0.64	0.64	0.64
Decision Tree	0.69	0.69	0.69	0.69
ANN	0.86	0.88	0.86	0.86
Random Forest	0.87	0.87	0.87	0.87
Our Method	**0.91**	**0.91**	**0.91**	**0.91**

to 30,0.001, 0.001, and 50, respectively, along with *Scaled Exponential Linear Units (SELU)* and *softmax* are activation functions for hidden and output layers. The three features mentioned above were extracted using the Librosa speech library and were used as an input to our proposed architecture McFee et al. (2015). Three different variations of the experiment namely *random splits, leave one speaker out*, and *gender-based setting* were selected for detailed analysis.

Our proposed architecture outperforms existing machine learning algorithms as shown in Table 9. The comparison features of SEMOUR$^+$ and the accuracy of our model are compared with the current dataset in Table 10. In the following subsections, we provide extensive details regarding the experimentation setup.
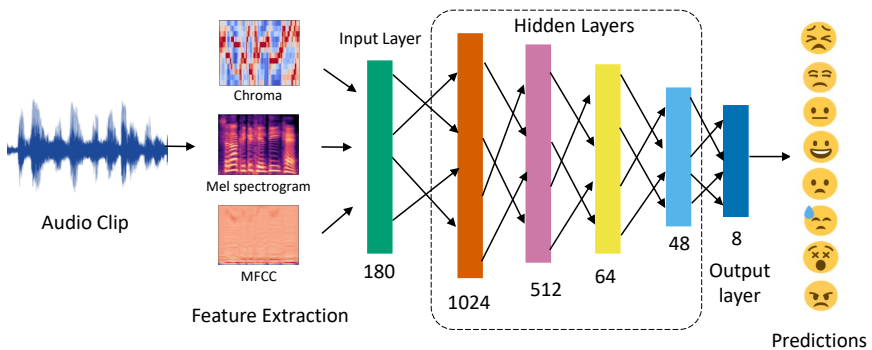
Figure 5: Proposed 5-layered neural network for Speech Emotion Recognition (SER). Three core features namely MFCCs, chromagram, and Mel-spectrogram are extracted from each audio clip and fed to a 5- dense-layered neural network to predict 8 complex emotions.

Table 10: Testing our proposed architecture and dataset (Dat.) against the number of emotions (No. of em.), number of instances (No. of in.) and performance metrics accuracy (Ac.), precision (Pre.), recall (Rec.), F1-score (F1.) with Latif et al. (2018) by taking 80% of SEMOUR$^+$ dataset for training and 20% for validation from speakers $1 - 8$.

Dat.	No. of em.	No. of in.	Acc.	Pre.	Re.	F1.
Latif et al. (2018)	4	400	0.21	0.3	0.21	0.23
SEMOUR$^+$	**8**	**15,040**	**0.91**	**0.91**	**0.91**	**0.91**

5.1.1 Stochastic Division of Train and Test sets

The first experiment we designed, was based on a stochastic split experiment, where we tested our accuracy on the complete repository. For this experiment, we used the architecture as explained in Figure 5. We trained our model for 100 epochs on a randomly selected 90% dataset ($13,536$ instances). Once trained, we tested it on the remaining 1504 audio clips and obtained the highest accuracy of 92% and an average accuracy of 90%. To validate the experiment, a 10-folds cross-validation technique was used. The results are visualized in Figure 6c. Accuracy for individual emotion class was also analyzed for the variance. We observed that our model performed exceptionally well on the boredom and neutral emotions with 99%, and 98% accuracy, respectively. The worse performing emotions were Fearful and Happiness which were identified with an accuracy of 86%, and 85%, respectively.

5.1.2 Gender-based Analysis among Actors

The second experiment, we conducted, was a gender-based experiment where we had a binary combinatorial testing technique. We have an equal distribution of male and female actors in our dataset, and therefore, in a random setting, one would expect a balancing classification accuracy. All four binary combinations were evaluated, i.e., binary choices are male and female actors and training and testing samples.

As shown in Table 11 the experiments on the same gender yielded excellent results whereas the cross-gender study resulted in significantly lower accuracy. We believe that it might be since each speaker primarily has an independent distribution that is not concerning a specific gender, but is rather unique to each individual. Henceforth, we conducted a leave-one-out experiment on the same gender and the accuracy dropped from 96% to 50% for the males and 92% to 45% for the females. The details of the leave-one-out experiment are discussed in the next subsection.

Table 11: SER results for different variations against our proposed neural network model.

Tech./Eval. Metrics	Accuracy	Precision	Recall	F1- score
Leave one out speaker	0.39	0.39	0.36	0.35
Gender Analysis-Male	**0.96**	**0.96**	**0.96**	**0.96**
Gender analysis -Female	0.92	0.92	0.92	0.92
Random splits	**0.92**	**0.93**	**0.92**	**0.93**

5.1.3 Speaker-Independent Cross-validation Experiment

In this experimentation setting, the model was trained on 7 actors and tested on the remaining actor with overall testing accuracy, recall, precision, and f1-score of 39%, 39%, 36%, and 35%, respectively, averaged over all actors as shown in Table 11. Predictions on actor 3 have relatively better results in the individual actor analysis. As mentioned earlier, this below-par accuracy is due to a significantly different distribution of features for individual actors, as seen in Figure 6a. The model fails on an unseen actor because of diversity in style to utter emotions. Training accuracy was observed to be 100%, even with extensive experimentation with reduction of layers, the addition of regularization, and low learning rate, the testing accuracy did not improve, which shows that a simple deep neural network is not suitable for this variation of the experiment. We propose that complex models like LSTMs and transformers should be tested to better represent heterogeneous distributions among speakers.

