International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v10i10.5734
***Article Received**: 28 July 2022 **Revised**: 05 September 2022 **Accepted**: 12 September 2022*

PEMO: A New Validated Dataset for Punjabi Speech Emotion Detection

Chaitanya Singla[1], Sukhdev Singh[2]
[1]Department of Computer Science,
Punjabi University, Patiala
Punjab, India
chaitanya.singla246@gmail.com
[2]Department of Computer Science,
Multani Mal Modi College, Patiala
Punjab, India
tomrdev@gmail.com

Abstract— This research work presents a new valid dataset for Punjabi called the Punjabi Emotional Speech Database (PEMO) which has been developed to assess the ability to recognize emotions in speech by both computers and humans. The PEMO includes speech samples from about 60 speakers with an age range between 20 and 45 years, for four fundamental emotions, including anger, sad, happy and neutral. In order to create the data, Punjabi films are retrieved from different multimedia websites such as YouTube. The movies are processed and transformed into utterances with software called PRAAT. The database contains 22,000 natural utterances. This is equivalent to 12 hours and 35 min of speech information taken from online Punjabi movies and web series. Three annotators categorize the emotional content of the utterances. The common label that is labelled by all annotators becomes the final label for the utterance. The annotators have a thorough knowledge of Punjabi Language. The data is used to determine the expression of emotions in speech in the Punjabi Language.

Keywords- Emotional Speech; Punjabi Speech database; emotional database; Punjabi; Emotion Detection; Emotion Recognition.

I. Introduction

Systems for detecting emotion in speech are aimed at identifying the fundamental affective condition of the speakers through its speech-related signals. They are able to be utilized for many different applications, from human-machine interaction to automated supervision as well as control over security system [1]. These systems are also utilized in the health field to monitor and detect the first indications of a depressive episode [2-3]. Another application is for criminal detection, where the state of mind of suspects who are accused of committing crimes (i.e., the degree to which they're really lying) is determined [4]. They are also used in car board systems that gather information on the state of mind of the driver to improve the safety of drivers [5]. In addition, identifying the emotional state of students within classrooms can assist instructors or even intelligent agents ensure that students receive the appropriate responses and enhance the quality of teaching as a result [6]. One important aspect to be taken into consideration prior to the development or implementing any of this speech recognition system is the accuracy of the data. The effectiveness for these devices (like any other model that uses statistics) is dependent upon the high quality of the training data [7].

However, there is often the absence of a good baseline emotional speech database for non-English languages like Punjabi. According to some studies [8-9] the connection between the content of a language and its emotion is dependent on the language of the speaker, which means translating from one language into another is typically difficult. That's why speech emotion detection systems are usually developed language-dependently. A handful of studies have examined Punjabi Speech Emotion Recognition and have introduced emotional databases [10-11]. It is the only database available in the Punjabi language that is constructed by recording sentences with various emotional states. This means that the data is developed in a controlled setting and doesn't provide superior performance than real data. Therefore, it is necessary to develop a natural database in the Punjabi language in order that it could be used in future speech emotion recognition applications. We describe an extensive validated, large-scale dataset for Punjabi known as the Punjabi Emotional Speech Database (PEMO). PEMO is a natural data set that includes emotional speech samples from various Punjabi speakers. According to the authors' knowledge it is the first comprehensive effort to create a massive, verified natural emotional speech data set that is suitable for Punjabi Language. This PEMO data will become made publicly accessible to aid research into Punjabi emotions in speech. The research has been reported in the literature that there isn't a Speech Emotion Recognition System exists for the Punjabi language. Our study focuses upon Speech Emotion Recognition System for Punjabi. To accomplish this a Punjabi emotions database is required that would include speech samples that show diverse emotions derived from different films or web series. The speech database that is presented in this article is the first to be designed for the Punjabi language which is a traditional language spoken by the Punjab State in India to study the fundamental emotions that are present in the spoken language. The database can analyze the emotion in light of gender, speaker, and the vulnerability of text. Punjabi is a regional language which means that its emotional and speech patterns are distinct from other languages. This database could be used for further research that focuses on the identification of emotions in Punjabi speech.

In Section. 2 we will review the various kinds of databases for emotional speech. We present the PEMO database and outline the data collection process as well as validation and annotation in Section. 3. In Sect. 4 We summarize our findings and offer suggestions for future direction using our data.

