INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Design and Development of a Multi-lingual Speech Corpora (TaMaR-EmoDB) for Emotion Analysis

Rajeev Rajan[1], *Haritha U.G, Sujitha A.C, Rejisha T.M*

Dept. of Electronics and Communication Engineering
College of Engineering, Trivandrum
Thiruvnanthapuram,Kerala, India
[1]rajeev@cet.ac.in

Abstract

This paper presents the design, the development of a new multilingual emotional speech corpus, TaMaR- EmoDB (Tamil Malayalam Ravula - Emotion DataBase) and its evaluation using a deep neural network (DNN)-baseline system. The corpus consists of utterances from three languages, namely, Malayalam, Tamil and Ravula, a tribal language. The database consists of short speech utterances in four emotions - anger, anxiety, happiness, and sadness, along with neutral utterances. The subset of the corpus is first evaluated using a perception test, in order to understand how well the emotional state in emotional speech is identified by humans. Later, machine testing is performed using the fusion of spectral and prosodic features with DNN framework. During the classification phase, the system reports an average precision of 0.78, 0.60, 0.61 and recall of 0.84, 0.61 and 0.53 for Malayalam, Tamil, and Ravula, respectively. This database can potentially be used as a new linguistic resource that will enable future research in speech emotion detection, corpus-based prosody analysis, and speech synthesis.

Index Terms: emotion, prosodic, corpus, machine testing, subjective

1. Introduction

Speech emotion recognition system extracts the emotional state of a speaker from his or her speech. The need for accurate detection of emotional speech has always been a challenging task due to the growing human-computer interaction. The numerous research attempts, being carried out in this area have necessitated the need for a robust speech emotion corpus. In this paper, we address the creation of a multilingual speech emotional corpus in primary emotions and its baseline-evaluation using a DNN framework. Speech emotion recognition has a wide range of applications such as interacting robots, smart call centers, intelligent spoken tutoring systems, assessing mental attitudes in psychiatric studies, the suicidal tendency analysis from stressed speech, robotics applications and computer tutorial applications[1].

Speech emotion recognition becomes a challenging task due to various reasons such as the choice of proper discriminating features, the acoustic variability, the presence of multiple emotions in the utterance and transience in the emotions. Since there are 300 emotional states in human behavior [1], the set of emotions to be addressed is also an important issue in developing robust speech emotion recognition systems. A detailed account of features and classifiers in an emotional speech recognition system can be seen in [1]. The need for an emotional speech corpus is very much essential for checking the robustness of new algorithms.

In general, the available emotional corpus may be of acted, elicited or spontaneous speech. Acted speech contains speech recordings by actors belonging to different age groups and a different gender. In the elicited speech, speakers are driven into a specific emotion situation and after which their voices are recorded. Spontaneous speech recording provides the best natural utterances as it records the real-world utterances.

1.1. Motivation

Most languages in the world lack the amount of text, speech and linguistic resources required to build good speech/speaker recognition models [2]. However, there have been many approaches incorporating linguistic knowledge into machine-learning based models, that can help in building systems for low resource languages. The use of language information in the upcoming algorithms and the non-availability of a standard database in the selected languages motivated us to create the proposed database. The significance of linguistic information has been effectively exploited in automatic recognition of a speaker's emotion [3]. Elfenbein and Ambady have observed that when emotions are expressed and recognized by the people of the same ethnic or regional group, emotion recognition accuracy is higher [3]. The language-specific paralinguistic patterns also influence the emotion perception [4]. Moreover, speech linguistic features carry information about the culture and the way emotions are expressed or perceived. In [5], a model selection technique based on language identification is proposed to improve speech emotion recognition accuracy. As per the census, tribal people make up about 8.2 percent of the India's total population. This diversity extends to languages as well. A survey done by linguists has established that there exist about 1635 native languages. It is worth noting that the emotional speech created in tribal dialect, Ravula, is among the first of its kind in India. The research outcomes on the acquired tribal dialect corpus can potentially be used to analyze similar language families.

