Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1538–1545
Marseille, 11–16 May 2020

IIIT-H TEMD Semi-Natural Emotional Speech Database from Professional Actors and Non-Actors

Banothu Rambabu, Botsa Kishore Kumar, Paidi Gangamohan, Suryakanth V Gangashetty
Speech Processing Laboratory
International Institute of Information Technology-Hyderabad, Telangana, India
{rambabu.b, kishore.botsa, gangamohan.p}@research.iiit.ac.in, svg@iiit.ac.in

Abstract

A fundamental essence for emotional speech analysis towards emotion recognition is a good database. Database collected from natural scenarios consists of spontaneous emotions, but there are several issues in collection of such database. Other than the privacy and legal related concerns, there is no control over environment at the background. As it is difficult to collect data from natural scenarios, many research groups have collected data from semi-natural or designed procedures. In this paper, a new emotional speech database named IIIT-H TEMD (*International Institute of Information Technology-Hyderabad Telugu Emotional Database*) is collected using designed drama situations from actors and non-actors. Utterances are manually annotated using a hybrid strategy by giving the context to one of the listeners. As some of the data collection studies in the literature recommend for actors, analysis of actors data versus non-actors data is carried out for their significance. The total size of the dataset is about 5 hours, which makes it an useful resource for the emotional speech analysis.

Keywords: Emotional speech, Natural database, Semi-natural database

1. Introduction

Research on emotions (especially) in speech communication has got more importance. This is due to emergence of technologies like smart phones, intelligent robots, spoken dialogue systems, and many human-machine speech based interactive systems. These technologies can be exploited effectively if they extract emotions from the speech signal. One of the major limitations in developing emotional speech based systems is lack of good databases. Ideally, a database consisting of ‘spontaneous’ emotions collected from natural scenarios is required. But it is difficult to collect such a database, mainly due to privacy concerns and legal issues.

Databases collected by several research groups can be broadly categorized into simulated parallel, semi-natural, and (near to) natural databases (Douglas-Cowie et al., 2003), (Koolagudi and Rao, 2012), (Schuller et al., 2011). Most of the simulated parallel and semi-natural databases were collected from speakers by asking them to elicit specific emotions through scripted scenarios. Performance of emotion recognition systems developed on such databases degrades in real-life scenarios (Busso et al., 2008), (Campbell, 2003). There are many issues, such as selection of speakers, design of scenarios, recording environment, and recording equipment, in collecting a simulated emotional speech database close to real-life scenarios.

Widely used simulated parallel databases such as IITKGP-SESC (in Telugu language)(Koolagudi et al., 2009) and German EMO-DB (Burkhardt et al., 2005) were collected through straight forward process, where the speakers were asked to emote a sentence in all emotions. The speakers involved in these recordings were professional voice talents. The major disadvantages of these kind of recordings are: 1) Speakers are asked to elicit sentences in all emotions, where there is no context. 2) These databases are developed with limited number of sentences and speakers, therefore they do not cover any variability in speaker or lexical information. 3) Annotation scheme is very limited, annotators know the apriori information of emotion. In most of the cases, annotators are asked to rate only the degree (or intensity) of elicitation of emotion. Also, there is no scope for utterances with mixed emotions. In natural human interactions, there exists mixed emotions such as fear-sad, surprise-happy, and anger-sad in a dialogue (Busso et al., 2008), (Izard, 1977), (Scherer, 1984), (Stein and Oatley, 1992), (Ekaman, 1992).

Databases such as Belfast natural database (Sneddon et al., 2011), Geneva airport lost luggage database (Scherer and Ceschi, 1997), and VAM (Grimm et al., 2008) were collected from near to natural cases. The Belfast natural database in English language was collected by segmenting 10-60 seconds long audio-visual clips from TV-talk shows (Sneddon et al., 2011). Geneva airport lost-luggage study database was collected by videotaping the interviews of passengers at lost-luggage counters (Scherer and Ceschi, 1997). The VAM German audio-visual emotion database (Grimm et al., 2008) was collected by segmenting the audio-visual clips from the talk-show “Vera am Mittag”. These databases involve speakers whom are not professional actors. The privacy and copyright issues arise when using the audio-visual clippings from these sources. There is no control over the acquisition of data and environment at the background. Hence, there is a challenge in processing the data for analysis (Busso et al., 2008). Also, limited number of emotions are available in these cases.

