International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Punjabi Emotional Speech Database:Design, Recording and Verification

Kamaldeep Kaur*[1], Parminder Singh[2]

Submitted: 31/08/2021 Accepted : 06/11/2021

Abstract: This paper introduces Punjabi Emotional Speech Database that has been created to evaluate the recognition of emotions in speech, by the humans and the computer system. The database has been designed, recorded and verified using various standards. The results set a standard for identifying emotions from Punjabi speech. Six emotions are simulated for the collection of speech corpus, including happy, sad, fear, anger, neutral and surprise. 15 speakers, with age group 20-45 years have participated in the recordings for this database. Finally, this database has been used to further design and develop the speech emotion recognition system for Punjabi language.

Keywords: *Emotions, Punjabi, Speaker, Recording, Evaluation, Emotion Recognition*

1. Introduction

Human emotions play an active role in extracting the inner plight of speaker's mind. They are the responses made towards any event happening internally or externally. The motive against the statement changes as the emotion changes. It becomes deceptive for humans also to judge the sentiments of the person in some situations. And this phenomenon can be used in Speech Emotion Recognition (SER) process [1][2]. Having a large number of real life applications, SER has been focus of research since 1970s [3]. The database creation is the first and one of the important and challenging steps in SER. The audio recordings of spoken language samples, are collected to form a speech corpus. The variations can be reflected due to differences in need, recording environment, speakers, device, etc. [4].

There are three standard methods of database collection, which classify the databases into three types, namely *actor* (simulated), *elicited* (induced) and *natural.* The simulated are collected from experienced professionals such as theatre or radio artists, where they express some pre-defined sentences in desired emotions. The variation in the degree of expressiveness can be considered by doing recording in different sessions, which would also address the variation in mechanism of speech production in different human beings. More than 60% of the databases collected are of this kind. The simulated emotions are considered as a good approach, in terms of the ability to express the emotions well. The elicited are composed, without knowledge of the speaker, by simulating artificial emotion situation. The speaker is taken into such environment/ situation where he/she would express the emotions needed for recording, but the speaker is unaware of the fact that he/she is being recorded. The elicited database is more natural than the simulated, but the speakers may not be able to express the emotions well. Natural emotions are not recorded under any simulated or induced environment, but the recordings from conversations of call centers, cockpit recordings of airplanes, talk between patient and doctor, etc. can be used for this type of database [5][6][7].

Databases for many foreign and Indian languages exist, namely SAVEE [8] and LDC [9] for English, EMODB [10] for German, SES [11] for Spanish, MES [12] for Mandarin, CASIA [13] for Chinese and DES [14] for Danish among foreign languages and IITKGP-SEHSC [15] for Hindi and IITKGP-SESC [16] for Telugu among Indian languages respectively.

It has been reviewed in literature that no Speech Emotion Recognition System exists for Punjabi language. Our research focuses on Speech Emotion Recognition System for Punjabi. For that purpose, a Punjabi emotion database is needed which would incorporate recordings in different emotions from few speakers. And no such database exists for Punjabi language. That is why, an additional Punjabi emotion database is required. The speech database presented in this paper is the first one created for Punjabi language, which is a native language of Punjab state in India, for investigating the basic emotions, which exist with the spoken material. This database is adequate to examine the emotions in view of gender, speaker and text vulnerability. As Punjabi is a regional tonal language, so its speech and emotion patterns are different from other languages. This reason justifies the need to create an emotional database for Punjabi language, which would be further used in our research focusing on recognition of emotions from a Punjabi speech.

The paper is organized as follows: Section 2 describes the design process of the database; Section 3 gives the detailed recording process and Section 4 discuss the verification and results and Section 5 highlights its applications and Section 6 concludes the paper.

2. Design

The Punjabi speech database of emotional speech has been recorded under the specific pre-defined conditions, followed by a standard procedure. The same utterance needed to be recorded systematically with different emotions, which makes it necessary to record clean, un-noised and un-echoed data, as it ought to be used for further recognition of emotions.

