Current State of Speech Emotion Dataset-National and International Level

Surbhi Khurana[1(✉)], Amita Dev[1], and Poonam Bansal[2]

[1] Indira Gandhi Delhi Technical University for Women, Delhi, India
{surbhi001phd20,vc}@igdtuw.ac.in
[2] Maharaja Surajmal Institute of Technology, GGSIPU, Delhi, India

Abstract. Research on emotion extraction from human speech is transitioning from a phase of exploratory study to one with the potential for significant applications, particularly in human–computer interaction. To achieve more accuracy while creating human-computer interaction, the computer system must be provided with good quality data that covers every aspect required for the interaction. The establishment of relevant databases is critical to progress in this area. This research will discuss the scope, naturalness, context, and descriptors of the dataset as the four primary challenges that must be taken care of while constructing databases for emotion embedded speech. Furthermore, the current state of the art is examined to get the status of available datasets for internally spoken languages like English, Dutch, French, and Chinese etc. and for Indian Spoken languages.

Keywords: Speech emotion recognition · Datasets · Context · Scope · Naturalness · International datasets · Indian datasets

1 Introduction

Human speech, par lingual, non verbal and facial expression are mainly the primary ways in which individuals communicate and connect with one another, with speech being the most effective way for exchanging information and thoughts. Speech is a multi-dimensional communication that includes language, speaker, emotion, and message. Understanding the effect of various emotions embedded in speech is critical as the presence of emotions makes communication more usual. If true emotion is included in a speech, human-robot contact can be improved, made more effective, and more natural. Finally, this aids in the artificial intelligence field.

Lie detection systems [1], audio/video retrieval [2], emotional bots , prioritization of customers calls, improvisation of medical tools, intelligent e-learning systems, language translation [3, 4], intelligent virtual games, smart cars, and categorization of voicemail are all examples of how emotion recognition can be utilised to impact human life. Emotion recognition from speech is a great study issue in the realm of voice processing because of these applications.

As demonstrated in Fig. 1, having a voice dataset is critical in the process of recognition emotions from speech. A lot of work has been done on dataset development at both

A. Dev et al. (Eds.): AIST 2021, CCIS 1546, pp. 232–243, 2022.
https://doi.org/10.1007/978-3-030-95711-7_21

the international and national levels. Speech corpora have been generated in a variety of languages, including Chinese, English, Italian, Japanese, Russian, and German. For official Indian languages such as Hindi, Telugu, Assamese, Gujarati, and Malayalam, there are just a few speech databases available. The idea of this swot is to summarize the emotional dataset requirement, as well as development challenges and a comparative analysis of the available dataset for SER systems.

It's crucial to understand the quality of the speech corpus that has been prepared, thus its analysis is on top priority [1]. Speech features are retrieved and analysed as detailed later. At the excitation source parameter, vocal tract structure, and linguistic levels, emotion-specific information is constantly present. Every emotion have its distinct impact on human speech, which can be noticed by evaluating numerous parameters/features such as MFCC [5], pitch, energy, and so on [2]. The structure of this paper is as follows: Sect. 1 states introduction, followed by Sect. 2 for the need of dataset, Sect. 3 describing the pillar for dataset, Sect. 4 for Dataset status at national and international level followed by Issues and then conclusion.

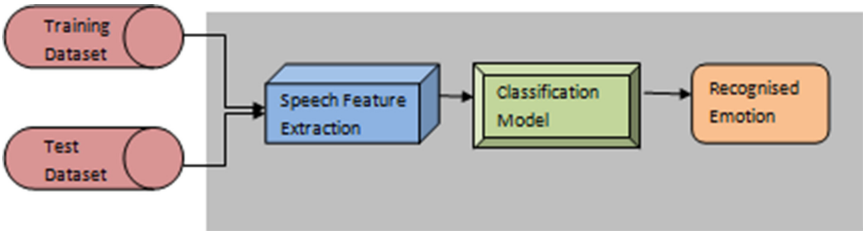

Fig. 1. Structure of SER model

2 Need of Dataset

As depicted in Fig. 1, it can be clearly noted that the emotion recognition from speech is dependent vastly on classification model, extracted speech features and moreover on the selection of good quality database. A sufficient emotional speech database is a requirement of any SER system. It is necessary to dig out speech features from the supplied dataset. Appropriate selection of features is critical since it conveys desired information and determines the system's overall efficiency.