Moreover, our model can only identify anger and surprise emotions for all actors and perform well while predicting anger emotions as compared to others, as shown in Figure 6b. Boredom, happiness, and sadness emotions cannot be discriminated against by all actors. Disgust has the lowest overall accuracy for all speakers. We conclude that there exists heterogeneity among speakers and emotions collectively, which is only natural considering diversity in emotion utterance for each individual, i.e., elicitation of similar emotions can vary speaker-wise.

(a)

(b)

(c)
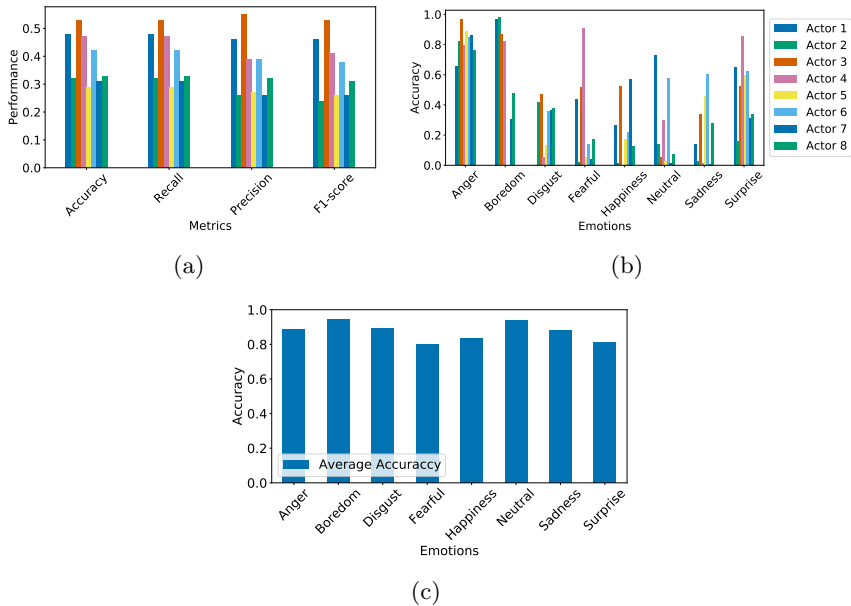
Figure 6: Summarized results for SER experiments with performance ranging from 0 to 1. (a) Performance analysis across actors against evaluation metrics. The mean value for accuracy, recall, precision, F1 score is $0.39, 0.39, 0.36$ and 0.36 respectively. The standard deviation for accuracy, recall, precision, F1 score is $0.09, 0.09, 0.1$ and 0.1 respectively. (b) Performance analysis across all emotions against each actor. The mean value for anger, boredom, disgust, fearful, happiness, neutral, sadness and surprise is $0.58, 0.52, 0.27, 0.27, 0.28, 0.24, 0.24, 0.23$ and 0.50 respectively. The standard deviation for anger, boredom, disgust, fearful, happiness, neutral, sadness and surprise is $0.09, 0.41, 0.18, 0.31, 0.21, 0.27, 0.22$ and 0.22 respectively. The legend for (a) and (b) is shown on the right side. (c) Average accuracy against all emotions for stochastic testing. The graph shows that average accuracy is not only very high, it is also stable and consistent among all emotions. The mean value for anger, boredom, disgust, fearful, happiness, neutral, sadness and surprise is $0.89, 0.92, 0.89, 0.80, 0.84, 0.91, 0.89, 0.84$ and 0.89 respectively. The standard deviation for anger, boredom, disgust, fearful, happiness, neutral, sadness and surprise is $0.03, 0.04, 0.04, 0.08, 0.06, 0.02, 0.06$ and 0.05 respectively.

5.2 Deep Learning

As discussed in the former section, our speaker-independent accuracy was not well enough. We improve our speaker-independent performance by diversifying accents among speakers. We enhance our dataset in terms of the number of speakers from 8 to 24. In this section, we utilize deep learning models such as

Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTMs), and Transfer learning on state-of-the-art models like VGG 19 (Simonyan & Zisserman, 2014) to improve our evaluation metrics by training on all speakers. The elaborated description of our deep learning experimentation is shown in Figure 7.
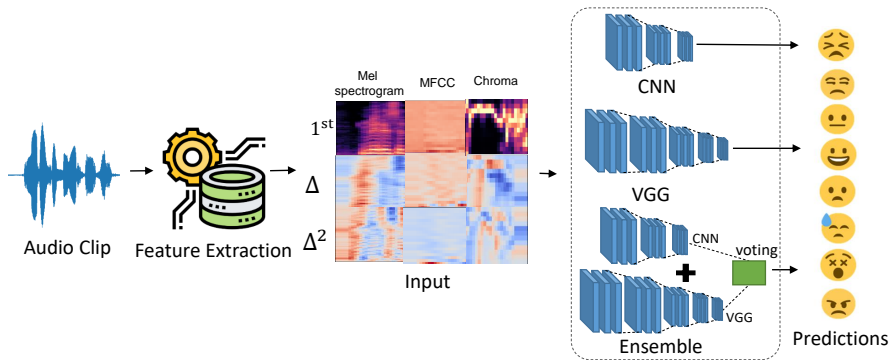
Figure 7: The figure elaborates on the model used to predict emotions. The raw audio clip is first converted into acoustic features namely Mel spectrogram, MFCCs and Chroma along with their deltas and double deltas to form the input for our proposed model. The output from each model is passed through eight-unit output layer for predicting emotions.