Addressing some of the limitations in collecting the databases discussed above, semi-natural databases such as IEMOCAP (Busso et al., 2008) and NIMITEK (Gnjatovic and Rosner, 2010) were collected. The NIMITEK (Gnjatovic and Rosner, 2010) database was developed by simulated human-machine interactions, in audio-visual mode. The IEMOCAP (Busso et al., 2008) database was collected from 10 professional actors in the form of dyadic inter-

Table 1: *Example depicting a drama situation.*

Three characters, namely, *X*, *Y*, and *Z* are involved. A cousin of character *X* is hospitalized, and urgently requires 2 units of B+ blood.
Dialogue 1: *X* angrily speaks over a phone to mediator (character *Y*) for 2 units of blood, as *Y* delays to get blood by 30 minutes. Dialogue 2: *Y* convinces that he will bring 2 units of blood in another 10 mins. Dialogue 3: *X* angrily replies over the phone about the negligence of mediator, and warns him about complaining to police. Dialogue 4: One of the hospital's nurse (character *Z*) comes to *X* and tells her that someone is donating the blood. Dialogue 5: *X* feels surprised, and happily requests *Z* to introduce the donor.

actions with markers on the face and hands, in two scenarios, namely, scripted and spontaneous. The recordings were done after taking a written consent from the speakers. There are three major limitations in the collection of these databases:

- The speakers are instructed to face towards the camera, which restricts the use of natural gestures.
- The fixed position of microphones also restricts the movement of speakers.
- The scripted scenarios make the speakers uncomfortable in remembering the text.

In the view of these limitations, a new procedure of emotional speech data collection is designed. The database named IIIT-H TEMD is collected from 38 (20 female and 18 male) speakers using drama situations. Total of 52 sessions are carried out, involving 4 speakers in each session. The collection of database covers some of the essential elements listed below:

- Drama situations are selected on a basis of real-life scenarios.
- There is no script: Speakers are asked to use their own dialogue and style which are suitable for natural elicitation.
- Movement of speakers: Recording equipment is chosen in such a way that enables free movement of the speakers while enacting the drama situation.
- Many studies in the literature recommend for professional actors (Busso et al., 2008). But there are also databases collected from non-actors. We collected data from both professional actors and non-actors, and analyzed differences among these categories.
- All annotations are assigned by human subjects, in a mixture of both context-dependent and context-independent schemes.

From design to annotation stages of the data, it took approximately 9 months. Different stages of development of this data are given in Section 2.. Statistics of the collected data are given in Section 3.. Analysis of professional actors data versus non-actors data is carried out in Section 4.. Finally, Section 5. gives summary and conclusions.

2. Collection of IIIT-H TEMD Database

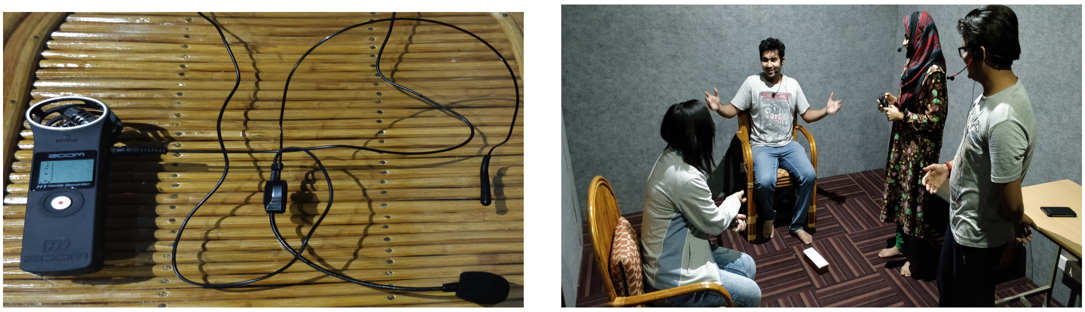
The database is developed in several stages, and each stage is described in detail in the following sections.