In most cases, three types of features are taken from the database. 1) LP residual, glottal excitation signal, and other properties of the excitation source 2) Features of vocal tracks such as MFCC and LPCC 3) prosodic characteristics such as pitch and formants 4) Features combining above [6].

Extracted features are used to train various classifiers such as Gaussian mixture model and Hidden Markov model, Machine Learning and Deep Learning which will determine the unique mood. Choosing a classifier for process is usually dependent upon experimental outcome. Linear classifiers (Naive Bayes classifier) and nonlinear classifiers are the HMM, GMM, deep learning [7] based classifier.

3 Pillar of Dataset

Research based on speech embedded with emotion is transitioning from a phase of exploratory study to one with the potential for significant applications, particularly in human–computer interaction. When creating a database, four primary pillars must be considered: the scope, naturalness, context and the types of descriptors used. This section explains the above mentioned term along with its significance and related issues while creating a speech emotion recognition corpus.

3.1 Scope

Scope in database can be used to state the diversification of dataset. It includes the number of different speakers in a dataset, as well as the language spoken by the speaker, the speaker's native language, different type of emotional states, and the social/environmental setting. Moreover, not only on the language spoken it is also dependent upon the actor's gender, actor's age. Any attempt to generalize the scope of dataset could hamper the kind of diversity.

According to the assessment, [8] some feature characteristics are very consistent throughout the investigation, while others varied. The results for happy and anger emotion appear to be consistent. However, most of the other emotions and related features that have been examined at all contain discrepancies. Sad emotion is often associated with a drop in mean F0, but there are exceptions. Fear is commonly associated with a hike in F0 value along with speech rate, however there is conflicting evidence for both variables (Figs. 2, 3, 4, 5, 6 and 7).

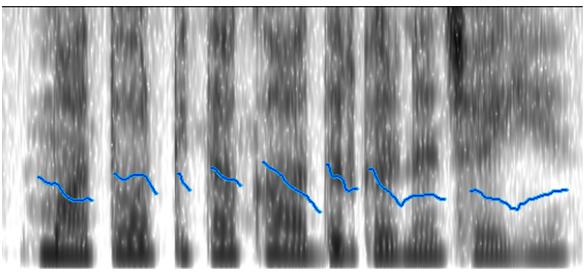

Fig. 2. F0 frequency for sad emotion-206.2 Hz

Some variances could simply be due to changes in process or misleading emotion categories, or among natural, elicitated and simulated data. Moreover, on the other hand, remaining parameters appear to reflect true changes in emotional vocal expression for eg- it varied for different speakers, from culture to culture, age to age, gender to gender and context to situation. Although there are few comparisons between languages and civilizations, they do reveal significant disparities.