5.2.1 Feature Selection

An essential part of a deep learning architecture is to appropriately select the feature set to be input to the model so that it can learn hidden feature representations. In most recent works on SER, many researchers use some enhanced variation of the speech features like Mel-frequency Cepstral Coefficients (MFCC), Mel-frequency spectrogram, ComParE , eGEMAPS, etc. (Cummins et al., 2017; Eyben et al., 2015; Schuller et al., 2016). The MFCCs, Mel spectrograms and chromagram all have a 2-D representation of an audio signal suggest the intuitive use of Convolutional Neural Networks (CNNs) to extract the salient hidden features in the spectrograms. We chose all these three features extracted from a fixed-sized audio sample. We pad/clip all the audio samples on 4s signal length and then extract the MFCC, Mel spectrogram, and chromagram features from these samples to obtain a consistently shaped feature set. We also get the first and second order deltas of all these three features to acquire more useful information. There are 128 Mel spectrogram features, 13 MFCCs, and 12 chromagram features making a total of 153. The 4s audio samples maps to a 126-shaped spectrogram. Concatenating the first and second-order deltas makes the overall feature set size of 153×378. After extensive experimentation, it is proved that including these deltas and the original features improve the model's emotion recognition accuracy, as shown

in Table 12 and Table 13 for the CNN and VGG-19 architectures, respectively. It can be seen from the tables that the deltas can increase up to 5% accuracy when used with the original features.

5.2.2 CNNs

Extensive experimentation was conducted on CNNs with variations in layers and filter size. The architecture was finalized with five convolution layers with batch normalization, max pooling, and drop-out operation in between each layer. The first layer contains a 2D convolution filter and kernel size of 4 and 5, respectively with a stride of size 1. Max-pooling includes a pool size of 2×2 with 2 as the stride size and the value 0.1 for the drop-out layer. The second layer contains a 2D convolution filter and a kernel size of 4 and 10 respectively with a stride of size 1. Max-pooling contains a pool size of 2×2 with a value of 1 as stride size and a value of 0.1 for drop-out layer. The third layer contains a 2D convolution filter and a kernel size of 8 and 15 respectively with stride size 1. Max-pooling contains a pool size of 6×6 with 1 stride size and 0.1 value for the drop-out layer. The fourth layer contains a 2D convolution filter and kernel size of 8 and 16 respectively with stride size 1. Max-pooling contains pool size of 6×6 with stride size 2 and 0.1 for the drop-out layer. The fifth layer contains 2D convolution filter and kernel size of 8 and 17 respectively with stride size 1. Max-pooling contains pool size of 6×6 with 1 stride size and value 0.1 for drop-out layer. Lastly, the extracted features are flattened, then followed by the output layer with eight emotions. Each convolutional layer uses Scaled Exponential Linear Units (SELU) and softmax as the activation functions for convolution blocks and predictions at the last layer respectively. $L - 2$ Kernel and bias regularizers are also utilized with the value set as 0.01. The parametric values for batch size, epochs, learning rate, decay and momentum are $50, 50, 0.0001, 1e - 6$, and 0.9, respectively.

There was a five percent decrease in performance while using CNNs on eight speakers as compared to simple NN. Upon training the model on 24 speakers independently, the averaged accuracy, recall, precision and F1-measure is $47, 47, 54$ and 44 percent, respectively. The bar charts for 24 individual speakers' performances are shown in Figure 8. Apart from speaker id. 9 and 10, the performance of the model is reasonably well, reaching upto 75% accuracy. There is a 20% increase in performance while taking the averaged accuracy of the first eight speakers trained on the whole dataset. The averaged accuracy, recall, precision and F1-measure for eight speakers trained on a smaller dataset are $34, 34, 40$ and 29 percent, respectively. The mean value for the first eight actors trained on 24 speakers is 53% supporting our claim that the extension of SEMOUR$^+$ is useful for generalization. The overall evaluation metrics are shown in Table 12. It can be observed from the graph that CNN performs poorly on the smaller repositories as compared to vanilla neural networks but performs well on SEMOUR$^+$.
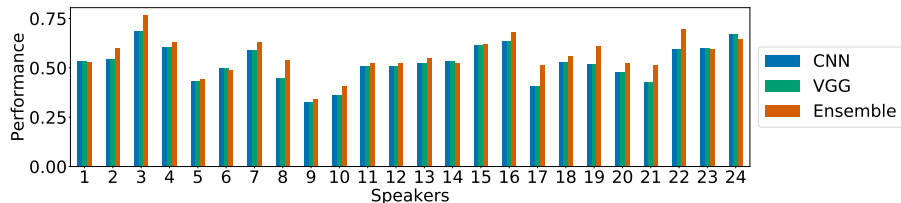
Figure 8: The figure elaborates the performance comparison between SEMOUR$^+$ dataset on our proposed CNN, VGG and Ensemble models. The speaker ids. are shown on the x-axis along with their accuracy measures along the y-axis. The range for performance is between 0.0 to 1.0.
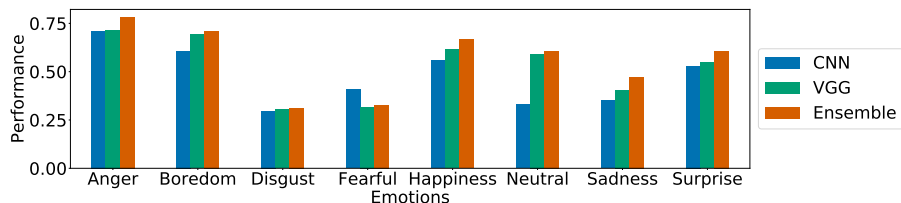
Figure 9: The figure elaborates emotion prediction results for 24 speakers trained and tested on our proposed models. The speaker ids. are shown on the x-axis along with their accuracy measures along the y-axis. The range for performance is between 0.0 to 1.0.

Table 12: The performance metrics for training on 8 speakers and 24- speakers using our 5-layered CNN model. All features include MFCCs, Mel spectrograms, Chromagrams and their deltas and double deltas whereas M+M+C represents training and testing on MFCCs, Mel spectrograms, Chromagrams only.