2.1. Stage 1: Design of drama situations

The objective of the data collection is to design the situations where speakers can emote themselves without having any script. The drama situations involving maximum of 4 characters are designed by taking the motivation from Telugu (Indian) language featured films, short films, and soap-operas. By watching several videos, 602 drama situations are designed over a period of time. The number of dialogues in these drama situations do not exceed more than ten. This is done by 5 subjects and supervised by 2 professionals. An example depicting such a situation is given in Table 1.

2.2. Stage 2: Selection of speakers and recording equipment

Data is collected from 38 Telugu native speakers in several sessions. Each session is carried out for a duration of 3 hours, involving 4 speakers. Professional actors with minimum experience of one year are engaged. Non-actors are selected after a few auditions. The auditions are supervised



(a)

(b)

Figure 1: *(a) Recording equipment used for data collection. (b) Speakers involved in the elicitation of a drama situation.*

by 2 professionals (assistant directors). The speakers' age ranges from 21 to 46 yrs, with average age of 33.5 yrs.

Hand portable ZOOM H1 handy recorders and head worn microphones are used for the recording purpose. The recording material is shown in Fig. 1 (a). Each individual speaker is given his/her own recording setup, as illustrated in Fig. 1 (b). The speakers involved in the recording sessions are not restricted to any gestures and movements.

2.3. Stage 3: Recording sessions

Recordings are done in Telugu language, using sampling rate (f_s) of 44.1 kHz and 16 bits/sample. The data is recorded in a professional recording studio available at International Institute of Information Technology-Hyderabad (IIIT-H). The dimensions of the room are 5.5 x 3.5 x 3.5 (length x breadth x height in meters), which gives enough space for movement of the speakers.

Each session included 4 speakers and a supervisor. Firstly, the supervisor explains the drama situation to be enacted. The factors involved in the drama situation are: Number of characters, assigning the speakers, dialogue counter sequence, and emotional states of the speakers. The speakers are also instructed to use their body gestures freely. A few rehearsal sessions are carried out, where the speakers are instructed to display the desired emotional situation by using their own sentence formation. In each session, approximately 15-20 drama situations are performed.

As each speaker has his/her recordings in the corresponding recorder, apart from the near field information of the speaker, the far field information of all the speakers is also captured.

2.4. Stage 4: Editing the raw data

In each recorded drama situation, data of a particular (near field) speaker is present in the corresponding recorder. After listening to the speech signals in the respective recorders, the utterances are manually segmented. The utterances are labeled (i.e., name of the wave file) with the session number, source information, and dialogue number. The labeling sequence comprising of 18 characters is given by the following example *"S01-MOV-C01-SPKM01"*, where

- *S01*: Refers to the session 1.
- *MOV*: Refers to the movie source.
- *C01*: Refers to the first drama situation from the source.
- *SPKM01*: Refers to a male (M) speaker (with given identity 01).

2.5. Stage 5: Annotation of the data

Annotation is a key step in developing an emotional speech database. There are various annotation strategies used in the literature. Two major procedures of annotation of an utterance in a semi-natural case are: Annotation with context and annotation without context. In annotation with context procedure, the information from various sources such as body gestures, facial expressions, and also total dialogue sequence is available. In annotation without context procedure, the annotation is done by having access only to an utterance.

Each utterance is annotated by 3 subjects, all of whom are native Telugu speakers. The context of the utterance, i.e., entire dialogue sequence and emotional state of the speaker, is given to one subject. Based on the judgment of the remaining subjects, the emotional state with a confidence level is assigned to the given utterance. The confidence level ranges from $1-5$. The approach followed for assigning the confidence level is given in the following steps:

- As the underlying state of the utterance is known to one subject, if none of the other two subjects agrees with the same emotion, then the confidence level is given as 1 or 2.
- If one of the other two subjects agrees with the same emotion, then the confidence level 2 or 3 is assigned based on the degree of elicitation (enacting performance).
- If all the three subjects agrees with the same emotion, confidence level 3, 4, or 5 is assigned based on the degree of elicitation (enacting performance).