II. Related Work

Due to the huge number of literature about emotional speech this section will concentrate on examining different kinds of databases on emotional speech.

A. Speech Emotion Databases

Databases of emotional speech can be classified based on their naturalness, emotionality, speaker, language distribution and so on [12-14]. Naturalness is among the primary factors to be taken into consideration when creating databases. Based on the level of naturalness the database can be classified into three categories that are natural, semi-natural and simulated [15-16].

In every machine-learning task, it is necessary to have a set of training examples; SER is no distinct from other tasks. The procedure of creating a training data set for SER requires human agents to identify the samples by hand. Individuals have different perceptions of emotions. For instance, one individual may perceive the emotion as angry, while others may view it as enthusiastic. To categorize utterances, it is essential to have several agents studying each sample and having a way to choose the proper label for each sample in a consistent manner. There are three kinds of databases designed specifically to recognize speech emotions, semi-natural, simulated, as well as natural. The simulated data sets are constructed by trained speech-reading experts who read the same text using different emotions [17]. The semi-natural collection is created by asking actors or people to read a story that has various emotions. Furthermore, natural datasets are taken from television series, YouTube videos, call centers, and so on, and then categorized by listeners to human voices [17]. They are completely natural and can be utilized to build systems for recognizing emotions without worrying about their being artificially created. However, the modeling and identification of emotions using these types of datasets could be difficult because of the continuous nature of emotions as well as their dynamic changes over the course of speech, the presence of multiple emotions at once as well as there is background noise. In addition, as the sources of data were not extensive, the range of emotions that can be found in these corpora is small. Additionally, there are possible privacy and copyright issues which arise with this type of corpus. The main issue with this kind of dataset is noise reduction. Databases that were created earlier for emotional speech have a limited number of samples and actors. However, more recent datasets tend to have a huge number of samples as well as a greater variety of speakers. Table 1 presents a short review of various kinds of databases, as described in the previous paragraphs, highlighting the different characteristics of each database, as well as an example of each type.

III. Punjabi speech emotion database

PEMO is a massive natural database of Punjabi that includes 12 hours and 35 min of speech information from 60 native-Punjabi speakers. There are 22,000 utterances available in .wav format 16 bits, 44.1 KHz, and in mono that encompass four primary emotions: anger, happiness sad, and neutral. The utterances are derived from Punjabi movies that are made available publicly on a variety of multimedia websites. In the next section we describe the various steps of creating PEMO including the pre-processing of data, annotation and testing reliability.

A. Pre-processing, annotation and reliability

The web series, movies are sourced from YouTube to create the data set as shown in table 2. The videos are first converted into Audio files (in .wav file format). Then, these audio files are separated to determine the emotion of the expressions with the help of software called PRAAT.

Article Received*: 28 July 2022* ***Revised****: 05 September 2022* ***Accepted****: 12 September 2022*

TABLE I: Comparison between different types of Databases

Features	Simulated	Semi-Natural	Natural
Description	created by speakers trained to read the same text, but with different emotional states	created by asking individuals and actors read a scene with a range of emotions	Extracted from YouTube videos, Call centers, TV shows etc.
Natural Emotions		□	□
Containing Contextual Information		□	□
Containing Situational Information		□	□
Widely Used	□		
Easy to model	□		
Large Number of emotions	□	□	
Used in real world emotion system			□
Examples	EMO-DB [18] DES [19] RAVDEES [20] TESS [21]	IEMOCAP [23] Belfast [24] NIMITEK [25]	VAM [25] Call centers [28,29] AIBO [27]