1.2. Related Work

The collection of a robust speech emotion database is a prerequisite for developing efficient algorithms for emotion synthesis and recognition. The majority of the databases record forced (simulated) emotional speech using professional artists, drama students or trained people. In the literature, there are numerous corpus designs with the English language as dominant followed by German in the second position. Cowie et al. [6] created a database in English with 5 emotional states anger, sadness, happiness, fear and neutral. Hansen et al.[7], has designed SUSAS database and it consists of 32 speakers. In addition, four military helicopter pilots were also recorded during flights. M. Edington [8] collected emotional speech databases

http://dx.doi.org/10.21437/Interspeech.2019-2034

Table 1: *Perceptual impressions of listeners for emotions*

Sl.No	Emotion	Impressions
1	Anger	loud, through the squeezed teeth, piercing, rage bursts out from words, voice raised
2	Happiness	Raised voice with positive energy, sound of laughter as if there is a smile, filled with joy
3	Anxiety	Changeable voice, trembling, interrupted voice, anxious voice, almost crying tone
4	Sadness	Slowly, monotonous, slowed down, very quietly, depressive, sadly
5	Neutral	Uniform, quiet voice, somehow identical tone

for training a voice synthesizer. The database contains speech in six emotional categories. ATR laboratories constructed an emotional speech database in Japanese with eight emotions[9]. Emotional databases in German [10], Spanish [11], Danish [12], Dutch [13], Hebrew [14], Chinese [15] are few of the prominent corpus available for research. In [16], IITKGP-SESC and IITKGP-SEHSC emotional speech corpora are used for the analysis of various emotion recognition tasks. In another attempt, K.Scherer has recorded a multilingual speech of 109 airline passengers, waiting at Luggage belt at Geneva Airport [17].

The rest of the paper is organized as follows: Section 2 describes the design, followed by the description of recording in Section 3. The performance evaluation is discussed in Section 4 and the analysis of results in Section 5. Finally, the conclusions are drawn in Section 6.

2. Design of Database

As mentioned earlier, the database consists of simulated speech utterances recorded in a studio environment. The database is designed to maintain the level of naturalness in utterances by avoiding over exaggeration. During the recording process in the studio, speakers are directed to try their best to simulate each emotional states under study. Speakers can simulate one sentence many times until they are satisfied with what they simulated. Finally, we recorded more than 2000 emotional speech sentences covering four emotions along with neutral speech state. The creation of the corpus and its evaluation is discussed in the subsequent sections.

2.1. Language, Emotions and Speakers

The database was designed in three languages namely, Malayalam [1], Tamil [2] and Ravula[3], a tribal dialect in South India. Currently, only two databases are available for Indian Languages, one for Telugu and other for Hindi [16]. Malayalam is a Dravidian language and one of 22 scheduled languages of India [18]. Tamil is predominantly spoken by the Tamil people of India and Sri Lanka, and by the Tamil diaspora. Ravula, known locally as Yerava or Adiyan, is a language spoken by the Ravulas, a tribal community in South India.

The selection of emotions for the corpus development was on the basis of most commonly observed emotions and those which were earlier used by the researchers in emotion recognition [19]. As considered in many emotional speech databases, we collected data samples from emotional states - anger, anxiety, happiness, sadness and neutral speech. As in most of the previous emotional corpus collections, the subjects were asked to read a sentence, expressing a given emotion, which is later used as the emotional label. The perceptual impressions of these emotions are shown in Table 1.

Table 2: *Corpus summary.F stands for Female, M for Male. Emotions, A: Anger, X: Anxiety, H: Happiness, S: Sadness, N: Neutral. Data size is given in number of audio files recorded*

Sl.No	Language	Particulars	Spec.
1	Malayalam	Data Size	1202
		Participants	12(6F, 6M)
		Utterances	[272A, 110X, 254H, 282S, 284N]
		Emotion	5(A,X,H,S,N)
2	Tamil	Data Size	476
		Participants	10(5F, 5M)
		Utterances	[133A, 80X, 92H, 88S, 83N]
		Emotion	5(A,X,H,S,N)
3	Ravula	Data Size	435
		Participants	5(3F, 2M)
		Utterances	[88A, 74X, 90H, 112S, 71N]
		Emotion	5 (A,X,H,S,N)

A total of 27 adult speakers in the age group 18 to 65, participated in the data collection. The speech data was collected from native speakers to reduce the acting effect and produce normal speech. Since the Ravula tribes are notoriously reclusive and are extremely reluctant to mingle with other subgroups of hill tribes, we could collect speech samples from only five speakers The speakers were asked to use their own everyday way of expressing emotional states and not the exaggerated emotional expression known from stage acting.