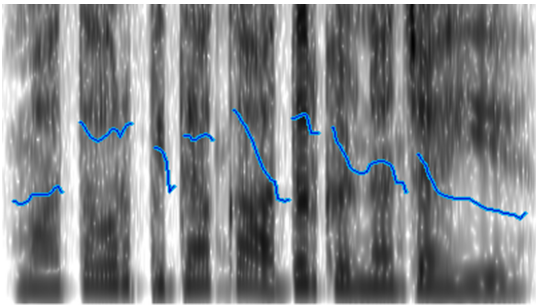

Fig. 3. F0 frequency for fear emotion-269.5 Hz

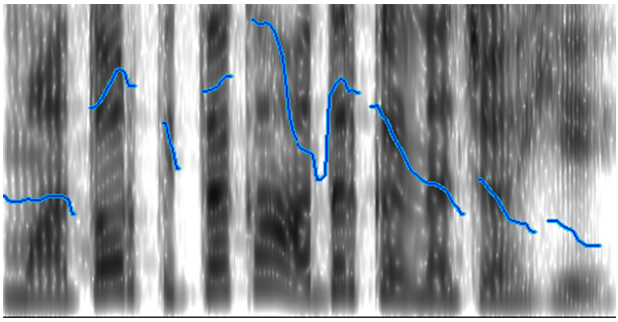

Fig. 4. F0 frequency for anger emotion-294.9 Hz

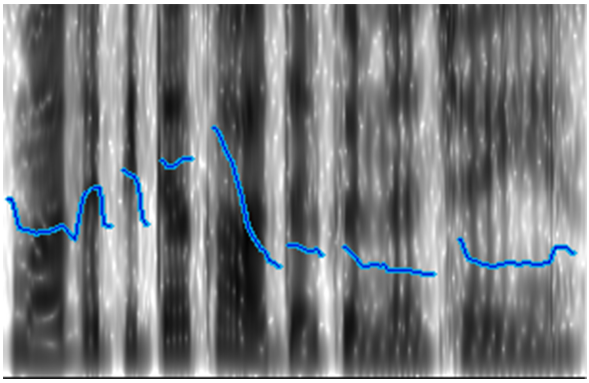

Fig. 5. F0 frequency for neutral emotion-214.6 Hz

Because speaking is primarily a cultural activity, signals of emotion embedded speech may be susceptible to cultural influences. Moreover, instead of expressing only basic emotions with maximum intensity, speech in everyday life tends to communicate intermediate emotional states. These observations show that the emotional breadth

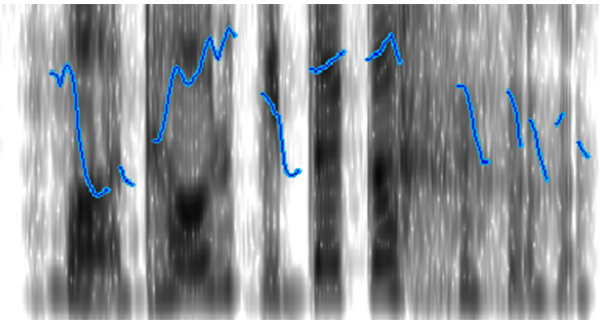

Fig. 6. F0 frequency for happy emotion-290.5 Hz

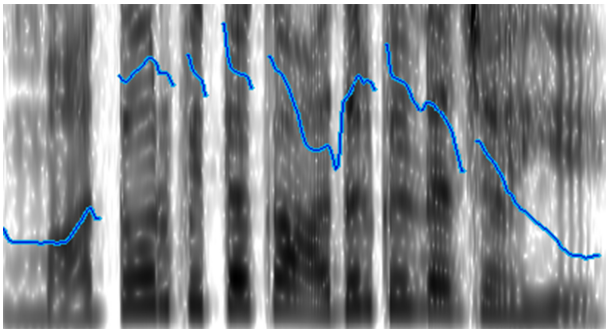

Fig. 7. F0 frequency for surprise emotion-308.8 Hz

of databases should be carefully considered. Because conventional lists of (non-basic) emotions comprise over a hundred words, the scope may need to be somewhat broad.

3.2 Naturalness

Having actors imitate emotional speech is the simplest approach to collect it. The problem with this method is that there is a surprising lack of emotional essence on the connection among performed data and spontaneous, ordinary emotional speech data. It is undeniably accurate that talented performers can produce speech that audiences can dependably categorize.

Acted speech audio is frequently not spoken instead actors read it only, and it is well known that read speech has specific qualities. Moreover, the traditional format is non-interactive, interpersonal consequences are yet not taken into account. Because the context is usually sparse, the spoken sentences do not show the vocal indicators of emotion development and how it fades over time, or how they are related to other parameter of signal.

Naturalness comes at a cost: a lack of control. Emotions are unpredictable, which makes collecting voice samples of speakers in a desired state, either elicitated or spontaneous or natural, problematic. Recognizing the expressed emotion becomes a significant

difficulty, especially if it is spontaneous. Some applications require data sets that are phonetically balanced, and its complex to envisage obtaining such balance with really natural speech. Bootstrapping, or using other pre verified content to direct is that is genuinely similar to nature, could be a long-term answer to such concerns.

3.3 Context

People use the context of words to establish the emotional meaning of voice qualities, according to direct evidence. As a result, datasets containing knowledge on the way vocal indicators connect to their surroundings are required if model wants to recognize or match human performance. There are three different sorts of contexts.

Semantic context: Emotionally charged words are more likely to appear in genuine emotional discourse. There is a clear possibility of content and vocal indicators interacting.

Structural context: Many indications of emotion, like as stress patterns and default intonation patterns, appear to be defined by syntactic structures. The hypothesis that emotion is signalled through differences in style that are conveyed in structural aspects of the spoken speech utterances is less well-known (length and duration of spoken phrases, total repetitions and gaps or interruptions in samples, etc).