TABLE II: Some Punjabi movies/ Web Series and number of clips taken from them

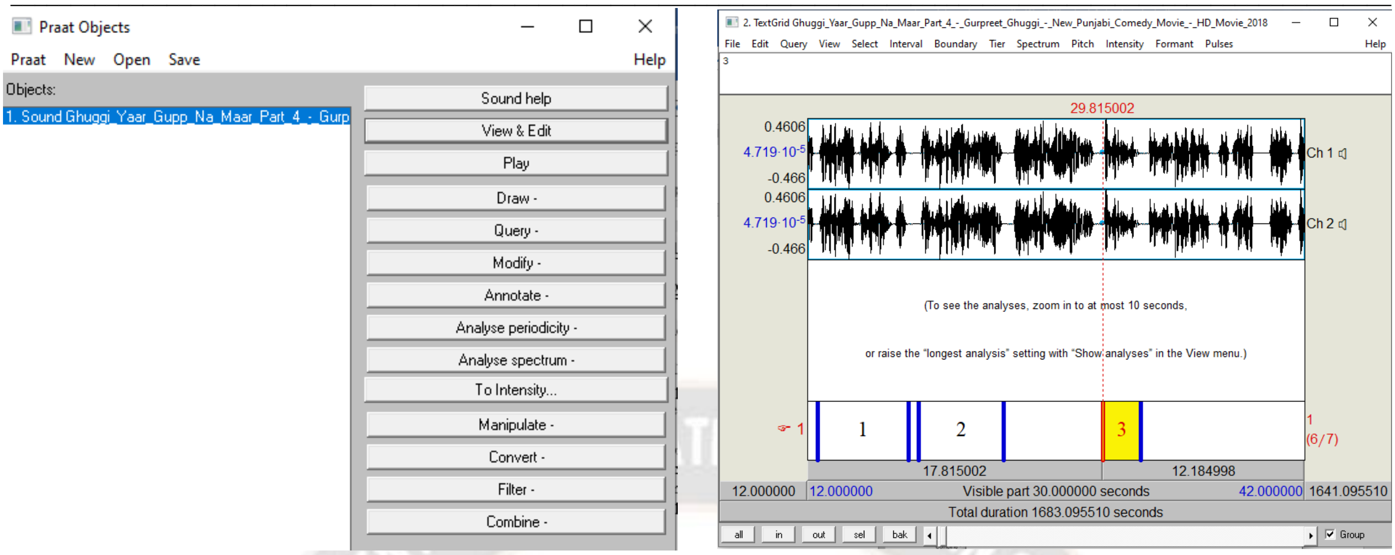
Sr. No	**Movie / Web Series Name**	**Number of Clips taken**
1	“Yaar jegree kasuti degree”	70
2	“Canada Jana hi Jana”	80
3	“Best of Gurchet Chitarkar”	215
4	"Puaada"	700
5	“Dheeth Jawaai Te 7 Salian”	340
6	“Adab Parahuna”	340
7	“Ardaas”	500
8	“Mr. and Mrs. 420”	660
9	“Rabb Da Radio”	450
10	“Carry on Jatta2”	1000
11	“Power Cut”	450
12	“Ghuggi Yaar Gupp Na Maar”	300

The uploading of the audio files and their segmentation using the PRAAT Software is shown in Figure 1.

We separated every stream in smaller segments so that each segment could be able to cover the speech sample of a one speaker, with minimal background noise or impact.

The utterances that are segmented are labelled by three annotators. The specifics of the annotators are shown in the table 3. The annotators categorize the segmented samples on a 5-point scale (including happy, angry, sad neutral, and none of the mentioned). They were native people who spoke Punjabi and had no hearing impairment or mental problems.

The utterances were played randomly in a calm environment. Because the utterances were derived from films it was not guaranteed that the words could be considered neutral emotionally. The most common label used for all annotators is the final label for the phrase. For utterances where the label isn't common to all the annotators were removed from the database as they contain several emotions that was expressed in an utterance or the emotion itself was not the one among the emotional states predefined. The neutral state is the one with the highest number of expressions, whereas sad states have only a small number of utterances in the database.



(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

a) Opening Movie/Web Series using Praat Software

b) Segmenting movie into clips

Figure 1 Annotating the movie/web series to single emotion clips using 1 a) and 1 b)

TABLE III. Annotator's information

Code	Gender	Age	Education
A1	Male	21	Undergraduate Student
A2	Female	30	PhD candidate
A3	Male	28	Master's degree

Thus, there could be certain situations where the affective state of the speaker derived in their speech could be completely different from the lexical contents of the speech. To clear up this confusion and to avoid confusion, the annotators were specifically instructed to classify the emotional content of the utterances solely based on how they had depicted it in the spoken word regardless of the content in the lexical context. The common label by all the annotators becomes the final label for that utterance. Some utterances along with its Punjabi transcript category wise is shown in table 4.