Model	Features	Accuracy	Recall	Precision	F1-Measure
NN	M+M+C	0.39	0.39	0.36	0.35
CNN-8	M+M+C	0.32	0.32	0.37	0.26
CNN-24	M+M+C	0.43	0.43	0.45	0.38
CNN-8	All	0.34	0.34	0.40	0.29
CNN-24	**All**	**0.47**	**0.47**	**0.54**	**0.44**

5.2.3 VGG-19

Building on the impressive results of CNN, next we propose the use of a pre-trained computer vision architecture namely VGG-19 (Simonyan & Zisserman, 2014) and fine-tune a few layers to achieve better results on SEMOUR$^+$. The feature set used for this architecture is the same one as used for the CNN model. Since the VGG-19 requires the RGB image as the input image, the architecture is modified so that the same 2-D feature set is copied thrice and is

concatenated together to feed to the pre-trained VGG-19 network. At the end of the network, the last layer is removed and the flattened layer is further fine-tuned with a following 1024 shaped dense and an output layer. At the output layer with 8-neurons and the softmax activation function, probabilities are obtained for each emotion which can be further processed to get the predicted emotion. The overall evaluation metrics for VGG-19 are shown in Table 13. It can be observed that VGG-19 can achieve 5% more average accuracy with the 24 actors speaker-independent configuration than of the CNN model. The parametric values for batch size, epochs, and learning rate are $64, 50$, and 0.0001 respectively.

Upon training the model on 24 speakers independently, the averaged accuracy, recall, precision and F1-measure are $52, 52, 56$ and 49 percent, respectively. The bar charts for 24 individual speakers' performances are shown in Figure 8. A total of 3% increase is observed while training on all 24 actors in the speaker-independent configuration. The averaged accuracy, recall, precision and F1-measure for eight speakers is $49, 49, 53$ and 46 percent respectively. As compared to training on a smaller set, the average accuracy value for the first eight actors trained on 24 speakers is 54% supporting the generalization claim of SEMOUR$^+$. The performance comparison is shown in Table 13 where a similar behavior to the CNN model is observed (refer to Table 12). A performance visualization is provided in Figure 9 based on the emotions of 24 actors. Aside from the fear emotion, the VGG-19 outperforms the CNN on the extended dataset.

5.2.4 Ensemble

Given different classification rules, majority vote classifiers or ensembles predict the class receiving the majority of predictions or votes (Murphy, 2012). Both the CNN and VGG-19 models for each speaker are joined together so that the overall prediction probabilities are generated by adding up the individual probabilities of both the models. The architecture is shown in Figure 7, where soft voting is performed to obtain the final probabilities of each class. Both the CNN and the VGG-19 models are given equal weights and whichever class gets the maximum probability is the predicted class by the ensemble model. In the speaker-independent configuration, separate CNN and VGG-19 models are trained for each actor and those models are run separately to obtain the soft probabilities for each of the emotion classes which is then summed up to get the ensemble predictions.

The overall average accuracy achieved through the ensemble classifier is 56% on the whole SEMOUR$^+$ dataset as shown in Figure 8. An overall comparison between the CNN, VGG-19, and the ensemble architectures is provided in Figure 8. As expected, the ensemble outperforms almost all the actors in comparison to the CNN and the VGG-19 models. From Figure 8, we can see that in some cases, the ensemble model is capable of achieving as high as 10% accuracy in comparison to both the CNN and the VGG-19 model. Besides, the ensemble is the only model that has achieved more than 70% accuracy

Table 13: The performance metrics for training on 8 speakers and 24- speakers using VGG 19 model. All features include MFCCs, Mel spectrograms, Chromagrams and their deltas and double deltas whereas M+M+C represents training and testing on MFCCs, Mel spectrograms, Chromagrams only.

Model	Features	Accuracy	Recall	Precision	F1-Measure
VGG-8	All	0.49	0.49	0.53	0.46
VGG-24	All	**0.52**	**0.52**	**0.56**	**0.49**
VGG-8	M+M+C	0.45	0.42	0.47	0.45
VGG-24	M+M+C	0.47	0.51	0.47	0.45

Table 14: The comparison between results of state-of-the-art, CNN-24, VGG-19 and the Ensemble model with all features.

Model	Accuracy	Recall	Precision	F1-Measure
NN	0.39	0.39	0.36	0.35
CNN-24	0.47	0.47	0.54	0.44
VGG-24	0.52	0.52	0.56	0.49
Ensemble	**0.56**	**0.56**	**0.60**	**0.53**

on an individual actor. The emotion-based performance of all three models is presented in Figure 9. Except for the Fearful emotion, for which the overall accuracy is already low, the Ensemble model outperforms both the rest of them. Taking the basic four emotions of anger, happiness, neutral and sadness, the Ensemble model achieves almost more than 50% on all four of them individually. Given the problem of 8-emotion classification, the four basic emotions are relatively more distinct than the rest of them in the SEMOUR$^+$ dataset.

6 Discussion

In this section, we elaborate on the design and construction of our dataset and the results of the human annotation process. We also discuss the limitations of our current work in terms of dialect representation, machine-based model generalization, and speech spontaneity.

6.1 Dataset Construction Process

This section summarizes the pursuit of gathering the high-quality databases followed by the authors to produce or extend any dataset. The first and foremost step is to select a language and study relevant published work to identify possible gaps in the resources available for that language. In light of our problem statement, a phonetically balanced script was designed enriched with words, phrases, and sentences to be elicited in diverse emotions and accents. Rich vocabulary sentences in the neutral speech were available (Raza et al., 2009) but not useful in our scenario as the vocabulary employed in

these is not used in daily routine. Hence a new script had to be designed to target the spoken Urdu language composed from frequently used words. The next step was to recruit actors by advertisement, following a strict screening process based on their language proficiency and performance experience. The script was distributed among the speakers for practice before recordings. A soundproof recording room was booked for actors to record their sessions. Actors with a performance lacking in the authenticity of emotions were asked to rerecord for better elicitation. Hence, recordings of each actor's emotion were procured and clipped according to the strict order and correct pronunciation of the script's instances.