[1] *Research Scholar, IKG Punjab Technical University, Punjab, India and Department of Computer Science & Engineering, Guru Nanak Dev Engineering College, Punjab, India*
ORCID ID: 0000-0002-3542-1214
[2] *Department of Computer Science & Engineering, Guru Nanak Dev Engineering College, Punjab, India*
ORCID ID: 0000-0003-4715-4966
** Corresponding Author Email: kamal.gndec@gmail.com*

2.1. Speakers

It has been investigated that if basic emotions are to be explored under the experiment, then the natural data makes it less possible to be successful [10]. On the other hand, the quality of database produced using non-professionals may not be suitable for evolving a vigorous speech system with respect to its accuracy [16]. As there is no standard fact about the success of a particular type of database, it was decided to create a database with non-professional speakers, who could speak the sentences in different emotions under supervision. Few points were kept into consideration while selecting the speakers. The number of speakers should be equitable and all the speakers should accomplish all the decided emotions to provide rationalization over the selected group. Same sentence should be uttered by all the speakers in order to permit the correspondence across speakers and emotions. The Punjabi Emotional Speech Database has been created with the help of 15 non-professional people. The 15 participants included 7 males and 8 females, ranging between 20-45 years of age. The reason behind taking this mid age group is the voice clarity. The voice clarity and consistency of speech is low in a child and an adult as compared to a mid-age person. Also, gender has an impact on speech features such as pitch, intensity, etc. That is why, a balance is maintained while gender selection, so that the features are not more biased towards particular gender.

2.2. Sentences for Recording

The text material is usually of two different types i.e. text including irregular series of letters or figures or fancy words [17] and normal sentences which could be used in everyday life. There is emotional neutrality in the first type, but the use of everyday conveyance has proved better [18], because this is the natural form of speech under emotional arousal.

For analyzing emotions from Punjabi language, 10 Punjabi sentences are used for recording. All the sentences are emotionally neutral in nature. The sentences used for the Punjabi database are as follows:

1) ਸੁਰਜੀਤ ਨੇ ਹੁਣ ਘਰ ਵਿਚ ਸ਼ੇਰ ਰੱਖ ਲਿਆ। (sʊrji:t ne: hʊɳ gʰr vɪc ʃe:r rakʰ lɪɑ:)
2) ਪਰੀਖਿਆਵਾਂ ਦੇ ਨਤੀਜੇ ਆਉਣ ਵਾਲੇ ਨੇ। (pri:kʰɪɑ:vɑ:ŋ de: nti:je: ɑ:ʊɳ vɑ:le: ne:)
3) ਡਰਾਈਵਰ ਪਤਾ ਨੀਂ ਕਿੱਥੇ ਚਲਾ ਗਿਆ। (ɖrɑ:i:vr ptɑ: ni:ŋ kɪatʰe: clɑ: gɪɑ:)
4) ਆਹ ਸਮਾਨ ਫਿਰ ਪੂਰਾ ਨਹੀਂ ਆ ਰਿਹਾ। (ɑ:h smɑ:n pʰɪr pu:rɑ: nhi:ŋ ɑ: rɪhɑ:)
5) ਬੱਚੇ ਨੱਚ ਟੱਪ ਕੇ ਸਾਰਾ ਸਮਾਂ ਕੱਢਣਗੇ। (bace: nac ʈap ke: sɑ:rɑ: smɑ:ŋ kaɖʰɳge:)
6) ਉਹ ਗੱਡੀ ਵੱਲ ਦੇਖ ਕੇ ਭੱਜ ਗਿਆ। (ʊh gaɖi: val de:kʰ ke: bʰaj gɪɑ:)
7) ਇੱਥੇ ਫੋਨ ਵਰਤਣ ਦੀ ਸਖ਼ਤ ਮਨਾਹੀ ਹੈ। (ɪatʰe: pʰo:n vrtɳ di: sxt mnɑ:hi: hɛ:)
8) ਪਾਣੀ ਬਰਬਾਦ ਕਰਨ ਤੇ ਹੁਣ ਭਾਰੀ ਜੁਰਮਾਨਾ ਪਵੇਗਾ। (pɑ:ɳi: brbɑ:d krn te: hʊɳ bʰɑ:ri: jʊrmɑ:nɑ: pve:gɑ:)
9) ਸਕੂਲ ਵਿੱਚ ਅੱਜ ਗਣਿਤ ਦੇ ਨਵੇਂ ਅਧਿਆਪਕ ਨੇ ਆਉਣਾ ਹੈ। (sku:l vɪac əaj gɳɪt de: nve:ŋ əʊdʰɪɑ:pk ne: ɑ:ʊɳɑ: hɛ:)
10) ਪੁਲਿਸ ਅਫ਼ਸਰ ਦੀ ਤਬਦੀਲੀ ਹੋਣ ਕਾਰਨ ਹੋਰ ਸਖ਼ਤਾਈ ਹੋ ਗਈ। (pʊlɪs əfsr di: tbdi:li: ho:ɳ kɑ:rn ho:r sxtɑ:i: ho: gi:)

The sentences that have been selected are not specific to any emotion. Neutral sentences are taken into consideration. The speakers had to speak each of these sentences in the said emotion, by embedding that particular emotion into it. This is done so that the sentence doesn't sound biased towards a particular emotion only and reflecting other emotions in it doesn't seem difficult.