TABLE IV. Some clips along with their Punjabi transcript and Emotion

SR. NO	CLIP FILE NAME	PUNJABI TRANSCRIPT	SPEAKER AGE	SPEAKER GENDER	EMOTION CATEGORY
1	H1	ਸਮਾਈਲ ਨੂੰ ਕਿਹੜਾ ਮੈਂ ਮਨਾਹੀ ਕੀਤੀ ਹੈ ਨਾਲੇ ਗੱਲਾਂ ਬੜੀਆਂ ਮਿੱਠੀਆਂ ਕਰਦੇ ਹੋ	30	Male	Happy
2	A1	ਉਹ ਬੇੜਾ ਬਹਿ ਗਿਆ ਉਹ ਕਿਸੇ ਦਾ ਅੱਜ ਕੋਈ ਮੇਰੀ ਮੱਛੀ ਲੈ ਗਿਆ ਕੋਈ ਚੱਕ ਕੇ	27	Male	Angry
3	H2	ਫੇਰ ਤਾਂ ਤੂੰ ਮੇਰੇ ਡੈਡੀ ਨੂੰ ਵੀ ਰਿਸ਼ਤੇ ਲਈ ਮਨਾ ਲਵੇਗਾ	28	Female	Happy
4	S1	ਬਿੰਦਰ ਮੇਰੇ ਸਾਹਮਣੇ ਤਾਂ ਚੁੱਪ ਚੁੱਪ ਜੀ ਰਹਿੰਦੀ ਹੈ ਪਰ ਮੈਂ ਜਾਨ ਦਾ ਜਿੰਨੀ ਉਹ ਸ਼ਾਂਤ ਦਿਸਦੀ ਏ ਓਨੀ ਹੈ ਨੀ	40	Male	Sad
5	S2	ਮੈਨੂੰ ਨਹੀਂ ਪਤਾ ਮੇਰੇ ਮਾਂ ਬਾਪ ਕੌਣ ਨੇ ਬਸ ਇੰਨਾ ਪਤਾ ਹੈ ਕਿ ਪਾਪਾ ਜੀ ਚਿੱਠੀਆਂ ਵੰਡਣ ਲਈ ਸੀ ਤੇ ਉਨ੍ਹਾਂ ਨੂੰ ਮੈਂ ਲੱਭ ਗਿਆ ਸੀ	30	Male	Sad

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v10i10.5734
***Article Received**: 28 July 2022 **Revised**: 05 September 2022 **Accepted**: 12 September 2022*

6	S3	ਮੇਰੀ ਮੰਮੀ ਚੋਰੀ ਚੋਰੀ ਰੋਂਦੀ ਰਹਿੰਦੀ ਐ ਉਹਨੂੰ ਦੂਜੇ ਬੱਚਿਆਂ ਦੀ ਮੰਮੀ ਵਾਂਗ ਹੱਸਣ ਲਾ ਦਿਓ	18	Female	Sad
7	A2	ਚਾਚਾ ਢਕਿਆ ਰਹੇ ਇੱਥੇ ਸਾਡੀ ਪਰਸਨਲ ਗੱਲ ਚੱਲ ਰਹੀ ਹੈ	33	Male	Angry
8	H3	ਜਦੋਂ ਤੁਸੀਂ ਬਨਾਣ ਲੱਗ ਗੇ ਅਸੀਂ ਖਾਲੀ ਗਲਾਸ ਲੈ ਕੇ ਰੋਜ਼ ਤੁਹਾਡੇ ਦਰਵਾਜ਼ੇ ਤੇ ਖੜ੍ਹ ਜਾਇਆ ਕਰਾਂਗੇ	38	Male	Happy
9	N1	ਇਹ ਵਿਆਹ ਚ ਬੰਦਾ ਖਾਣ ਪੀਣ ਲਈ ਏ ਜਾਂਦਾ ਹੁੰਦਾ	25	Male	Neutral
10	N2	ਜੀਜਾ ਚਾਹ ਪੀ ਸਫਰ ਚੋਂ ਆਇਆ ਹੈਂ ਤੂੰ	28	Female	Neutral
11	A3	ਉਹ ਹੈ ਨੀ ਕੱਲ੍ਹ ਸਕੂਟਰ ਚੁੱਕਿਆ ਗਿਆ ਉਹ ਦੇਖ ਲਓ ਅੱਜ ਸਾਈਕਲ ਚੁੱਕਿਆ ਗਿਆ ਉਹ ਮੈਂ ਲੁੱਟਿਆ ਗਿਆ ਓਏ	32	Male	Angry
12	A4	ਤੂੰ ਐਂ ਕਰ ਇਹਨੂੰ ਇਹ ਵੀ ਦੱਸਦੇ ਕਿ ਮੇਰੇ ਮੁੰਡੇ ਦਾ ਰਿਸ਼ਤਾ ਟੁੱਟ ਗਿਆ ਜੇ ਤੇਰਾ ਫੇਰ ਵੀ ਨ੍ਹੀਂ ਸਰਦਾ ਫਿਰ ਗਲ ਵਿੱਚ ਢੋਲ ਪਾ ਲੈ	38	Male	Angry
13	S4	ਸੁੱਖੀ ਮੁੱਕ ਗਿਆ ਹੋਵੇ ਤਾਂ ਰੱਬ ਦਾ ਭਾਣਾ ਐ ਪਰ ਧੋਖਾ ਬਰਦਾਸ਼ਤ ਨੀ ਹੋਣਾ ਮੇਰੀ ਧੀ ਤੋਂ	36	Male	Sad

