AMLED: Design, Collection and Preliminary Investigation of a Multilingual Emotional Speech Database

[1]Lobsang Darge, [2]Bomken Kamdak, [3]Bhaskar Jyoti Chutia, [4]Mrinal Jyoti Sarma and [5]Utpal Bhattacharjee[*]

Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India. Pin – 791112

[1]lobsangdarge@gmail.com,
[2]bomken.kamdak@rgu.ac.in[3]bhaskar.chutia@rgu.ac.in[4]mrinaljyotisarma@gmail.com[5]utpal.bhattacharjee@rgu.ac.in

Abstract:

In this paper, we are introducing a database for speaker recognition in multilingual and emotional conditions. The paper describes the design, collection and preliminary investigations carried out on a multilingual emotional speech database, named Arunachali Multilingual Emotional Database (AMLED), developed using English, Hindi and the languages of Arunachal Pradesh, namely, Adi, Nyishi, Galo, Apatani and Monpa, spoken by the natives of Arunachal Pradesh, a north east frontier state of India. The speech corpus consists of two different categories of emotions– simulated or acted and emulated or induced. The speech corpus has been recorded from 100 native speakers of Arunachal Pradesh. Each speaker has recorded for four different types of emotions– angry, happy, sad and neutral. The neutral voice of each speaker is recorded in three different languages– English, Hindi and the mother tongue of the speaker, which is any one of the above five Aruanchali languages. The proposed database will be useful for speaker verification at emotional and multilingual conditions. Further, the database can be used for identifying acoustic cues that distinguish true emotions from the acted emotions. To determine the quality of the emotions present in the speech signals, subjective listening test has been conducted. Statistical analysis of the features extracted from the speech signals recorded at different emotional and linguistic conditions have been carried out to determine the changes at parameter level with change in emotional and linguistic conditions.

Keywords: ***Emotional Speech Corpus; Multilingual Speech Corpus; Simulated Emotion; Emulated Emotion; Subjective Listening Test; Statistical Analysis.***

DOI Number:10.48047/nq.2021.19.1.NQ21046 NeuroQuantology2021;19(1):350-361

[*] Corresponding author

1. Introduction

In recent years, speech-based technology has progressed to the level of commercialization. However, mitigating the performance issues arising out of mismatched acoustical environments is still a major challenge for speech-based systems. Speech signal, which is a complex signal is produced by the involvement of two time-varying systems, the vocal tract and the excitation [1]. The speech signal is usually used to convey the messages in human-to-human communications. In addition, it contains a lot of other information such as the language spoken, speaker identity and even some indications about the age, gender, social-cultural background and emotional state [2]. During verbal communication, human uses linguistic as well as other non-linguistic information, such as emotion to frame a natural and efficient communication with its human counterpart. However, this multimodality of speech signal possesses serious challenges for the implementation of automatic speech-based systems. The human-machine communication is still far from such naturalness. Speech processing system can reach human equivalent performance only when such systems can detect the underlying emotion of a speech signal effectively [3]. Speech emotion recognition can be used to extract useful semantics from speech to improve the performance of the speech-processing system [4]. Speech emotion recognition is particularly useful in systems which require man-machine interaction where the response to the user depends on the emotional state of the user [5]. It can also be used as a diagnostic tool for treating various disorders by therapists [6]. Improving the communication behaviour of robots is an urgent requirement for the development of human-friendly personal robots. Robots have to be spontaneous and polite and must react according to the emotional state of the user [7]. Natural communication between human and robot is very difficult without an emotional feedback system [8][9]. In automatic translation systems, the emotional state of the speaker plays an important role in communication between parties [10]. In case of cross language speech-to-speech translation, the emotion in the source language needs to be properly recognized and synthesize the same in the target language [11]. Emotion recognition systems may be used in an on-board car driving systems to keep the driver alert during driving to avoid any accident [12]. Speech emotion recognition system may also be useful in call centre applications and mobile communications. Analysis of the conversation may be helpful in behavioural study of the call attendant with their customers and help to improve the quality of services [13]. In aircraft cockpits, a speech recognition system trained with emotional speech perform better than a system trained with normal speech [14]. Call analysis of emergency services like ambulance and fire brigade may help in evaluating the genuineness of those requests [15]. The task of speech emotion recognition is highly challenging for the following reasons:

a) The term emotion is highly intuitive in nature. It is individual instinctive feelings that arise spontaneously without conscious efforts. It is inherently subjective in nature and it is difficult to have a well-accepted objective measure for it.
b) Emotion is recognized by analysing the acoustic difference of the utterances at different emotional conditions. The features which are used for developing an emotion recognition system are influenced by speaker and language dependent information.
c) Emotion is dynamic in nature. In a single utterance, more than one emotion may be present. Moreover, there may be some predominant emotions based on the prolonged mental condition of the speaker and there may be some transient emotions which occurs momentarily during the conversation.
d) Method of expressing emotion

depends on socio economic and cultural background of the speaker.

e) Noise robustness of emotion detection systems is another issue that needs to be addressed for effective emotion detection.

In this paper, we have described the design, recording and verification of an emotional speech database for speaker verification research. The database is recorded using English, Hindi and the languages of Arunachal Pradesh in North East India. We called it Arunachali Multi Lingual Emotional Database (AMLED). Arunachal Pradesh in North East India is home to thirty major tribal languages and many distinct dialects and subdialects thereof.The vast majority of indigenous languages in Arunachal Pradesh belong to the Tani language family, which is a branch of the Tibeto Burman language family. In addition to these non Indo European languages, Indo-European languages like Assamese and Hindi are spoken extensively by the natives of Arunachal Pradesh.

2. Scope and Types of Emotional Databases

A major question in the development of an emotional database is about the scope of the database. The term scope covers several kinds of variabilities, the number of speakers, languages spoken, gender of the speaker, etc[16]. These variabilities are important due to the fact that emotion is subjective in nature, and signs of emotion are highly inconsistent across individuals and occurrences. An ideal emotional database should be able to capture all possible occurrences of an emotion. It must be able to capture the variability in expressing an emotion across individual, language and socio-cultural background. The scope of the database also depends on the research goal of developing the database. Another important design decision for the development of an emotional database is about the nature of the emotional speech materials. Generally three methods are used for collecting the emotional speech namely– simulated or acted, induced or emulated and natural. The simulated emotional speech material is collected from professional actors. These actors are asked to express the emotion in otherwise emotionally neutral sentences. The emotions generated by simulated means are intense and contains most of the aspects that are required for expressing an emotion [17]. The emotion collected through simulated means are fully blown and are expressive in nature [10][11][18]. The simulated emotion provides the benchmarking characteristics about an emotion; however, they are far from the actual emotions occurred in natural conversations. The emulated or induced emotions are generated by creating artificial emotional environment without the knowledge of the speaker. Speakers are gradually involved in emotional conversations by an anchor. To emulate different types of emotion, different contextual situations are created without the knowledge of the subject. The emotions in the emulated databases are less expressive compared to their simulated counterparts and more close to the natural emotions. The emulated emotions are recorded in a controller acoustical environment; therefore, it is normally free from environmental noises. However, since the speakers are aware of the fact that they are going to be recorded, most of the times, the emotions are expressed in a very controlled manner. Therefore, careful observation is required for the segmentation and labelling of the emotions. Natural emotions may be recorded from real world voice conversations. Some of the sources for natural emotions are recorded telephonic conversations to police stations, fire services, ambulance services, call centers, spot interviews of different news channels etc. Such emotions are naturally presents in those conversations. The natural emotions are mildly expressed and difficult to clearly recognize. This type of emotion is also called underlying emotions. The most challenging task for the development of a natural emotional database is the identification and segmentation of the soft emotions present

in the speech signal. The major advantage of natural emotional database is that the emotions are naturally expressed and useful for real-world emotion modelling. However, only a limited number of emotions are present in such a database. The emotions are overlapping and not clearly distinct from each other. Moreover, due to uncontrolled acoustical environment, background noises are present in the speech samples of the natural databases.