It is important to note that there are a few cases with multiple emotional states. For an instance, there is a possibility of emotion combinations such as surprise-happiness, frustration-anger, and anger-sadness. The subjects are instructed to tag both the emotions in the labeling. The label of an utterance gives most of the information. For example, *"S01-MOV-C01-SPKM01-AN5-XXX"* gives the entire labeling scheme, where

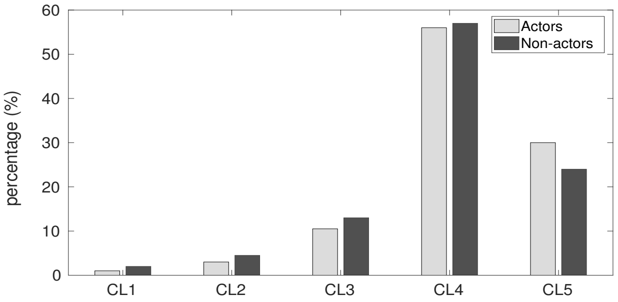


Figure 2: *Distribution of confidence levels across the emotional states for actors and non-actors data.*

- *AN5*: Refers to anger emotion with confidence level 5.

- *XXX*: Refers to nonexistence of second emotion category. If another emotion exists in the utterance, then these characters are replaced by the emotion category.

3. Statistics of the data

Data is collected from 2 sets of speakers, namely, acting professionals and non-actors. The set of acting professionals consists of 19 speakers (12 female and 7 male), and the set of non-actors consists of 19 speakers (8 female and 11 male). On the whole, the database is composed of 5317 annotated utterances, of which 2341 utterances are of female speakers and 2976 utterances are of male speakers. The statistics of the data of both the sets as per emotion and confidence levels are given in Tables 2 and 3. Fig. 2 shows the distribution of confidence levels across the emotional states for acting professionals and non-actors. From this figure, it can be observed that the acting professionals data have comparatively higher percentage of utterances with confidence level 5. This indicates the ability of acting professionals in enacting the drama situations.

The database consists of other categories of expressive voices such as laughter and cry. The laughter and cry utterances are also annotated with various confidence levels. The laughter and cry categories comprises of 84 and 204 utterances, respectively. In addition to the utterances with single emotion, the database also consists of 275 utterances with multiple emotions in a dialogue. Table 4 gives the number of utterances with multiple emotional states.

Apart from the emotion recognition application, the database is useful in several applications such as speech recognition and speaker recognition in emotional environments. Other than emotion-based applications, the data is also useful in analysis of near field and far field speech analysis.

4. Analysis of actors vs non-actors data

The objective of this study is to understand the differences in the data among actors and non-actors. Perceptual (qualitative) and acoustic (quantitative) analyses are carried in this section.

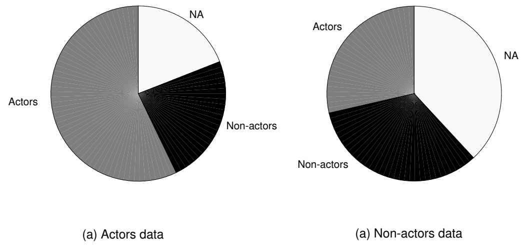


Figure 3: *(a) Percentage of actors set (of utterances) labeled as $actors$, $non\text{-}actors$, or $indistinguishable$ (NA). (b) Percentage of non-actors set labeled as $actors$, $non\text{-}actors$, or $indistinguishable$ (NA).*

4.1. Perceptual analysis

Perceptual analysis is carried out through listening tests with 18 student listeners from International Institute of Information Technology-Hyderabad (IIIT-H). These subjects are native Telugu speakers. A set of utterances corresponding to a speaker (actor/non-actor) is given to a subject, and asked to identify whether the speaker is actor or non-actor. Five emotion categories, namely, anger (AN), happiness (HA), sadness (SA), neutral (NU), and surprise (SU) are considered for the analysis. Two utterances (with confidence levels 5 or 4) of each emotion category of each speaker, i.e., a set of 10 utterances of each speaker is used for perceptual evaluations. For a given set of utterances of a speaker, each subject is asked to label the speaker as one of the following $actor$, $non\text{-}actor$, and $indistinguishable$. Hence, for each speaker we obtain 18 judgments. Figs. 3 (a) and (b) show the percentage of judgments for actors set and non-actors set, respectively, in terms of categories $-actors$, $non\text{-}actors$, or $indistinguishable$. From these figures, it can be observed that there is a bias in the decision towards $actors$. This is mainly due to the following: The non-actors are selected (out of 31) before the recordings. The utterances used for listening tests correspond to confidence levels 4 and 5.