2.2. Corpus Design and Labeling

The summary of the database is given in Table 2. The sentences were all of short duration with a maximum of 10-30 sec. In the text material, we make sure that all the sentences should be interpretable in the considered emotions. The speakers were asked to pronounce each word while keeping in mind the emotion under which the word is categorized. A few text materials with their meanings, label and emotional states are given in Table 3.

Each emotion utterance is named with a 7-digit alphanumeric code. The first two symbols denote the speaker number. The next letter is for gender label, F for female and M for male. The next letter indicates the emotional state - A for anger, X for anxiety, H for happiness, N for neutral and S for sadness. The next letter represents the language code, M for Malayalam, T for Tamil and R for Ravula language. The number following the third letter gives the utterance number. The last letter denotes the attempt in which the emotion is recorded. A few utterances were recorded twice due to improper tone or rendering of the dialogue. The first attempt was noted as a and second attempt was noted as b. For example, the audio label 01FAM1a denotes the first utterance by speaker one, a female, expressing the emotion anger in the Malayalam language, during the first recorded attempt.

3. Recording

The recording of the designed database was made at a professional recording studio. The speech was recorded in separate sessions to prevent influencing the speaking style of others. Prior to the recordings, the speakers were given a very brief introduction to the recordings and the prompting text. A high-quality microphone was used for the recordings and the actors were in a standing position in front of the microphone at the

[1] https://en.wikipedia.org/wiki/Malayalam

[2] https://en.wikipedia.org/wiki/Tamil-language

[3] https://en.wikipedia.org/wiki/Ravulalanguage

Table 3: *Few text materials used in the creation of the corpus in three languages*

Sl.No	Emotion	Label	Utterances in text (Meaning in English is given in bracket)
1	Anger	03FAM1a	Erangi poda ente veetil ninnu (Get out from my house !)
		02FAT1a	De..shumma peshame irikkaraya (Can you please be silent?)
		01FAR1b	Ni ente karyathiliu thile idanda (You should not interfere in my case)
2	Anxiety	01MXM1a	Raghavante kuttikku asukham veendum vanno entho? (Whether Ragavans son got sick again?)
		05MXT1a	Ayyao.. En kuzhaenthekku Ennachu?... (Oh..What happend to my kid?)
		03MXR1a	Pachakkarikkokke enganothe theebileyaneki engane jeevepa? (How will we survive if price for the vegetable is going up like this?)
3	Sadness	04FSM1a	Njan Eppozhum Ottakkanu. Aarum illandai... (I am alone.. No one is there to help me)
		02FST1a	Appadi.neeyum enne thaniye uttitt Poyidaya (You also left me alone)
		01FSR1b	Nannu enganokke pechadichumu enakku oru jolimu kittuvakani (Even though I tried my level best, I am not able to get a job)
4	Happiness	02FHM1b	Alla, Aara ithu..Kure kalamayallo kandit ..Evidarunnu.. (Hi.. Who is this? Where were you ?..)
		03FHT1a	Ithu yaru. Romba naalai unkale pathittu (Hi..Who is this..I haven't see you!)
		03FHR1a	Ente magakku oru joli kittinaye (My son got a job)
5	Neutral	04MNM1a	PNi nale kalyanathinu pokunnundo? (Are you going for the marriage tomorrow?)
		04MNT1a	Adutha vaaram delhiku touruporem (Next week, going to Delhi for a tour program)
		04FNR1b	Nanu naleya mattannalo veruvaye (I will be coming tomorrow or day after tomorrow)

distance of about 25 cm, which permitted them a certain gesticulation in emotion expression. Each utterance was recorded as a separate .wav file.