3. A Brief Review of the Emotional Database

In this paper, we have reviewed 16 (Sixteen) emotional databases. The salient characteristicfeatures of each database have been presented below: Faculty of Electrical Engineering and Computer Science, university of Maribor, Slovenia [19] developed an emotional database for English language. The database consists of six emotions namely disgust, surprise, joy, fear, anger and sadness. Two natural emotions, fast loud and slow soft were also included. The emotions are simulated by the speakers. The database consists of 186 utterances per emotional category. Emotional prosodic speech and transcripted database was developed by Liberman at the University of Pennsylvania [20]. The database consists of 15 emotions namely neutral, hot anger, cold anger, panic, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt. The recording was done by professional actors. The database consists of 9 hours of simulated emotional speech in English language. At Macquire University, Pereira constructed a simulated emotional database in English language [21]. The database consists of 40 different sentences uttered by 2 actors. There were two repetitions of these 40 sentences. Hence, the database consists of 80 sentences. The utterances were rated by 31 normal hearing subjects. The database consists of 14 emotions. M.Edgington at BT Labs, UK collected a simulated emotional database in English language for voice synthesizer [22]. The database was recorded by trained male actors. Thirteen raters judged the identity of the emotions with 79.3% score rate. The database consists of the six emotions: anger, happiness, fear, sadness, boredom and neutral. Polzin and Waibel at the Carnegie Mellon University constructed simulated emotional database [23] in English. The database was recorded by 5 drama students. The database comprised of 291 words taken per emotion per speaker. Emotion considered were happy, sad, angry and afraid. In addition, it contains neutral pronunciations for all 50 sentences. The subjective evaluation of the database was conducted and it showed 70% recognition accuracy. Sendlmeier et al [24][25] at the Technical University of Berlin collected a simulated emotional database in German language. The database was recorded by the participation of 10 professional speakers. The emotions considered in the database were–hot anger, happiness, fear (panic), sadness (sorrow), boredom, disgust and neutral. Subjective tests were conducted with the participation of 25 judges and they classified each emotion with a score rate of 80%. Nakatsu et al [26] at ATR laboratories constructed a simulated emotional database in Japanese language. The database was recorded with the participation of 100 native speakers (50 male and 50 female). The reference material was generated by one professional radio artist. The professional artist was told to read 100 neutral words in 8 emotions. The ordinary speakers mimicked the reference material generated by the professional speaker. The emotions considered were neutral, anger, sadness, happiness, fear, surprise, disgust and teasing. Niimi et at [27] at Kyoto Institute of Technology developed a simulated emotional database in Japanese language for the emotion anger, sadness and joy. The database consists of VCV (vowel-consonant vowel) segments expressed in different emotions. The VCVs were collected from 400 linguistically unbiased utterances, uttered by acoustically trained male speakers. Twelve participated in the subjective test and they recognized each emotion with 84% accuracy. Lida and Campbell [28] at ATR