2.3. Emotions

For analyzing the emotional speech process of Punjabi language, six basic emotions are taken into account. The six emotions included are happy, neutral, sad, anger, fear and surprise. The evolutionary theory defines the big six model, which is the most fundamental set of emotions including anger, happiness, neutral, sadness, fear and surprise [19]. So, these standard emotions are taken into consideration for this database. There is acoustic difference and similarities between different emotions. Studies show that happy/anger and sad/neutral share similar acoustic properties. Speech associated with anger and happiness are characterized by longer utterance duration, shorter inter-word silence, higher pitch and energy values with wider range. There is slightly higher pitch with wider range in sad speech, as compared to neutral. RMS energy, inter-word silence and speaking rate are useful in distinguishing sadness from others. RMS energy is found to be the only single parameter that is significantly different among the all-emotion classes. Acoustic separability between anger and happiness is poor. There is also similarity between fear/ sadness and happiness/ surprise.

All the speakers had to utter all the 10 sentences in 6 basic emotions in one session. So, the total number of utterances in the database is 900 (10 sentences × 6 emotions × 15 speakers × 1 session). Each emotion has 150 utterances. The number of words in a single statement is ranging from 6-10 respectively.

3. Recording

3.1. Recording Environment

To achieve a high-quality audio, recordings have to be done in a pre-defined environment. It was necessary to record clean, un-noised and un-echoed audios, as it was to be further used for SER process. These conditions or parameters are necessary to be followed while recording. If there is a lot of disturbance in the background environment, or the hardware to be used for recording is not appropriate, or some other software obstacle in the recording phase, then the results would be affected. So, proper care should be taken about the parameters used for recording, so as to achieve good results.

For Punjabi emotional speech corpus, all the recordings were done at a studio, trying to be done without obstacles in the recording path. The recordings were done using a Sennheiser e835 microphone and audacity 2.2.2 software. The sampling rate was taken as 16KHz, represented as 16-bit numbers, with mono-channel recording.

3.2. Recording the sentences

A complete detail of the experimentation process was given to all the speakers at the start, and they were asked about their ease with the sentences, word's pronunciations and emotions. The speakers were insisted to use their own usual way of demonstrating emotions, rather than the overemphasized emotional utterance as stage acting. One emotion at a time was considered for all the utterances. The whole recording process was monitored, with a single session, along with feedback and instructions throughout the process. The speakers were asked to recall an actual circumstance from their past when they had experienced this emotion. Following this method, the speakers tried to re-capture the emotions by developing the same physiological effects as in the real situation.

The speakers produced each of the sentences as many times as they liked with several variants of a sentence. And out of those, one was selected for further process.

4. Verification

4.1. Subjective Evaluation

The subjective evaluation was planned to be carried out for authentication of the recordings that were done. A listening test was carried out, after the recording session, to test if normal listeners could recognize the category/class of emotion with which the utterances had been recorded. The listening subjects were not given any training before the test and they were not given any feedback throughout the process of evaluation. As we wanted to build a realistic system, so in order to capture natural sounding emotions, recordings were done by native Punjabi speakers. The verification of the results of proposed system is done by a group of native Punjabi speakers who are well versed with pronunciation and emotions in Punjabi. It was necessary to carry out this procedure so as to incorporate any changes, if required. If some problem is found with emotion prediction of any of the sentences, the re-recording could be done.

The procedure followed for subjective evaluation consisted of following steps:

i) 20 subjects (10 male and 10 female) were selected with an average age of 35 years.

ii) The listeners were asked to judge the emotional contents of the utterances. The listeners were allowed to hear the utterance 2-3 times before deciding the emotional category. They could not go back to compare with previous utterances and they were not allowed to change any choice made earlier.

iii) After the listening test, the subjects were asked to state the emotion. They were also asked to give suggestions, if any, with respect to the speakers or the recordings, so that improvements could be incorporated.