4. Performance Evaluation

The evaluation of the database was carried out in two phases: subjective evaluation and machine testing.

4.1. Subjective Evaluation

Subjective evaluation [10, 20, 16] was conducted to ensure the emotional quality and naturalness of the utterances using a perception test. In the perception test, subjects (20 normal hearing persons) were asked to recognize the emotion of recorded speech by listening to the subset of the corpus consisting 100 audio samples per language. The listeners were asked to judge the emotional contents of the utterance by forced choice and to mark them in the previously prepared form. The listeners were given prior information to recognize the emotion solely based on the acoustic content and not on the meaning of the sentence. The perception results are plotted in Figure 1 for all the subsets in the database. It is observed that, for the Malayalam language, the perception accuracy for three emotions, anger, sadness, and neutral is greater than 90%. In the case of Tamil, it is 95%, 84%, and 74% respectively. The anger emotion is well perceived in perception test as compared to other emotional states for Ravula language.

4.2. Machine Testing

In machine testing, emotion classification is performed using the fusion of spectral and prosodic features with DNN. A few researchers are of opinion that continuous prosodic features and spectral features convey much of the emotional content of an utterance [1]. In the front-end, spectral features (20 dim mel-frequency cepstral coefficents(MFCC)) and prosodic features (5 dim) are frame-wise computed from the audio files. Prosodic features namely short-time energy, zero crossing rate are computed from the audio files with framesize of 30ms and hopsize of 10ms. Utterance level prosodic features such as standard deviation of pitch, skewness and kurtosis of pitch distribution, are also computed and appended to the frame-level extracted features in all frames. Due to the ability to capture "global" spectral envelope properties, MFCCs are employed in numerous perceptually motivated audio classification tasks, despite their widespread use as predictors of perceived similarity of timbre [21]. DNN based classifier is used in the classification phase. 60% files in the corpus are used for training and the rest for the testing. Our proposed DNN architecture is based on a three hidden-layered feed-forward neural networks(100 nodes per layer), which was randomly initialized with Adam optimization algorithm. Rectified linear units (ReLUs) have been chosen as the activation function for hidden layers and soft-max function for the output layer. In the DNN experiment, the system was trained for 500 epochs with learning rate of 0.002 and batch size 64. The proposed system is implemented using Tensorflow with Keras as front-end.

5. Results and Discussion

We performed two types of evaluation-subjective and machine testing on the designed corpus as described in above sections. As explained in Section 4.1, the perception test is conducted to test the naturalness of the emotions. Initially, the listeners are allowed to listen to a few audio samples from all the classes to get better acquainted with the peculiarities of each emotion under test. The perception test result shows that the anxiety emotional speech created confusion in choosing the right emotion as compared to other emotional states. It is observed that listeners recognized the emotional state anger easily in all the three languages (more than 90%). Based on the feedbacks from the listeners, it is found that the sound modulation and the tone changes in dialogue delivery have played a key role in choosing the right emotion from the choices. It is quite understandable that semantic content might have also aided in making a decision during listening test. It is also important to note that semantic content is not at all considered during machine testing. On an average, the subjective evaluation resulted in an overall accuracy of 89.6%, 84.6% and 81.6%, for Malayalam, Tamil and Ravula respectively.

As discussed in Section 4.2, machine testing is also performed using a feature-fusion-DNN framework. The confusion matrix of the experiment is given in Tables 4, 5 and 6 for Malayalam, Tamil, and Ravula, respectively. In the evaluation of Malayalam database, the feature-fusion- DNN framework gives least accuracy of 51.35% for anxiety emotional state. This is