Also, Figs. 3 (a) and (b) indicate that the actors identified as $actors$ is more when compared to non-actors identified as $non\text{-}actors$. The confusion is more in the case of non-actors. The explanation for the above observation from the listeners is following

- Clarity and presentation of the linguistic information

- Quality of voice and its modulation for various emotions

In the next section, the acoustic analysis of actors and non-actors data is carried out. The parameters related to the excitation source component of the speech production are analyzed.

4.2. Excitation source analysis

The production of speech by humans is a complex phenomenon, which is interpreted as coupling effect of the

Table 2: *Annotation of actors data. Number of utterances per emotional state, with confidence levels (CL1 to CL5).*

	CL5	CL4	CL3	CL2	CL1	Total
Anger (AN)	125	240	63	14	1	443
Happiness (HA)	81	106	8	-	1	197
Sadness (SA)	56	77	1	1	8	143
Neutral (NU)	323	494	83	55	12	967
Surprise (SU)	36	78	24	4	-	142
Fear (FE)	22	40	15	-	-	77
Disgust (DI)	7	22	4	1	-	34
Sarcastic (SR)	6	46	14	1	-	67
Frustrated (FR)	16	102	14	-	-	132
Relaxed (RE)	1	6	1	-	-	8
Worried (WO)	25	79	20	1	-	125
Shy (SH)	13	6	-	-	-	19
Excited (EX)	4	38	10	2	-	54
Shout (SO)	12	28	2	-	-	42

Table 3: *Annotation of non-actors data. Number of utterances per emotional state, with confidence levels (CL1 to CL5).*

	CL5	CL4	CL3	CL2	CL1	Total
Anger (AN)	90	165	60	12	1	328
Happiness (HA)	41	112	18	3	-	174
Sadness (SA)	42	67	9	6	6	130
Neutral (NU)	229	434	115	61	26	865
Surprise (SU)	14	67	6	-	-	87
Fear (FE)	6	20	10	1	-	37
Disgust (DI)	1	10	-	-	-	11
Sarcastic (SR)	10	41	10	-	-	61
Frustrated (FR)	18	77	9	-	-	104
Relaxed (RE)	3	8	3	1	-	15
Worried (WO)	9	66	4	1	-	80
Shy (SH)	2	8	-	-	-	10
Excited (EX)	2	21	1	1	-	25
Shout (SO)	4	11	5	1	-	21

dynamic (or time-varying) vocal tract system and time-varying excitation source. The excitation source component constitutes of two primary modes of larynx, namely, voiced and unvoiced. In voiced mode, the repeated process of vocal folds abduction and adduction gives rise to the vocal folds vibration. These modes are voluntarily changed by the speaker in producing voiced and unvoiced sounds. In voiced and unvoiced modes, the shape of the vocal tract system changes due to movement of various supraglottal articulators, to produce different sounds. Thus the vocal tract system component mainly contributes to the linguistic information.

During production of speech in emotional state, there are certain deviations in components of the speech production

Table 4: *Number of utterances with different combinations of emotional states. Anger (AN), happiness (HA), sadness (SA), surprise (SU), fear (FE), sarcastic (SR), frustrated (FR), and worried (WO).*

	AN	HA	SA	SU	FE	SR	FR	WO	SO
AN	-	0	20	21	2	17	33	10	49
HA		-	2	6	0	3	1	1	8
SA			-	0	4	1	6	8	4
SU				-	2	3	2	10	4
FE					-	0	0	9	2
SR						-	2	0	2
FR							-	10	24
WO								-	9