The main motive of this evaluation process is to create a Punjabi Emotional Speech Database, that would incorporate the specified emotions well, so that it performs well with various applications, including Speech Emotion Recognition for Punjabi language.

4.2. Results

After the recording process is done, the subjective evaluation is carried out to authenticate the process. The results of subjective evaluation are shown in this section. Two performance metrics have been calculated and presented graphically, namely Recognition rate and Confusion matrix.

The recognition rate defines the correctly classified emotions in sentences as compared to the total number of sentences. The recognition rate results of evaluation process are shown in fig. 1. The rates are good for anger, neutral and happy emotions as 97.33%, 96% and 93.3% respectively. The rates for other emotions are bit lower as 91.33% for surprise, 88% for sad and 85.30% for fear respectively. On the whole, the recognition rates are good for all the six emotions.

A confusion matrix, also known as error matrix, summarizes the performance of a classifier. It summarizes the correct and incorrect number of predictions with count values and broken down by each class. The confusion matrix shows the ways in which classification model is confused while making predictions. The confusion matrix is shown in fig. 2. It was also observed that the Anger emotion was the most discriminating and easily recognizable. Fear and Sad emotions were often confused. There was also confusion between Happy and Surprise emotions, but to a very lesser degree. Neutral emotion was also easily distinguishable as compared to others.

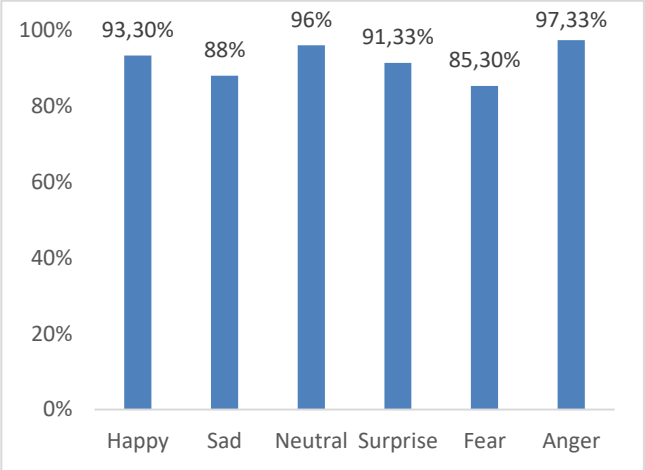

Fig. 1. Recognition Rate for all Emotions

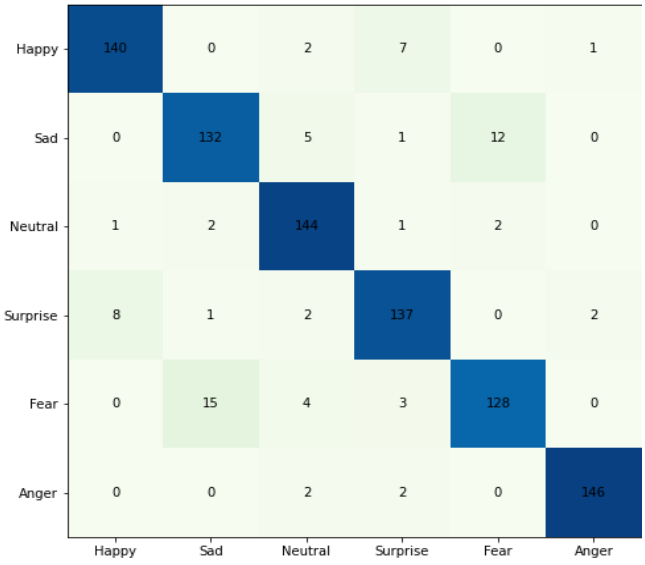

Fig. 2. Confusion Matrix

5. Applications

This Punjabi Emotional Speech Database would be further used for the process of Speech Emotion Recognition for Punjabi language, that can also be used in various application areas such as in medical industry for predicting fatigue, detecting depression with a patient, fear detection in anomalous conditions, psychiatric diseases, detecting child's emotions. It can also be used in e-learning, distance education, detection of lies, car navigation systems, etc.

6. Conclusion

The aim of this work was to create the emotion database for Punjabi language in the form of sound recordings so as to contain real emotional content. The database was designed, recorded and verified using various standards. 15 speakers were chosen with 20-45 years of age and 10 sentences were recorded in each of the six basic emotions namely happy, neutral, sad, fear, anger and surprise, with a total of 900 utterances. All the database specifications have been summarized in Table 1.