Intermodal context: Humans transmit a great set of emotions over the phone demonstrates that the study based solely on voice is feasible. Speech, on the other hand, is frequently used as a supporting parameter to other sources of emotional information instead of as a standalone feature.

Descriptors and Accessibility.

Building a dataset necessitates the use of tools for describing the emotional and linguistic content of the material along with spoken speech samples on the other end.

Two difficulties stand out in terms of speech descriptors. To begin, coding must take into account the complete set of elements involved in vocal expressions of emotion, that includes supra segmental features like prosody, speech quality and non-linguistic features. Second, it must define the characteristics that are associated with emotion. The decision between continuous variables and categorical descriptors is crucial. The advantages of these different types have yet to be determined. Additional forms of labels (e.g. facial, age, gender, gestural) may be required if databases are multi-modal.

When a database is accessible to the entire speech community, it eliminates the need for duplication of effort, allows algorithms to be compared on the same data, and so on. Format and ethics are two major factors that influence availability.

The data files must be in a standardize and portable format. This is not only necessary for raw material coding formats (such as wav), but also to descriptor coding.

Moreover, copyright and ethical problems, particularly with innate data, are more fundamental. Natural emotional content is frequently highly personal, and speakers may oppose to its widespread dissemination. Radio, you tube, television, conversation shows, documentaries, and other programming, provide vast sources, but accessing them poses major copyright issues. It is evident that creating and characterizing datasets of the type that fit the needs we define is difficult.

4 Dataset Status at National and International Level

This section tries to summarise the current state of emotional speech datasets. Three of the pillars discussed above scope, naturalness and context are used to define different datasets. A generic name is used to identify a dataset. The scope of the project includes a variety of topics, emotions to examine, and the language used that indicate the cultural diversity of the dataset. Naturalness of the dataset is divided into three categories: simulated, elicitated, and natural; whether the content is scripted or not; and content structure (words, sentences or numbers). Elicitated refers to a range of strategies that produce results that fall somewhere between simulation and naturalness. Study also tries to look for any attempt to address the topic of emotional development and change through time, as well as whether the data is audio or multimodal i.e. audio–visual in nature.

Table 1 and Table 2 are used to summarise the status of dataset for International languages and Indic languages resp.

Table 1. Analysis of available speech emotion dataset for internationally spoken languages.

Identifier	Scope			Naturalness			Context
	Subject	Emotion	Spoken language	Dataset type	Scripted dataset	Linguistic structure	Modality
Danish emotional database [9]	4	Anger, Happiness, Neutral, Sadness, Surprise	Danish	Simulated	✔	2 words, 9 sentences and 2 passages	Audio
EMO-DB [10]	10	Anger, Boredom, Disgust, Fear, Happiness, Neutral, Sadness	German	Simulated	✔	500 Utterances	Audio
SAVEE	4	Anger, Boredom, Disgust, Fear, Happiness, Neutral, Surprise	English	Simulated	✔	480 Utterances	Audio-Video
RECOLA [11]	46	Arousal, Agreement, Dominance, Engagement, Performance, Rapport, Valence	French	Natural		3.8/2.9 h Of annotated audio visual/multimodal data	Audio-Visual
SAMAINE	150	Activation, Expectations, Power, Valence	English, Greek	Natural		959 Conversation	Audio

(*continued*)

Table 1. (*continued*)

Identifier	Scope			Naturalness			Context
	Subject	Emotion	Spoken language	Dataset type	Scripted dataset	Linguistic structure	Modality
eNTERFACE'05 [12]	42	Anger, Disgust, Fear, Happiness, Sadness, Surprise	English, Spanish, and French	Elicited	✔	English utterances-186, Spanish utterances-190, French utterances-175	Audio-Visual
IEMOCAP [13]	10	Anger, Frustration, Happiness, Neutral, Sadness	English	Elicited	✔	5 Sessions with conversation between male and female speaker	Audio-Visual
FAU AIBO [14]	51 German childrens and 30 English childrens	Anger, Boredom, Emphatic, Helpless, Joy, Motherese, Neutral, Reprimanding, Rest, Surprised, Touchy	English, German	Natural		51,393 words in German 5,822 words in English	Audio
BAUM speech dataset [15]	31	Anger, Boredom, Bothered, Contempt, Concentration, Disgust, Fear, Happiness, Interested, Surprise, Sadness, Thoughtful	Turkish	Acted and Natural		1184 Video clips	Audio-Visual
Chinese speech dataset [16]	8	Anger, Fear, Neutral, Happy, Sadness	Chinese	Simulated	✔	2400 Utterances	
Emotional speech database for corpus based synthesis [17]	2	Anger, Disgust, Fear, happiness, Sadness, Surprise, Neutral	Basque	Elicited	✔	702 Sentences per emotion	Audio
TV series ally McBeal [18]	6	Cold anger, Fear, Hot anger, Happy, Sadness, Neutral	English, German, Japanese	Simulated	✔	135 Utterances (45 utterances per language)	Audio–Visual