After completing the labeling process, the dataset is then verified by some Punjabi Language known persons. To increase the accuracy of the dataset, a fully validated dataset must be needed. After verifying the dataset, the recognition rate for the various states is shown in the Figure 2. The main motive of this verification is performance of the dataset as this dataset can be further used in various applications.

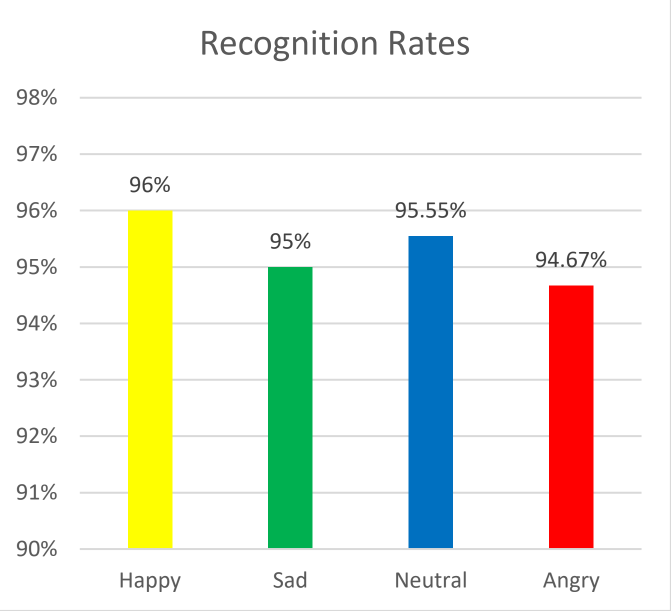


Figure 2 Recognition Rates for dataset validation

IV. Conclusion and Future Work

The purpose of this research was to develop the database of emotions for Punjabi language, in the form of sound recordings, so as to include real emotion. The next step is increasing the frequency of utterances for sadness. We are also planning to expand the test results to incorporate other techniques for classification like deep neural networks, which is the latest technique for Speech emotion identification. The labeling of the data according to emotion and valence is a possible future enhancement. In the near future we will also be able to note the strength of the emotional impact of the speech. We are also working to increase the number of utterances that are in line with more emotions.

References

[1] Huahu, Xu, Gao Jue, and Yuan Jian. "Application of speech emotion recognition in intelligent household robot." In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, pp. 537-541. IEEE, 2010.

[2] Dickerson, Robert F., Eugenia I. Gorlin, and John A. Stankovic. "Empath: a continuous remote emotional health monitoring system for depressive illness." In *Proceedings of the 2nd Conference on Wireless Health*, pp. 1-10. 2011.

[3] Heni, Nazih, and Habib Hamam. "Design of emotional educational system mobile games for autistic children." In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 631-637. IEEE, 2016.

[4] Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor. "Emotion recognition in human-computer interaction." *IEEE Signal processing magazine* 18, no. 1 (2001): 32-80.

[5] Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." In *2004 IEEE international conference on acoustics, speech, and signal processing*, vol. 1, pp. I-577. IEEE, 2004.

[6] Kort, Barry, Rob Reilly, and Rosalind W. Picard. "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion." In *Proceedings IEEE international conference on advanced learning technologies*, pp. 43-46. IEEE, 2001.