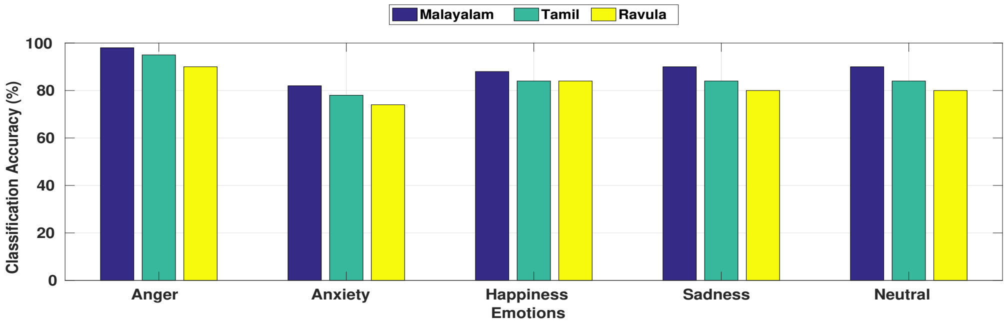

Figure 1: *Perception test result*

Table 4: *Confusion matrix of Feature fusion-DNN framework for Malayalam subset*

Class	Anger	Anxiety	Happiness	Sadness	Neutral
Anger	**79**	2	25	3	0
Anxiety	3	**19**	10	5	0
Happiness	33	8	**44**	14	3
Sadness	4	10	2	**74**	29
Neutral	1	2	3	27	**87**

Table 5: *Confusion matrix of Feature fusion-DNN framework for Tamil subset*

Class	Anger	Anxiety	Happiness	Sadness	Neutral
Anger	**50**	0	2	1	0
Anxiety	2	**18**	0	5	7
Happiness	15	1	**17**	4	0
Sadness	2	1	1	**31**	0
Neutral	1	1	0	2	**29**

Table 6: *Confusion matrix of Feature fusion-DNN framework for Ravula subset*

Class	Anger	Anxiety	Happiness	Sadness	Neutral
Anger	**29**	3	3	0	0
Anxiety	11	**18**	6	0	0
Happiness	14	3	**17**	2	0
Sadness	3	14	6	**20**	2
Neutral	1	10	0	8	**14**

also reflected in the perception test. The perception test reports least accuracy of 88%, 82% for happiness and anxiety, respectively. In addition, the evaluation shows that happiness is strongly confused with other emotions, especially with anger. A possible cause for this is the similarity in raised voice during the dialogue delivery. But, on the contrary, the emotional state, anger is well perceived in both machine testing and listening test with classification accuracy of 72.47%.

Similarly in the evaluation of the Tamil database, it appears that happiness is frequently confused with other emotions, especially anger. In the meanwhile, the anger emotion shows good classification accuracy (94.3%). It appears that sadness and neutral are also better-perceived emotions (88.57 %, 87.87%) in machine testing while happiness performs the worst(46.0%). The classification accuracy for the emotional states, anxiety and happiness is far less than the perception results. It is quite clear that the addition of more linguistic or prosodic features could potentially improve the system performance.

Table 7: *Precision (P) and Recall (R) for three subsets of TaMaR-DB; Malayalam, Tamil and Ravula.*

	Malayalam		**Tamil**		**Ravula**	
Class	**P**	**R**	**P**	**R**	**P**	**R**
Anger	0.71	0.94	0.66	0.72	0.50	0.83
Anxiety	0.85	0.56	0.46	0.51	0.30	0.43
Happiness	0.85	0.96	0.52	0.43	0.53	0.47
Sadness	0.72	0.88	0.60	0.65	0.80	0.44
Neutral	0.81	0.88	0.76	0.73	0.88	0.5

The evaluation of the Ravula database shows that happiness and sadness are confused more with other classes. As in the case of other subsets, the emotional state of anger was better perceived (82.85%) but neutral, the worst (42.4%). It is noted that the anxiety class was reported the least classification accuracy in perception test, which is also correlated with machine testing outcome. Due to the unequal samples from each test classes(unequal priors), we analyze the results using precision and recall. The Table 7 shows the precision and recall of all the classes for three databases. It is found that, average precision of 0.78. 0.60 and 0.61 are obtained for Malayalam,Tamil and Ravula data subsets, respectively. Average recall of 0.84, 0.61 and 0.53 are also reported for the databases. It is believed that a proper selection of features significantly affects the classification performance. The performance can be improved by combining other types of features such as linguistic, discourse information, or facial features. To end the discussion, the proposed databases represent a new linguistic resource that will allow the study of emotional speech in Malayalam, Tamil and Ravula. A web interface with full annotation is being developed to make available the corpus for the research community. Sample audio files can be accessed via the internet: http://mca.rit.ac.in/CASP/downloads.html.