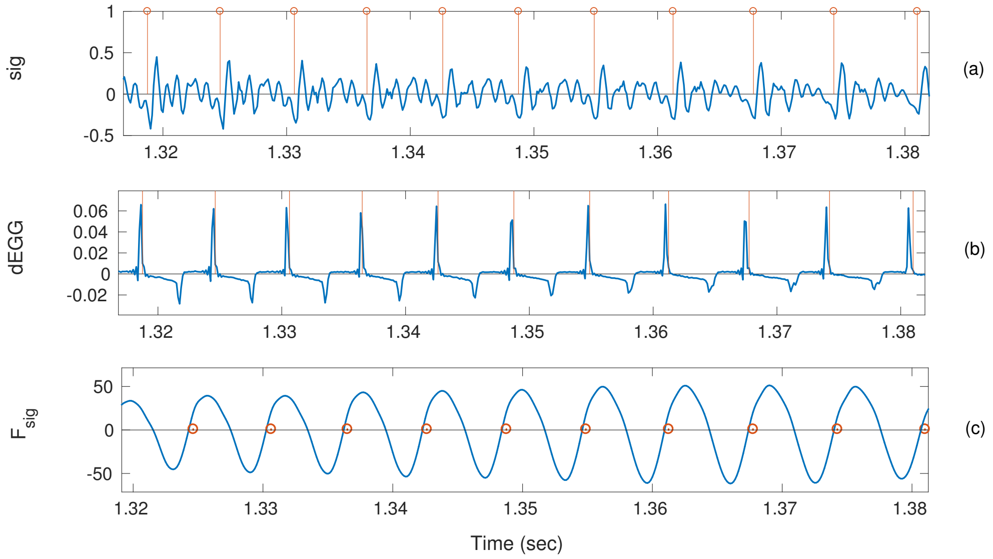
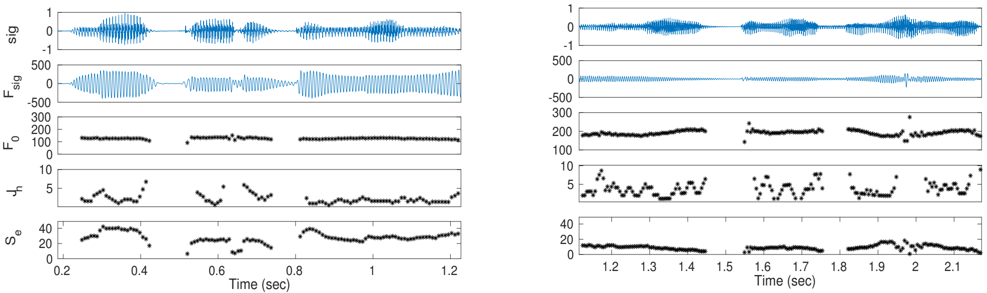


Figure 4: *(a) Speech segment (sig). (b) Differenced EGG signal of the corresponding speech segment (dEGG). (c) Zero frequency filtered signal (F_{sig}). Epoch locations are highlighted with circles ('o') in (c)*

mechanism with respect to neutral state (Gangamohan et al., 2013). From several studies in the literature, it is observed that the emotion characteristics are reflected more in the excitation source component, and to a lesser extent in the vocal tract system component (Koolagudi and Rao, 2012; Gangamohan et al., 2013). The objective of this paper is to analyze deviations in the parameters related to the excitation source component in an emotion-specific manner, between the actors and non-actors data.

Three parameters, namely, fundamental frequency (F_0), jitter (J_h), and strength of excitation (S_e) are used for analysis. These parameters correspond to the excitation source component, and are derived using the zero frequency filtering (ZFF) approach (Murty and Yegnanarayana, 2008).