Table 2. Analysis of available speech emotion dataset for nationally spoken languages.

Identifier	Scope			Naturalness		Context	
	Subject	Emotion	Spoken language	Dataset type	Scripted dataset	Linguistic structure	Mode
IITKGP-SEHSC [19]	10	Anger, Disgust, Fear, Happy, Neutral, Sad, Sarcastic and surprise	Hindi	Simulated	✔	12000 (15 text prompts × 8 emotions × 10 speakers × 10 sessions)	Audio
Malayalam language speech emotional dataset [20]	16	Neutral, Happy, Sad and Anger	Malayalam	Acted	✔	20 sentences	Audio
SUST Bangla emotional speech corpus [21]	20	Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise	Bangla	Acted	✔	7000 Utterances (20 speakers × 10 sentences × 5 repetitions × 7 emotions)	Audio
IIIT-H TEMD [22]	19 speakers (12 female and 7 male) Non Actors-19 speakers (8 female and 11 male)	Anger, Happiness, Sadness, Neutral, and Surprise	Telugu	Natural-Drama speech	✔	5317 Annotated utterances	Audio
Gujarati speech emotional dataset [23]	9 (6 male and 3 Female)	Anger, Sadness, Surprise, Disgust, Fear, Happiness	Gujarati	Acted	✔	24 Different words	Audio
Emotional Hindi speech database [24]	28	Neutral, Happy, Anger, Sad, Sarcastic, Surprise	Hindi	Simulated	✔	6048 Utterances (28 speakers X6 emotions X12 statements X3 repetitions)	Audio

5 Issues

Database plays a vital role while working with any model based upon machine learning and deep learning algorithms. Model parameters like accuracy, precision are moreover, depends upon the quality of dataset used. However, if the model is used for detecting the presence of any specific emotion, the role of database increases drastically. Even not only upon the quality, model result will also vary on the speaker age group, gender of the speaker and the language along with their speaking style. However, during the database review following are the list of issue that researches have faced during database creation.

- The majority of the study on emotion embedded with speech is only supported by datasets instead of on databases. Term dataset are often refers to small compilations of content created to investigate a certain problem and sometime are not publicly available.
- Some of the speech dependent applications require data sets that are phonetically balanced [25], and hence it is challenging to envisage obtaining such balance with really natural speech.
- Naturalness in the database comes at a cost of lack of control. As, emotion is unpredictable, which makes collecting speech samples of actors in a desired state, problematic. Naturalness may interact with the need for proper labeling of emotional content. Acted content can be appropriately defined using category labels like sad, furious, glad, and so on.
- Some variances could simply be due to changes in process or misinterpretation of emotional grouping, or between real, elicitated or simulated data.
- Quality of database varies from speaker to speaker, their culture, speaker gender, age and context to situation, this diversity must be taken care of to reflect the true changes in emotional vocal expression.

6 Conclusion

The aim of the review is to understand the need and significance of dataset in determining emotions accurately using speaker speech samples. This paper emphasis on a comprehensive analysis of notably available speech emotion datasets for internationally spoken languages as well for nationally spoken Indic languages. The dataset used by SER systems must covers all the four aspects of scope, naturalness, context, and accessibility. However, there are still many issues that need to be solved while designing and developing SER dataset. Although much has been done to address the major challenges, there is still much more to be done. Various methods have been created, which address the issues that arise and point to future paths for the development of emotional databases.