Laboratories constructed a simulated database in Japanese language. Special care was given to ensure that emotions are not exaggerated. The database consists of monologue texts collected from newspapers, webs, self-published autobiographies of disabled people, essays and columns. Some emotionally biased texts were added in appropriate places in order to enhance the expressiveness of each emotion. The emotions considered were– joy, anger and sadness. One non-professional male and one non professional female speaker were selected for recording. The subjective tests were conducted by 18 university students on emotionally neutral emotions. It was observed that emotions were correctly recognized at significant level. Mozziconacci [29][30] collected a simulated emotional database in Dutch language to study the relationship between speech rate and emotion. The speech material consists of 315 utterances from 3 speakers (2-male and 1-female). Each speaker uttered 5 sentences which were semantically neutral. The emotions considered were neutral, joy, boredom, anger, sadness, fear, indignation. Two intonation experts labelled the emotions. Subjective evaluation of the truth values of these labelling were established with the participation of twenty-four subjects. Montero et al [31] constructed Spanish Emotional speech database (SES). The corpus consists of 3 short sentences and 30 isolated words recorded by a professional male actor in an acoustically treated studio. The emotions present in the database were– neutral, happy, surprised, sad and angry. The labelling of the emotion was semi-automatic and it was judged 85% correctly by 15 subjects. Engberg et al at centre for person Kommunikation at Aalbra University recorded Danish Emotional speech (DES) database [32]. The database consists of 5 simulated emotions. The database was recorded with the participation of 4 actors, familiar with radio theatre (2 male, 2-female). The emotions present in the database are– neutral, surprise, happiness, sadness and anger. The database was evaluated by twenty subjects with 67% accuracy. Abelin et al [33] recorded Sweden Emotional database. The database was recorded with single male speaker. The speaker was not a professional actor. The emotions considered were anger, fear, joy, sadness, surprise, disgust, dominance and shyness. Emotionally neutral sentences were considered for recording. The utterances were evaluated by 35 native Swedish and 78 Swedish immigrant speakers to evaluate the cross-lingual interpretation of emotions. Russian language Affective Speech (RUSSLANA) was collected at Meikai University in Japan by Makarova and Petrushin [34]. The database consists of 3660 sentences from 61 speakers (12-male and 49 female). The database consists of 10 sentences with different syntactic, structural and discoursal types, which were read by 61 (12-male, 49 female) native speakers of Russian portraying six emotions namely– neutral, surprise, happiness, anger, sadness and fear. Indian Institute of Technology Kharagpur Simulated Emotional Speech Corpus (IITKGP-SESC) was recorded using 10 professional radio artists (5-male and 5 female)[35] in Telugu language of India. Fifteen emotionally neutral sentences were considered for recording. Each of the artists had to speak 15 sentences in 8 basic emotions– anger, compassion, disgust, fear, happy, neutral, sarcastic and surprise. Each speaker was recorded for 10 times in 10 different sessions for all the emotions. The total number of utterances in the database is 12000. The subjective evaluation was conducted by 25 students of the Institute. The average emotion recognition rate for male and female speakers were 61% and 66% respectively. Belfast Naturalistic database was recorded by Cowie et al[36] at Queen's University of Belfast in English language. Two kinds of recordings were taken. One was recorded in the studio and the other directly from the TV programs. The clip length is taken to be quite long in order to study the emotional dynamics of speech episodes. The database consists of 298 audio visual clips from 125 speakers (31 male and 94 female). For each speaker there is one clip which the selector

judged relatively neutral and one judged relatively emotional. Clips from the first 100 speakers, totaling 86 minutes of speech, were labelled psychologically and acoustically. The clips were stored in MPEG files with audio data extracted into .wav format. The studio recording part consists of conversations between students, who knew each other well. The topics used for conversation were usually about their past experiences for a particular situation. The TV recordings were done from chat shows, religious programs, programmes tracking individual lives over time and current affairs. Indian Institute of Technology Guwahati, Department of Electronics and Electrical Engineering has developed a simulated stressed speech database in Hindi language of India called Speech Under Simulated Stress Condition (SUSSC)[37]. It contains isolated words and continuous text that does not evoke any stress by themselves. The database consists of three induced stresses– angry, sad and Lombard along with neutral speech. The database was recorded by 10 male and 8 female non-professional speakers. Fifteen (7-male, 8-female) listeners, not participating in the recording process and fluent in Hindi evaluated the stress associated with the utterances. Listeners were allowed to listen the recording until their decision was final. The Speech Under Simulated and Acted Stress (SUSAS) database is a collection of utterances recorded under simulated and actual stress conditions [38]. The database consists of single-word utterances from 9 male speakers, spoken in eleven styles: angry, clear, cond50, cond70, fast, Lombard, loud, neutral, question, slow and soft. The cond50 and cond70 styles are recorded by engaging the speakers in tracking tasks under different levels of workload. The Lombard class was recorded from the speakers listening to pink noise presented binaurally through headphone at 85 dB SPL. Subjective evaluation of the database was conducted and the listeners identified the question speaking style at 85% accuracy whereas the other styles were identified only at 53% of times. From the above study it has been observed that:

(a) Though natural emotional database is the most appropriate database for emotional speech processing, creating a natural emotional database is a difficult process as there are lots of legal and moral aspects associated with it. Further, such a database cannot be recorded in a controlled environment. Therefore, there are always issues like ambient conditions, voice quality etc.