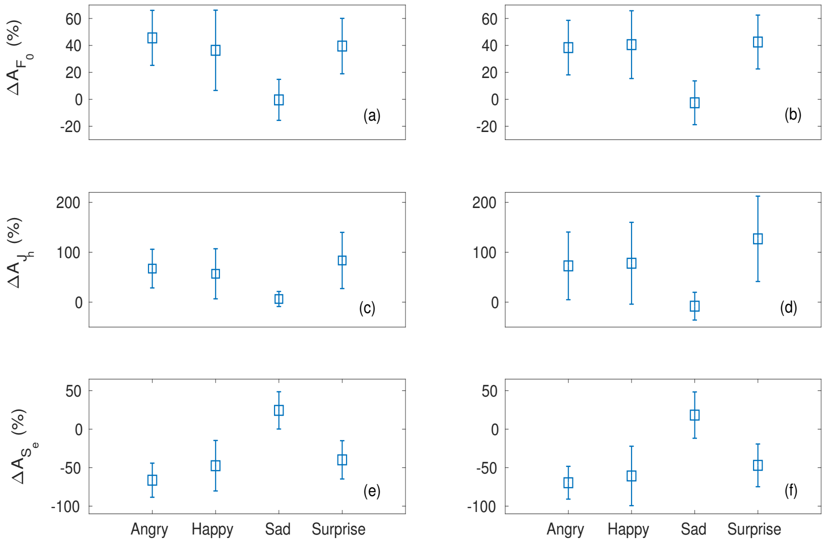
Epoch locations refer to instants of significant excitation of the vocal tract, which occur at the glottal closure instants (GCIs). In the ZFF method (Murty and Yegnanarayana, 2008), the speech signal is passed through a cascade of two ideal zero frequency resonators. Due to the polynomial growth/decay in the output signal, the trend removal operation is required. The length of the window for the trend removal operation is chosen to be 1.5 times the average pitch period. The negative to positive zero crossing instants in the filtered signal correspond to the epochs. The filtered signal along with the original speech signal and differenced electroglottograph (dEGG) signal are shown in Fig. 4. The interval between the successive epochs gives the value of the instantaneous pitch period (T_0), and the instantaneous fundamental frequency ($F_0 = \frac{1}{T_0}$). The slope of the filtered signal at each epoch (S_e parameter) is observed to be proportional to the magnitude of the corresponding peak in the dEGG signal (Yegnanarayana and Murty, 2009). Absolute jitter (J_h) is given by the average absolute difference



(a) Neutral speech

(b) Angry speech

Figure 5: *(a) Speech segment (sig), zero frequency filtered signal (F_{sig}), F_0 contour, S_e contour, and J_h contour of neutral speech. (b) Speech segment (sig), F_{sig}, F_0 contour, S_e contour, and J_h contour of angry speech.*



(Actors data)

(Non-actors data)

Figure 6: *(a), (c), and (e) show the deviation in F_0, J_h, and S_e parameters corresponding to the actors data. (b), (d), and (f) show the deviation in F_0, J_h, and S_e parameters corresponding to the non-actors data*

between consecutive instantaneous F_0s,

$$J_h = \frac{1}{N-1}\sum_{i=1}^{N-1} |F_{0_i} - F_{0_{i+1}}| \qquad (1)$$

where N is the number of considered instantaneous F_0s, and F_{0_i}s are the extracted instantaneous F_0s. The F_0, J_h, and S_e contours of natural and angry speech segments are shown in Fig. 5.

As reported in the studies (Gangamohan et al., 2013; Murray and Arnott, 1993), there are deviations in F_0, J_h, and S_e parameters in emotion-specific way. For example, there is increase in F_0 and J_h, and decrease in S_e in the case of angry and happy speech when compared to neutral speech. In this study, we are interested in the degree of deviation in these parameters between the actors and non-actors data. The degree of deviation in these parameters in an emotion-specific way with respect to neutral speech for actors and non-actors data is analyzed. An illustration of deviation (ΔA_{F_0}) in the average F_0 (A_{F_0}) of angry speech with respect to neutral speech is given by

$$\Delta A_{F_0}(in\%) = \frac{A_{F_{0_{angry}}} - A_{F_{0_{neutral}}}}{A_{F_{0_{neutral}}}} * 100 \qquad (2)$$

where $A_{F_{0_{angry}}}$ and $A_{F_{0_{neutral}}}$ are the average F_0 of angry and neutral utterances, respectively. Similarly, ΔA_{J_h} and ΔA_{S_e} are computed for the average J_h and S_e values of emotional speech with respect to neutral speech. For this analysis, five utterances with confidence levels 5 or 4 of each emotion category of each speaker are considered. Fig. 6 shows the box plots of ΔA_{F_0}, ΔA_{J_h}, and ΔA_{S_e} for four emotions (angry, happy, sad, and surprise) in the cases of actors and non-actors data. In each box plot, *square* gives the mean value and the vertical line gives the standard deviation.