[7] Nuria Rabanal, & Prof. Dharmesh Dhabliya. (2022). Designing Architecture of Embedded System Design using HDL Method. Acta Energetica, (02), 52–58. Retrieved from http://actaenergetica.org/index.php/journal/article/view/469

[8] Busso, Carlos, Murtaza Bulut, Shrikanth Narayanan, J. Gratch, and S. Marsella. "Toward effective automatic recognition systems of emotion in speech." *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds* (2013): 110-127.

[9] Feraru, Silvia Monica, and Dagmar Schuller. "Cross-language acoustic emotion recognition: An overview and some tendencies." In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 125-131. IEEE, 2015.

[10] Sagha, Hesam, Pavel Matejka, Maryna Gavryukova, Filip Povolný, Erik Marchi, and Björn Schuller. "Enhancing multilingual recognition of emotion in speech by language identification." (2016).

[11] Anatoliy Goncharuk, & Ahmed F. Mohamed. (2022). Analytical Approach of Laboratory for the Microelectronics Fabrication. Acta Energetica, (02), 08–14. Retrieved from http://actaenergetica.org/index.php/journal/article/view/463

[12] Kaur, Kamaldeep, and Parminder Singh. "Punjabi emotional speech database: design, recording and verification." *International Journal of Intelligent Systems and Applications in Engineering* 9, no. 4 (2021): 205-208.

[13] Singla, Chaitanya, and Sukhdev Singh. "Punjabi Speech Emotion Recognition using Prosodic, Spectral and Wavelet Features." In *2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22)*, pp. 1-6. IEEE, 2022.

[14] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern recognition* 44, no. 3 (2011): 572-587.

[15] Singla, Chaitanya, Sukhdev Singh, and Monika Pathak. "Automatic Audio Based Emotion Recognition System: Scope and Challenges." In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*. 2020.

[16] Singla, Chaitanya, and Sukhdev Singh. "Databases, Classifiers for Speech Emotion Recognition: A Review." (2019).

[17] Fahad, Md Shah, Ashish Ranjan, Jainath Yadav, and Akshay Deepak. "A survey of speech emotion recognition in natural environment." Digital Signal Processing 110 (2021): 102951.

[18] Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." *Sensors* 21, no. 4 (2021): 1249.

[19] Douglas-Cowie, Ellen, Roddy Cowie, and Marc Schröder. "A new emotion database: considerations, sources and scope." In *ISCA tutorial and research workshop (ITRW) on speech and emotion*. 2000.

[20] Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. "A database of German emotional speech." In *Interspeech*, vol. 5, pp. 1517-1520. 2005.

[21] Engberg, Inger S., Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard. "Design, recording and verification of a Danish emotional speech database." In *Fifth European conference on speech communication and technology*. 1997.

[22] Livingstone, Steven R., and Frank A. Russo. "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, no. 5 (2018): e0196391.

[23] Dupuis, Kate, and M. Kathleen Pichora-Fuller. "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set." *Canadian Acoustics* 39, no. 3 (2011): 182-183.

[24] Cao, Houwei, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. "Crema-d: Crowd-sourced emotional multimodal actors dataset." *IEEE transactions on affective computing* 5, no. 4 (2014): 377-390.

[25] Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation* 42, no. 4 (2008): 335-359.

[26] Sneddon, Ian, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. "The belfast induced natural emotion database." *IEEE Transactions on Affective Computing* 3, no. 1 (2011): 32-41.

[27] Gnjatović, Milan, and Dietmar Rösner. "Inducing genuine emotions in simulated speech-based human-machine

interaction: The nimitek corpus." *IEEE Transactions on Affective Computing* 1, no. 2 (2010): 132-144.

[28] Grimm, Michael, Kristian Kroschel, and Shrikanth Narayanan. "The Vera am Mittag German audio-visual emotional speech database." In *2008 IEEE international conference on multimedia and expo*, pp. 865-868. IEEE, 2008.

[29] Steidl, Stefan. Automatic classification of emotion related user states in spontaneous children's speech. Berlin, Germany: Logos-Verlag, 2009.

[30] Morrison, Donn, Ruili Wang, and Liyanage C. De Silva. "Ensemble methods for spoken emotion recognition in call-centres." *Speech communication* 49, no. 2 (2007): 98-112.

[31] Lee, Chul Min, and Shrikanth S. Narayanan. "Toward detecting emotions in spoken dialogs." *IEEE transactions on speech and audio processing* 13, no. 2 (2005): 293-303.