6. Conclusions

In this paper, we have proposed an emotional speech corpus recorded in Tamil, Malayalam, and Ravula. The emotions considered for developing the corpus are anger, anxiety, happiness, sadness and neutral. The quality of the emotions present in the developed emotional speech corpus is first evaluated using a subjective listening test. Later, a machine testing is also performed on the developed TaMaR-Emo-DB corpus using feature-fusion-DNN framework. The performance metrics, precision and recall demonstrate the promise of the designed corpus for further research. It is our hope that the database may aid the research work in speech modification, emotion recognition, corpus-based prosody, and speech synthesis.

7. References

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

[2] E. Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," *Ph.D. Thesis, Massachusetts Institute of Technology*, June 2016.

[3] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis," *Psychological bulletin,*, vol. 128, pp. 203–235, 2002.

[4] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology,*, vol. 32, pp. 76–92, 2001.

[5] H. Sagha, P. Matejka, M. Gavryukova, F. Povolny, E. Marchi, and B. Schuller, "Enhancing multilingual recognition of emotion in speech by language identification," *in proceedings of Interspeech*, September 2016.

[6] R. Cowie, E. Douglas-Cowie, S. Savvidou, and E. McMahon, "FEELTRACE: an instrument for recording perceived emotion in real time," *in proceedings of ISCA workshop on Speech and Emotion*, pp. 19–24, 9 2000.

[7] J. H. L. Hansen, "Linguistic data consortium," *SUSAS LDC99S78. Web Download. Philadelphia*, 1999.

[8] M.Edgington, "Investigating the limitations of conatenative synthesis," *in proceedings of Eurospeech conference*, pp. 593–596, 1997.

[9] R. Nakatsu, J. Nicholson, and N. Tosa, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities," *in proceedings of IEEE Third Workshop on Multimedia Signal Processing*, pp. 439–444, Sep. 1999.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *in proceedings of 6th Annual Conference of the International Speech Communication Association*, 2005.

[11] J. Montero, J. Arriola, J. Cols, E. Enrquez, and J. M. Pardo, "Analysis and modelling of emotional speech in Spanish," *in proceedings of International Congress of Phonetic Sciences*, pp. 957–960.

[12] A. V. Hansen and I. S. Engberg, "Documentation of the Danish emotional speech database," *Internal AAU report, Center for PersonKommunikation,Department of Communication Technology, Institute of Electronic Systems, Aalborg University Denmark*, 1996.

[13] www.elra.info, "Website," *European Language Resources Association*, 1999.

[14] N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," *in proceedings of ISCA workshop on speech and emotion : A conceptual framework for research, Belfast*, pp. 19–24.

[15] Y. Feng, C. Eric, X. Ying, and S. Heung, "Emotion detection from speech to enrich multimedia content," *in proceedings of 2nd IEEE pacific-Rim conferense on multimedia, Beijing*, pp. 550–557.

[16] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "IITKGP-SEHSC : Hindi speech corpus for emotion analysis," *in proceedings of International Conference on Devices and Communications (ICDeCom)*, pp. 1–5, Feb 2011.

[17] K. R. Sherer, "Emotion effects on voice and speech :paradigms and approaches foe evaluation," *in proceedings of ISCA workshop on speech and emotion : A conceptual framework for research,Belfast*, pp. 77–81.

[18] "Official language part XVII," *The constitution of India*, pp. 212–217, 1950.

[19] Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of database," *Speech Communication*, vol. 40, pp. 33–60.

[20] S. T. Jovicic, Z. Kasic, M. Dordevic, and M. Rajkovic, "Serbian emotional speech database: Design, processing and evaluation," *in proceedings of 9th Conference on Speech and Computer, St. Petersburg,Russia*, pp. 77–81.

[21] G. Richard, S. Sundaram, and S. Narayanan, "An overview on perceptually motivated audio indexing and classification," *in proceedings of the IEEE*, vol. 101, pp. 1939–1954, 2013.