Table 1. Database Specifications

Specification	Description
Speakers	15 (7 males and 8 females)
Age group	20-45 years
Emotions	6 (happy, sad, neutral, angry, surprise, fear)
Sentences	10 sentences × 6 emotions × 15 speakers × 1 session=900 total (150 sentences per emotion)
Environment	Studio
Hardware	Sennheiser e835 microphone
Software	Audicity 2.2.2
Sampling Rate	16 KHz
Channel	Mono

As we decided to make the system realistic, choosing speakers was a challenging task, who could speak Punjabi and utter sentences in all emotions. The sentences selected had be neutral in nature. So, sentence selection was also challenging. The speakers were asked to repeat with each sentence number of times, so that sentences recorded should contain the emotion well. This database can be further extended by including more emotions or adding more recordings so that the confusion that is occurring while recognizing happy/surprise and fear/sad can be lowered and recognition rates can be further improved.

This database created for Punjabi language with emotional speech would be further used for the research purpose of Speech Emotion Recognition for Punjabi Language.

Acknowledgements

This work was supported by Guru Nanak Dev Engineering College, Ludhiana, Punjab (India) and IKG Punjab Technical University, Kapurthala, Punjab (India). The authors are thankful to these organizations for their support in this research work. The authors are also thankful to all those who have participated in the recording and verification of this database. We are highly obliged to all the speakers and listeners for their active participation throughout the whole process.

References

[1] I. Luengo, E. Navas, and I. Hernáez, "Feature Analysis and Evaluation for Automatic Emotion Identification in Speech," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, 2010, doi: 10.1109/TMM.2010.2051872.

[2] S. Kuchibhotla, H. D. Vankayalapati, R. S. Vaddi, and K. R. Anne, "A comparative analysis of classifiers in emotion recognition through acoustic features," *International Journal of Speech Technology*, vol. 17, no. 4, pp. 401–408, 2014, doi: 10.1007/s10772-014-9239-3.

[3] P. Chandrasekar, S. Chapaneri, and D. Jayaswal, "Automatic speech emotion recognition: A survey," in *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications, CSCITA 2014*, 2014, pp. 341–346, doi: 10.1109/CSCITA.2014.6839284.

[4] S. Bansal and A. Dev, "Emotional hindi speech database," *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*, pp. 1–4, 2013, doi: 10.1109/ICSDA.2013.6709867.

[5] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012, doi: 10.1007/s10772-011-9125-1.

[6] J. Gomes and M. El-Sharkawy, "i-Vector Algorithm with Gaussian Mixture Model for Efficient Speech Emotion Recognition," *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 476–480, 2015, doi: 10.1109/CSCI.2015.17.

[7] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 1, pp. 115–126, 2015, doi: 10.1109/TASLP.2014.2375558.

[8] P. Jackson and S. ul haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) database*. 2011.

[9] M. Liberman, "Emotional prosody speech and transcripts," *LDC2002S28,University of Pennsylvania*, 2002. .

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology*, 2005, vol. 5, pp. 1517–1520.

[11] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: design and collection of a multilingual emotional speech database," 2002.

[12] X. Mao and L. Chen, "Speech emotion recognition based on parametric filter and fractal dimension," *IEICE Transactions on Information and Systems*, vol. E93-D, no. 8, pp. 2324–2326, 2010, doi: 10.1587/transinf.E93.D.2324.

[13] C. academic of science Institute of automation, "CASIA-Chinese Emotional Speech Corpus," *Chinese Linguistic Data Consortium (CLDC)*, 2005. .

[14] I. S. Engberg, A. V Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," in *5th European Conference on Speech Communication and Technology, Rhodes, Greece*, 1997, pp. 1–4.

[15] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "IITKGP-SEHSC : Hindi speech corpus for emotion analysis," *2011 International Conference on Devices and Communications, ICDeCom 2011 - Proceedings*, 2011, doi: 10.1109/ICDECOM.2011.5738540.

[16] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "IITKGP-SESC: Speech Database for Emotion Analysis," in *Contemporary Computing*, 2009, pp. 485–492.

[17] R. Banse and K. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *Journal of personality and social psychology*, vol. 70, pp. 614–636, Apr. 1996, doi: 10.1037/0022-3514.70.3.614.

[18] K. Scherer, "Speech and emotionnal states.," in *Speech Evaluation in psychiatry*, 1981.

[19] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech and Language*, vol. 25, no. 3, pp. 556–570, 2011, doi: 10.1016/j.csl.2010.10.001.