(b) Simulated or acted emotional database is a good source of expressive emotional speech material. When professionally trained actors are used to express the emotions, they express the emotions in an intense manner, considering all aspects which are required for expressing a particular type of emotion. Such emotions are very rarely available in natural conversations. Simulated database recorded by professionally trained actors may be considered as a good source of reference material for emotional speech. However, for practical consideration, such database has little applicability. However, if non-professional actors are used to express the emotions, because of their inexperience, they will not be able to consider the aspects which are required for expressing a particular emotion. As a result, the emotion will not be fully blown and will be more closure to the real world emotions.

(c) Emulated emotional databases are created by involving the subject into emotionally intense situations without their knowledge. However, due to their consciousness about the fact that their voices are being recorded, the subjects may not express their emotion freely. However, the transient nature of human concentration can be suitably exploited in this regard. If a subject is involved in a long-term conversation with the anchor, after a given point of time, the subject will start to respond freely. The anchor will have to closely monitor the psychological state of the subject. Once the recoding is over, the anchor along with the participation of some selectors judged the

portions which contain naturally expressed emotion from the recordings.

(d) Subjective evaluation of the emotional databases was conducted to ascertain the emotion class present in a speech recording. Normally subjective evaluations were conducted by listening tests. Playback of the recordingswere made to belistened by a group of subjects and they assigned each recording to an emotion class. In most of the cases,a forced assignment strategy is used. Therefore, each subject must have to assign one and only one emotion class to each recording from a group of predefined emotion classes. The subjects may be professional experts or naïve listener. If the professional experts are used, they may exploit their heuristic knowledge in labelling the recordings. Therefore, naïve group of listener is a good option for subjective test. Further, the socio cultural and linguistic knowledge of the subjects also play important role in their recognition accuracy. If the subjects are familiar with the language of the recordings, they may use the linguistic contents of the recording to make a decision about the emotion present in it. Again, if the socio-cultural background of the subject and the speaker are completely different, they may wrongly interpret the underlying emotion of a recording.

To study the impact of emotional variability on speaker recognition task, AMLED was recorded in four emotional conditions, namely Neutral, Happy, Angry and Sad. Each speaker was recorded for three different languages English, Hindi and a local language which must be the mother tongue of the informant. Each recording was of 4-5 minutes duration. Speech data were recorded in parallel across microphone and portable voice recorder. The speech samples were recorded in reading and conversation mode. The speech data collection was done in controlled acoustical environment. To record the neutral voice from the speaker in English, Hindi and mother language, the speaker was asked to read a story from the school book or newspaper. In the next session, the speakers were requested to act on some emotionally neutral scripts to simulate the angry, happy and sad emotion in their mother language. Finally, to emulate the emotion, the speakers were shown some video clips to emulate the required emotion and at the same time the anchor talks to the speakers over Skype from a remote location to emulate different types of emotions. From the entire recordings, 3-4 minutes portion that contains the desired emotion were identified and selected. The Table.1 gives the specifications for the database

4. Recording and Validation of the Database

Table 1 Recording specifications for the AMLED database

Number of Speakers	100 (Male=58, Female=42)
Number of sessions	3
Intersession interval	At least 1 week
Data types	Speech
Type of Speech	Conversation & reading
Sampling rate	16 KHz
Sampling format	Mono 16 bits resolution
Applications	Multiple
Speech Duration	3-4 minutes per sample
Acoustic environment	Controlled environment
Language	Multilingual

4.1 Subjective Listening Test of the Database

To establish the validity of the database, especially to ascertain whether a particular segment that has been kept in the database to represent a particular emotion actually convey that emotion or not, human perceptual listening test has been conducted. From each recorded speech file, a 20-second segment was extracted and assigned a serial number to conceal the actual emotion associated with the file from the human evaluators. Each segment were played to 15 persons, 5 from the same linguistic group and remaining from the same socio-cultural background but of different linguistic groups. Each listener identifies the emotion associated with the audio clips. Only those files which were identified correctly by more than 60% of listeners were considered for further processing.