Once the sound clips were ready, 18% of the repository from the first eight actors was tagged by annotators to obtain human accuracy for SEMOUR$^+$. A user-friendly application to aid annotation was designed to achieve this goal. Annotators were asked to tag discrete categorical and continuous attributes along with the authenticity of sound clips. Extensive experimentation to address accuracy and performance measures were performed along with comparative analysis for providing a fine benchmark for further explorations on the repository. Figure 1 elaborates the aforementioned steps for procuring this dataset.

The authors would like to highlight that each step of acquisition posed diverse challenges. Instances of the script were modified continuously until a reasonable balance in phonemes and their frequencies were achieved as compared to the existing language corpus and most frequently used word list. Moreover, various re-takes were performed to ensure correct pronunciations and authentic emotion elicitation by actors. Post-processing of audio clips after successful recording sessions was indeed a tedious task. Gaps of silences were removed then clips were pruned and renamed according to instances of the script. During this process, a list of mistakes was maintained for the actors to elicit the mispronounced instances again. Experimentation on the dataset was only made possible once all the instances were correctly saved.

Lastly, gathering speakers with diverse hometowns and accents was also a laborious task. It was made specific that the speakers spent their childhood years in their respective hometowns irrespective of their current residence to ensure accent diversification. The six accents of the Urdu language were categorized with respect to provinces. Then we classified speakers with their accents according to the cities belonging to these provinces. This extension was found helpful as the speaker-independent results were improved by 17%. Thus the more samples (belonging to some particular area) we collect, the greater will be the model's performance.

6.2 Interpreting Human Subjects Study

As discussed in section 4 we used human subjects to evaluate 18% of the instances in SEMOUR$^+$. The goal of this study was to verify whether the expressions in these instances represent the corresponding emotion labels.

Among the randomly chosen subset, 78% of the instances were correctly classified by a human evaluator on average. We concluded that these 78% instances contain enough information to detect a correlation between the acted expression and the corresponding emotion label. For the remaining 22% instances, there are two possibilities. Either those instances were not uttered or recorded with the correct emotional expression, or, the instances contain the correct acted expression but it was missed by the human evaluator, and they classified it incorrectly. We designed an experiment to investigate the cause of these misclassifications by human evaluators as follows. We trained a simple Neural Network on the instances that were correctly classified by the evaluators. So, the model learned to distinguish the correct expression based on the labels of instances on which human evaluators agreed with dataset labels. We tested this Neural Network model on the remaining instances that were either acted incorrectly by actors or misclassified which human evaluators. The Neural Network achieved a test accuracy of 92% on these instances. This indicated a significant correlation between the emotion-labels and the features of respective utterances. We conclude that these correlations may have been missed by the evaluators.

There is also a concern about whether the uttered instances contain exaggerated expressions of emotions which may lead to a system that won't generalize to spontaneous speech. We asked our human subjects to tag whether a given acted expression was perceived as natural or exaggerated. Of the 9977 sample instances for which we received a vote, $> 84\%$ were reported to have a natural expression while the remaining $< 16\%$ were tagged with a fake or exaggerated expression. We conclude that most of the acted expressions in SEMOUR$^+$ successfully mimic the corresponding natural expressions in spontaneous speech.

Regarding speakers that participated in the recording, the first eight speakers had some experience in the acting, while the others with diverse accents (speaker ids: $9-24$) had no experience in acting. Studying the results reported by our best model: Ensemble, speaker with id. 9 has the lowest accuracy, i.e., 34%, and the best performance is delivered by actor id. 3 with 76% accuracy, as shown in Figure 8. Studying their individual emotions, the model predicts the boredom and sadness emotion with 83% and 99% accuracy respectively. All the other emotions have very low accuracy values. The accuracy values for anger, boredom, disgust, fearful, happiness, neutral, sadness and surprise are $78, 70, 31, 32, 66, 60, 47$ and 60 percent respectively, as shown in Figure 9. The disgust, fear and sadness emotions have low values indicating that the speakers did not quite utter these emotions. It could be the case that these emotions are not manifested in daily routine by the speakers. Apart from actors $4, 5, 9, 10$, all the other speakers have performances greater or equal than 50%. We can conclude that anger, boredom, happiness neutral and surprise can be correctly classified by our model.

6.3 Limitations

SEMOUR$^+$ is the first dataset for the Urdu Language recorded in eight complex emotions. Naturally, there are several limitations in this work that offer interesting avenues for improvement. Below, we discuss these limitations and provide future directions for this work.

Studies on speech recognition systems have noted several disadvantages to using acted speech as compared to data from spontaneous conversations in natural settings (Batliner, Fischer, Huber, Spilker, & Nöth, 2000; Douglas-Cowie et al., 2005). As the ultimate goal of a recognition system is to classify uncontrolled speech, a model trained on natural datasets is expected to generalize better. Similarly, spontaneous speech also provides a context that is lacking in the acted counterpart (Cauldwell, 2000). Despite these benefits, most of the works on speech emotion recognition, including SEMOUR$^+$, are based on datasets that use acted speech expressions as mentioned in Swain, Routray, and Kabisatpathy (2018). The general drawback in building any potential datasets of spontaneous speech is that the audio samples are unlabelled by nature and depend on either the subjective opinion of the people tagging the dataset or some machine-assisted predictions to establish the ground truth. Secondly, the collection and cleaning process of spontaneous data requires considerably more effort (Douglas-Cowie, Campbell, Cowie, & Roach, 2003). For instance, the voice quality in the natural datasets has a high variance that results in a significant loss of accuracy, as studied in Scherer (2003). Therefore, one needs more advanced methods for noise removal. As one expects, natural datasets are also highly unbalanced for emotion classes. An overwhelming majority of the instances need to be discarded because they contain redundant information that is unproductive for model training. The construction of a natural speech dataset while overcoming these challenges is an open problem. Meanwhile, there is ample evidence to suggest that acted speech expressions provide a good approximation to spontaneous speech (Jürgens, Grass, Drolet, & Fischer, 2015). Also, acted speech expressions are ideal for training for certain applications. One such potential application is to understand the extent of any inherent emotional bias towards a community in movies. Since unobserved target utterances are acted expressions, it is a good idea to have a model train on acted instances.