The following are the observations that can be drawn from Fig. 6.

- The trends in the feature deviations are emotion-specific in both the cases of actors and non-actors data.
- On comparison of Figs. 6 (a) and 6 (e) with Figs. 6 (b) and 6 (f), the deviations in the average F_0 and average

S_e are similar in the cases of actors and non-actors data.

- On comparison of Fig. 6 (c) with Fig. 6 (d), it is observed that the jitter is more in the cases of non-actors data when compared to actors data.

The above observations convey that the actors portray emotions with increase or decrease in F_0 with less jitter when compared to non-actors, this may be due to their efforts in presentation of the linguistic message. In realistic scenarios, we often are interested in detecting emotions of persons who are non-actors. Therefore, a database developed from non-actors with critical annotations seems to be useful.

5. Summary and conclusions

This paper presents a new database named IIIT-H TEMD, which is collected from actors and non-actors by enacting the drama situations. The database provides a potential resource to analyze emotions in the speech signal. From the design of drama situations to annotation, there are 5 stages involved in the data collection process. A different annotation procedure is adopted by the mixture of context-dependent and context-independent ways. There are some utterances in the database where there is an occurrence of multiple emotions. Apart from the suitability of the database for emotion classification, it is also useful in applications such as speech recognition and speaker verification for emotional speech.

Perceptual and acoustic analyses are carried out to understand differences in the data among actors and non-actors. Perceptual studies indicate that the actors intend to convey emotions in speech without affecting the linguistic message much. Deviations in the acoustic parameters related to the excitation source component in an emotion-specific manner are analyzed. It is observed that there is more jitter in non-actors data when compared to actors data.

Acknowledgements

The authors would like to thank Sumedh, Bhavana, Nikhil, Sai, and Varun for their invaluable efforts in collecting the drama situations. The authors would also like to thank Lalitha, Gopi Madhav, and Samarasimha Reddy for providing the contacts of actors related to the Telugu cinema industry, and also handling the recording sessions.

6. Bibliographical References

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of German emotional speech. In *Proc. INTERSPEECH*, pages 1517–1520.

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

Campbell, N. (2003). Databases of emotional speech. In *Proc. ITRW on Speech and Emotion (ISCA)*, Northern Ireland, UK.

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

Ekaman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6:169–200.

Gangamohan, P., Reddy, K. S., and Yegnanarayana, B. (2013). Analysis of emotional speech at subsegmental level. In *Proc. INTERSPEECH*, pages 1916–1920.

Gnjatovic, M. and Rosner, D. (2010). Inducing genuine emotions in simulated speech-based human-machine interaction: The nimitek corpus. *IEEE Transactions on Affective Computing*, 1(2):132–144.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *Proc. Int. Conf. Multimedia and Expo*, pages 865–868.

Izard, C. E. (1977). *Human emotions*. Plenum Press.

Koolagudi, S. and Rao, K. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117.

Koolagudi, S. G., Maity, S., Vuppala, A. K., Chakrabarti, S., and Rao, K. S. (2009). IITKGP-SESC: Speech database for emotion analysis. In *Proc. Communications in Computer and Information Science*, pages 485–492.

Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Am. (JASA)*, 93(2):1097–1108.

Murty, K. S. R. and Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, and Language Processing*, 16(8):1602–1613, Nov.

Scherer, K. R. and Ceschi, G. (1997). Lost luggage: A field study of emotion-antecedent appraisal. *Motivation and Emotion*, 21(3):211–235.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer et al., editors, *Approaches to emotion*. Lawrence Elbraum, Hillsdale, N.J.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087.

Sneddon, I., McRorie, M., McKeown, G., and Hanratty, J. (2011). The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41.

Stein, N. and Oatley, K. (1992). Basic emotions: Theory and measurement. *Cognition and Emotion*, 6:161–168.

Yegnanarayana, B. and Murty, K. S. R. (2009). Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Trans. Audio, Speech, and Language Processing*, 17(4):614–624, May.