References

1. Rao, K.S., Shashidhar, G.K.: Emotion Recognition using Speech Features. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-5143-3
2. Rachman, F.H., Sarno, R., Fatichah, C.: Music emotion classification based on lyrics-audio using corpus based emotion. Int. J. Electr. Comput. Eng. (IJECE) **8**(3), 1720 (2018)
3. Kumari, R., Dev, A., Kumar, A.: An efficient adaptive artificial neural network based text to speech synthesizer for Hindi language. Multimedia Tools Appl. **80**(16), 24669–24695 (2021). https://doi.org/10.1007/s11042-021-10771-w
4. Bhatt, S., Jain, A., Dev, A.: Continuous speech recognition technologies—a review. In: Singh, M., Rafat, Y. (eds.) Recent Developments in Acoustics. LNME, pp. 85–94. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-5776-7_8
5. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. **21**(1), 93–120 (2018). https://doi.org/10.1007/s10772-018-9491-z

6. Koolagudi, S.K.: Recognition of emotions from speech using excitation source features. Procedia Eng. **38**, 3409–3417 (2012)
7. Bhatt, S., Jain, A., Dev, A.: Feature extraction techniques with analysis of confusing words for speech recognition in the Hindi language. Wirel. Pers. Commun. **118**(4), 3303–3333 (2021). https://doi.org/10.1007/s11277-021-08181-0
8. Cowie, R., et al.: Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. 18, 32–80 (2001)
9. Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P.: Design, recording and verification of a Danish emotional speech database, pp. 1–4 (1997)
10. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech (2005)
11. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interaction (2013)
12. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE' 05 audio-visual emotion database. In: IEEE Conference Publication, no. 1, pp. 2–9 (2019). https://ieeexplore.ieee.org/abstract/document/1623803
13. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6
14. Batliner, A., et al.: You stupid tin box – children interacting with the AIBO robot: a cross-linguistic emotional speech corpus (2004)
15. Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E.: BAUM-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans. Affect. Comput. **8**(3), 300–313 (2017). https://doi.org/10.1109/TAFFC.2016.2553038
16. Zhang, S., Ching, P., Kong, F.: Automatic recognition of speech signal in Mandarin (2006)
17. Saratxaga, I., Navas, E., Hernáez, I., Luengo, I.: Designing and recording an emotional speech database for corpus based synthesis in Basque. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, pp. 2126–2129 (2006)
18. Braun, A., Katerbow, M.: Emotions in dubbed speech: an intercultural approach with respect to F0. In: 9th European Conference on Speech Communication and Technology, pp. 521–524 (2005). https://doi.org/10.21437/interspeech.2005-331
19. Koolagudi, S.G., Reddy, R., Yadav, J., Rao, K.S.: IITKGP-SEHSC : Hindi speech corpus for emotion analysis. In: 2011 International Conference on Devices and Communications, ICDeCom 2011 - Proceedings, pp. 1–5 (2011). https://doi.org/10.1109/ICDECOM.2011.5738540
20. Rajisha, T.M., Sunija, A.P., Riyas, K.S.: Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM. Procedia Technol. **24**, 1097–1104 (2016). https://doi.org/10.1016/j.protcy.2016.05.242
21. Sultana, S., Rahman, M.S., Selim, M. R., Iqbal, M.Z.: SUST Bangla emotional speech corpus (SUBESCO): an audio-only emotional speech corpus for Bangla. PLoS One **16**(4) 1–27 (2021). https://doi.org/10.1371/journal.pone.0250173
22. Rambabu, B., Kumar, B.K., Gangamohan, P., Gangashetty, S.V.: IIIT-H TEMD semi-natural emotional speech database from professional actors and non-actors. In: LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, pp. 1538–1545, May 2020
23. Tank, V.P., Hadia, S.K.: Creation of speech corpus for emotion analysis in Gujarati language and its evaluation by various speech parameters. Int. J. Electr. Comput. Eng. **10**(5), 4752–4758 (2020). https://doi.org/10.11591/ijece.v10i5.pp4752-4758

24. Bansal, S., Dev, A.: Emotional Hindi speech database. In: 2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013, pp. 5–8 (2013). https://doi.org/10.1109/ICSDA.2013.6709867
25. Kumari, R., Dev, A., Kumar, A.: Automatic segmentation of Hindi speech into syllable-like units. Int. J. Adv. Comput. Sci. Appl. **11**(5), 400–406 (2020). https://doi.org/10.14569/IJACSA.2020.0110553