Urdu has four dialects and several accents that depend on the demographics of the speaker (Ghulam & Soomro, 2018). Due to a lack of resources and accessibility, utterances in SEMOUR$^+$ have been limited to a single *(Urdu)* dialect spoken by the native people of Lahore. An extension covering rich dialect representation is required in this repository. Our current work also lacks the study of the impact of demographic effects on actors' emotion elicitation as well as taggers' human annotation for the rest of the dataset. SEMOUR$^+$ lacks Sindhi accent and contains only one speaker from the Balochistan province. With the further addition of the missing accents and more speakers, there is a better chance of improving the SER accuracy as observed from the experiments.

Lastly, We report a modest 39% accuracy on the leave-one-out experiment performed on the first eight speakers and improve it to 56% by training on the full dataset (24 speakers). Training a model to predict emotions for a previously unknown speaker is, indeed, a challenging problem. For example, a CNN-BLSTM based model on the IEMOCAP dataset reports 52% accuracy for 6 emotions (Zhao et al., 2019). This also highlights a problem regarding diversity in speaker profiles. Along with utilizing a deep model, learning better speaker embeddings might improve speaker-independent accuracy.

For future direction, we plan to diversify our repositories to cover all the major dialects as well as the Sindhi accent of the Urdu Language. With this goal in mind, we invite the research community to expand the dataset further to help create a universal speech emotion recognition system for Urdu. The dataset and other relevant metadata are available at https://tinyurl.com/yc499z7k through a request form. We allow the research community to download and independently extend and improve the dataset for non-commercial research purposes. We have also devised a system where the researchers can help improve the current version of our datasets by supplementing the recorded samples in different dialects and accents. The researchers can get the dataset and the script from https://tinyurl.com/yc499z7k and have their uttered collection appended in the current version after quality checks. We also intend to expand the current version on our own to develop a more generalized speech emotion recognition system and make it available for the public on our website, where users can upload an audio clip and predict their emotions.

7 Conclusion

In this work, we provide a first of its kind, gender, and phonetically balanced, diverse in accents, scripted emotional speech datasets for the Urdu Language namely SEMOUR$^+$ recorded by 24 actors in eight complex emotions with $27,840$ unique instances worth of 14-hours collectively. A uniformly selected sample from 18% of SEMOUR$^+$ is manually tagged and validated, resulting in human accuracy of 78% and high inter-evaluator correlation scores. We also provide evaluation results of our datasets for speech emotion recognition task with variations of experiments on SEMOUR$^+$ in comparison with classical machine learning techniques. Our model performs with an average accuracy of 56% for the speaker-independent experiment when trained on the whole dataset.

References

Ali, H., Ahmad, N., Yahya, K.M., Farooq, O. (2012). A medium vocabulary urdu isolated words balanced corpus for automatic speech recognition. *2012 international conference on electronics computer technology (icect 2012)* (pp. 473–476).

Atta, F., van de Weijer, J., Zhu, L. (2020). Saraiki. *Journal of the International Phonetic Association*, 1–21.

Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *3rd international conference on learning representations, iclr 2015.* San Diego, CA, USA: Computational and Biological Learning Society.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. *Isca tutorial and research workshop (itrw) on speech and emotion.* Newcastle, Northern Ireland, UK: International Speech Communication Association.

Batliner, A., Steidl, S., Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. *Proc. of a satellite workshop of lrec* (p. 28). Marrakesh, Morocco: European Language Resources Association.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B. (2005). A database of german emotional speech. *Ninth european conference on speech communication and technology* (p. 1517-1520). Lisbon, Portugal: International Speech Communication Association.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., … Narayanan, S.S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., Provost, E.M. (2016). Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, *8*(1), 67–80.

Cai, X., Yuan, J., Zheng, R., Huang, L., Church, K. (2021). Speech Emotion Recognition with Multi-Task Learning. *Proc. interspeech 2021* (pp. 4508–4512). 10.21437/Interspeech.2021-1852

Castillo, J.C., Fernández-Caballero, A., Castro-González, Á., Salichs, M.A., López, M.T. (2014a). A framework for recognizing and regulating emotions in the elderly. L. Pecchia, L.L. Chen, C. Nugent, & J. Bravo (Eds.), *Ambient assisted living and daily activities* (pp. 320–327). Cham: Springer International Publishing.

Castillo, J.C., Fernández-Caballero, A., Castro-González, Á., Salichs, M.A., López, M.T. (2014b). A framework for recognizing and regulating emotions in the elderly. L. Pecchia, L.L. Chen, C. Nugent, & J. Bravo (Eds.), *Ambient assisted living and daily activities* (pp. 320–327). Cham: Springer International Publishing.

Cauldwell, R.T. (2000). Where did the anger go? the role of context in interpreting emotion in speech. *Isca tutorial and research workshop (itrw) on speech and emotion.* Newcastle, Northern Ireland, UK: International Speech Communication Association.

Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., … Narayanan, S. (2019). Data augmentation using gans for speech emotion recognition. *Interspeech* (pp. 171–175). Graz, Austria: International Speech Communication Association.

Chen, J., She, Y., Zheng, M., Shu, Y., Wang, Y., Xu, Y. (2019). A multimodal affective computing approach for children companion robots. *Proceedings of the seventh international symposium of chinese chi* (pp. 57–64).