5. Preliminary Investigationof the database

A baseline emotion recognition system has been developed to analyze the quality of emotions present in the AMLED database. We utilized two feature vectors, namely Mel Frequency Cepstral Coefficient (MFCC) and prosodic features. A 20-bank mel-filter bank was used to extract the 19-dimensional cepstral coefficients. To account for the time-varying characteristics of the speech signal, we included the first-order and second-order derivatives, resulting in a 57-dimensional MFCC feature vector. We extracted the fundamental frequency (F0) and energy (E) of each frame to form the prosodic features along with their first- and second-order derivatives, which were concatenated with the MFCC feature vector. Thus, we got a 63-dimensional feature vector. To model the emotions, we developed a Gaussian Model (GMM) based emotion recognizer, trained with 60% of the speech data and tested with the remaining 40%. We utilized 256 mixtures and diagonal covariance matrices for the GMM models. The Expectation-Maximization (EM) method has been used to train the models and convergence criterion is set with a tolerance of 1e-6 and a maximum of 1000 iterations with a regularization parameter of 1e-6 to positive definite covariance matrices.The model performance was evaluated through Receiver Operating Characteristic (ROC) and Area under Curve (AUC) parameters, demonstrating good discrimination for AUC of 0.98 and an Equal Error Rate (EER) of 0.0583. The ROC curve is shown in Fig. 1.

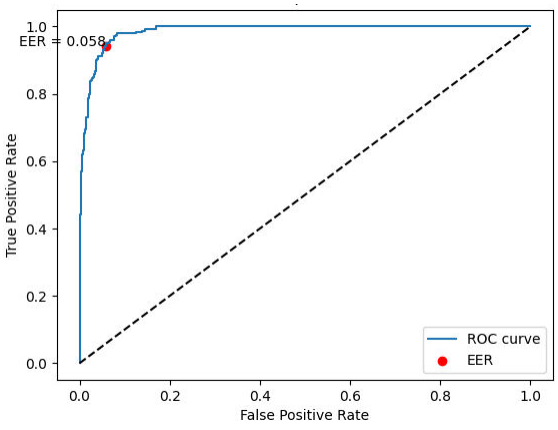

Fig. 1: ROC Curve and Equal Error Rate (EER) for the GMM-based Emotion Classifier

The performance of the GMM based recognizer was further evaluated using the parameters accuracy, precision, recall and F1-score. The confusion matrix is shown in Fig. 2 and the corresponding performance report in Table-2.

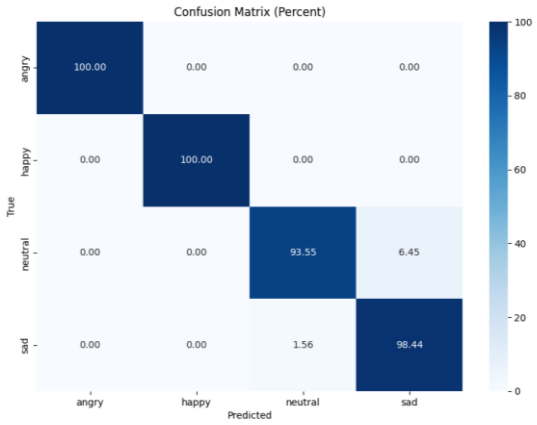

Fig. 2: Confusion matrix for the GMM-based Emotion Classifier

Table 2: Overall Performance of the GMM for Emotion Recognition

Metric	Score (in %)
Accuracy	97.90%
Precision	98.08%
Recall	97.98%
F1-score	97.63%

The GMM model performed well for emotion recognition with the AMLED database, with an overall accuracy of 97.9%. High precision, recall and F1 scores indicate that the model is highly effective in distinguishing between different emotions. At the same time, it reasserted the comprehensive representation of emotion by the AMLED database. The confusion matrix indicates that the model has a very small number of misclassifications.