Chen, M., He, X., Yang, J., Zhang, H. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440–1444.

Costantini, G., Iaderola, I., Paoloni, A., Todisco, M. (2014). Emovo corpus: an italian emotional speech database. *International conference on language resources and evaluation (lrec 2014)* (pp. 3501–3504). Reykjavik, Iceland: European Language Resources Association.

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., Schuller, B.W. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. *Proceedings of the 25th acm international conference on multimedia* (pp. 478–484).

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech communication*, *40*(1-2), 33–60.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C. (2005). Multimodal databases of everyday emotion: Facing up to complexity. *Ninth european conference on speech communication and technology* (p. 4). Lisbon, Portugal: International Speech Communication Association.

Eberhard, D.M., Simons, G.F., Fennig, C.D. (2020). *Ethnologue: Languages of the world. 23rd edn. dallas* (Vol. 23).

Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P. (1997). Design, recording and verification of a danish emotional speech database. *Fifth european conference on speech communication and technology.*

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., … others (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190–202.

Fayek, H.M., Lech, M., Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, *92*, 60–68.

Ghulam, S.M., & Soomro, T.R. (2018). Twitter and urdu. *2018 international conference on computing, mathematics and engineering technologies (icomet)* (p. 1-6). Sukkur, Pakistan: IEEE. 10.1109/ICOMET.2018 .8346370

Grimm, M., Kroschel, K., Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. *2008 ieee international conference on multimedia and expo* (pp. 865–868). Hannover, Germany: IEEE.

Han, W., Jiang, T., Li, Y., Schuller, B., Ruan, H. (2020). Ordinal learning for emotion recognition in customer service calls. *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6494–6498). Barcelona, Spain: IEEE.

Haris Bin Zia, A.A.R., & Athar, A. (2018, may). Pronouncur: An urdu pronunciation lexicon generator. *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).* Paris,

France: European Language Resources Association (ELRA).

Ijaz, M., & Hussain, S. (2007). Corpus based urdu lexicon development. *the proceedings of conference on language technology (clt07), university of peshawar, pakistan* (Vol. 73, p. 12). Pakistan: Academia.

Jackson, P., & Haq, S. (2014). *Surrey audio-visual expressed emotion (savee) database.*

Jürgens, R., Grass, A., Drolet, M., Fischer, J. (2015). Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of nonverbal behavior*, *39*(3), 195–214.

Kabir, H., & Saleem, A.M. (2002). *Speech assessment methods phonetic alphabet (sampa): Analysis of urdu.*

Koolagudi, S.G., Maity, S., Kumar, V.A., Chakrabarti, S., Rao, K.S. (2009). Iitkgp-sesc: speech database for emotion analysis. *International conference on contemporary computing* (pp. 485–492).

Koolagudi, S.G., Reddy, R., Yadav, J., Rao, K.S. (2011). Iitkgp-sehsc: Hindi speech corpus for emotion analysis. *2011 international conference on devices and communications (icdecom)* (pp. 1–5).

Kostoulas, T., Mporas, I., Ganchev, T., Fakotakis, N. (2008). The effect of emotional speech on a smart-home application. *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 305–310). Wroclaw, Poland: Springer, Berlin, Heidelberg.

Kumar, P., Kaushik, V., Raman, B. (2021). Towards the Explainability of Multimodal Speech Emotion Recognition. *Proc. interspeech 2021* (pp. 1748–1752). 10.21437/Interspeech.2021-1718

Kumawat, P., & Routray, A. (2021). Applying TDNN Architectures for Analyzing Duration Dependencies on Speech Emotion Recognition. *Proc. interspeech 2021* (pp. 3410–3414). 10.21437/Interspeech.2021-2168

Latif, S., Qayyum, A., Usman, M., Qadir, J. (2018). Cross lingual speech emotion recognition: Urdu vs. western languages. *2018 international conference on frontiers of information technology (fit)* (pp. 88–93). Islamabad, Pakistan, Pakistan: IEEE. 10.1109/FIT.2018.00023

Leem, S.-G., Fulford, D., Onnela, J.-P., Gard, D., Busso, C. (2021). Separation of Emotional and Reconstruction Embeddings on Ladder Network

to Improve Speech Emotion Recognition Robustness in Noisy Conditions. *Proc. interspeech 2021* (pp. 2871–2875). 10.21437/Interspeech .2021-1438

Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., … Chen, X. (2000). Cass: A phonetically transcribed corpus of mandarin spontaneous speech. *Sixth international conference on spoken language processing* (p. 485-488). Beijing, China: International Speech Communication Association.

Li, B., Dimitriadis, D., Stolcke, A. (2019). Acoustic and lexical sentiment analysis for customer service calls. *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 5876-5880).

Li, J.-L., & Lee, C.-C. (2019). Attentive to individual: A multimodal emotion recognition network with personalized attention profile. *Interspeech* (pp. 211–215). Graz, Austria: International Speech Communication Association.

Li, Y., Tao, J., Chao, L., Bao, W., Liu, Y. (2017). Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, *8*(6), 913–924.

Liu, J., & Wang, H. (2021). Graph Isomorphism Network for Speech Emotion Recognition. *Proc. interspeech 2021* (pp. 3405–3409). 10.21437/ Interspeech.2021-1154

Livingstone, S.R., & Russo, F.A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, *13*(5), e0196391.

Makarova, V., & Petrushin, V.A. (2002). Ruslana: A database of russian emotional utterances. *Seventh international conference on spoken language processing*.

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18–25). Austin, Texas: Academia.

Meddeb, M., Karray, H., Alimi, A.M. (2017). Building and analysing emotion corpus of the arabic speech. *2017 1st international workshop on arabic script analysis and recognition (asar)* (pp. 134–139). Nancy, France:

IEEE.   10.1109/ASAR.2017.8067775

Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Enriquez, E., Pardo, J.M. (1999). Analysis and modelling of emotional speech in spanish. *Proc. of icphs* (Vol. 2, pp. 957–960).

Moriyama, T., Mori, S., Ozawa, S. (2009). A synthesis method of emotional speech using subspace constraints in prosody. *Journal of Information Processing Society of Japan*, *50*(3), 1181–1191.

Murphy, K.P. (2012). *Machine learning: a probabilistic perspective.* MIT press.

Nezami, O.M., Lou, P.J., Karami, M. (2019). Shemo: a large-scale validated database for persian speech emotion detection. *Language Resources and Evaluation*, *53*(1), 1–16.

Oflazoglu, C., & Yildirim, S. (2013). Recognizing emotion from turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*(1), 26.

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., Hofer, G. (2019). Analysis of deep learning architectures for cross-corpus speech emotion recognition. *Interspeech* (pp. 1656–1660). Graz, Austria: International Speech Communication Association.

Qasim, M., Nawaz, S., Hussain, S., Habib, T. (2016). Urdu speech recognition system for district names of pakistan: Development, challenges and solutions. *2016 conference of the oriental chapter of international committee for coordination and standardization of speech databases and assessment techniques (o-cocosda)* (p. 28-32). Bali, Indonesia: IEEE. 10.1109/ICSDA.2016.7918979

Ramakrishnan, S., & El Emary, I.M. (2013). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, *52*(3), 1467–1478.

Raza, A.A., Athar, A., Randhawa, S., Tariq, Z., Saleem, M.B., Bin Zia, H., … Rosenfeld, R. (2018). Rapid collection of spontaneous speech corpora using telephonic community forums. *Proc. interspeech 2018* (pp. 1021–1025). Retrieved from http://dx.doi.org/10.21437/Interspeech.2018-1139   10.21437/Interspeech.2018-1139

Raza, A.A., Hussain, S., Sarfraz, H., Ullah, I., Sarfraz, Z. (2009). Design and development of phonetically rich urdu speech corpus. *2009 oriental cocosda international conference on speech database and assessments* (pp. 38–43).

Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. *2013 10th ieee international conference and workshops on automatic face and gesture recognition (fg)* (pp. 1–8). Shanghai, China: IEEE. 10.1109/FG.2013.6553805

Russell, J.A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273 - 294. Retrieved from http://www.sciencedirect.com/science/article/pii/009265667790037X

https://doi.org/10.1016/0092-6566(77)90037-X

Sager, J., Shankar, R., Reinhold, J., Venkataraman, A. (2019). Vesus: A crowd-annotated database to study emotion production and perception in spoken english. *Interspeech* (pp. 316–320). Graz, Austria: International Speech Communication Association.

Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech communication*, *54*(4), 543–565.

Santoso, J., Yamada, T., Makino, S., Ishizuka, K., Hiramura, T. (2021). Speech Emotion Recognition Based on Attention Weight Correction Using Word-Level Confidence Measure. *Proc. interspeech 2021* (pp. 1947–1951). 10.21437/Interspeech.2021-411

Sarfraz, H., Hussain, S., Bokhari, R., Raza, A.A., Ullah, I., Sarfraz, Z., … Parveen, R. (2010). Speech corpus development for a speaker independent spontaneous urdu speech recognition system. *Proceedings of the O-COCOSDA, Kathmandu, Nepal.*

Scherer, K.R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40*(1-2), 227–256.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J.K., Baird, A., … others (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. *17th annual conference of the international speech communication association (interspeech*

2016), vols 1-5 (pp. 2001–2005). San Francisco, CA, USA: International Speech Communication Association.

Scott, K.M., Ashby, S., Hanna, J. (2020). "human, all too human": Noaa weather radio and the emotional impact of synthetic voices. *Proceedings of the 2020 chi conference on human factors in computing systems* (p. 1–9). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3313831.3376338   10.1145/3313831.3376338

Sebastian, J., & Pierucci, P. (2019). Fusion techniques for utterance-level emotion recognition combining speech and transcripts. *Interspeech* (pp. 51–55). Graz, Austria: International Speech Communication Association.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stevens, S.S., Volkmann, J., Newman, E.B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190.

Swain, M., Routray, A., Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*(1), 93–120.

T, M.K., Sanchez, E., Tzimiropoulos, G., Giesbrecht, T., Valstar, M. (2021). Stochastic Process Regression for Cross-Cultural Speech Emotion Recognition. *Proc. interspeech 2021* (pp. 3390–3394).   10.21437/Interspeech.2021-610

Vashistha, A., Garg, A., Anderson, R., Raza, A.A. (2019). Threats, abuses, flirting, and blackmail: Gender inequity in social media voice forums. *Proceedings of the 2019 chi conference on human factors in computing systems* (p. 1–13). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3290605.3300302 10.1145/3290605.3300302

Walker, K., Ma, X., Graff, D., Strassel, S., Sessa, S., Jones, K. (2015). *RATS Speech Activity Detection.* Abacus Data Network. Retrieved from https://hdl.handle.net/11272.1/AB2/1UISJ7   11272.1/AB2/1UISJ7

Xu, X., Deng, J., Cummins, N., Zhang, Z., Zhao, L., Schuller, B.W. (2019). Autonomous emotion learning in speech: A view of zero-shot speech emotion recognition. *Interspeech* (pp. 949–953). Graz, Austria: International Speech Communication Association.

Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E. (2016). Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, *8*(3), 300–313.

Zhang, J.T.F.L.M., & Jia, H. (2008). Design of speech corpus for mandarin text to speech. *The blizzard challenge 2008 workshop* (p. 4). Brisbane, Australia: International Speech Communication Association.

Zhao, J., Mao, X., Chen, L. (2019). Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, *47*, 312–323.