6. Conclusion

The efficacy and robustness of emotion representation in the Arunachali Multilingual Emotional Database (AMLED) were demonstrated through its evaluation utilising a Gaussian Mixture Model (GMM) for spoken emotion recognition. The AMLED, developed to capture the angry, happy, sad and neutral expressions of the multilingual people in Arunachal Pradesh, has undergone both subjective evaluations and quantitative analysis using the GMM-based model. The confusion matrix and overall accuracy statistics demonstrate the excellent performance of the GMM in accurately identifying emotions from the AMLED dataset. The database demonstrates a high level of reliability for emotion classification tasks, with an impressive overall accuracy of 97.9% and flawless precision, recall, and F1-scores for the Angry and Happy emotions. The tiny misclassifications observed in the confusion matrix between Neutral and Sad emotions indicate small areas where improvement could be made, although they do not have a significant negative impact on the overall performance. The subjective assessments conducted in conjunction with the quantitative examinations provide additional evidence of the emotional representation capability of the AMLED. The tests have verified that the database effectively captures and communicates the desired emotions, which is consistent with the excellent performance metrics acquired from the GMM reviews. The AMLED has been confirmed as a very efficient tool for recognizing speech emotions, capable of delivering precise and dependable emotional information. The database's strength in emotion representation is highlighted by the

combination of subjective validation and quantitative analysis, making it a useful asset for future study and applications in emotion identification systems.

References

1. Bishnu S Atal. Automatic recognition ofspeakers fromtheirvoices. *Proceedings of the IEEE*, 64(4):460–475,1976.
2. MariusVasileGhiurcau,CorneliuRusu,andJaakko Astola. A study of the effect of emotional state upon text-independent speaker identification. In *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 4944– 4947. IEEE,2011.
3. O Douglas and O Shaughnessy. Speech communications: Human and machine. *IEEEpress, Newyork*, pages 367–433,2000.
4. Joy Nicholson, Kazuhiko Takahashi, and Ryohei Nakatsu. Emotion recognition in speech using neural networks. *Neural computing & applications*, 9(4):290–296,2000.
5. Björn Schuller, Gerhard Rigoll, and ManfredLang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–577. IEEE, 2004.
6. Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837,2000.
7. Javier G Rázuri, David Sundgren, Rahim Rahmani, Antonio Moran, Isis Bonet, and Aron Larsson. Speech emotion recognition in emotional feedbackfor human-robot interaction. *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, 4(2):20–27,2015.
8. Hooman Aghaebrahimi Samani and Elham Saadatian.Amultidisciplinaryartificialintelligence model of an affective robot. *International Journal of Advanced Robotic Systems*, 9(1):6,2012.
9. Javier G Rázuri, Pablo G Esteban, and David Ríos Insua. An adversarial risk analysis model for an autonomous imperfect decision agent. In *Decision Making and Imperfection*, pages 163–187. Springer, 2013.
10. Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587,2011.
11. K Sreenivasa Rao and Shashidhar G Koolagudi. *Emotion recognition using speech features*. Springer Science & Business Media,2012.
12. Jianhua Ma, Hai Jin, Laurence Tianruo Yang,and Jeffrey JP Tsai. Ubiquitous intelligence and computing. In *Third International Conference, UIC*, volume 4159. Springer,2006.
13. Chul Min Lee and Shrikanth SNarayanan. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303,2005.
14. John HL Hansen and Douglas A Cairns. Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments . *Speech communication*, 16(4):391– 422,1995.
15. Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99– 117,2012.
16. Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach.

Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1-2):33–60,2003.
17. Marc Schröder. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*,2001.
18. Carl E Williams and Kenneth N Stevens.Emotions and speech: Some acoustical correlates. *TheJournal of the Acoustical Society of America*, 52(4B):1238– 1250,1972.
19. Daniel Cen Ambrus. Collecting and recording of an emotional speech database. *Maribor, Slovenia: University of Maribor*,2000.
20. Linguistic data consortium (LDC). http://www.ldc. upenn.edu/.
21. CécilePereira.Dimensions of emotionalmeaningin speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*,2000.
22. Mike Edgington. Investigating the limitations of concatenative synthesis. In *Fifth European Conference on Speech Communication and Technology*,1997.
23. Thomas S Polzin and Alex Waibel. Detecting emotions in speech. In *Proceedings of the CMC*, volume 16,1998.
24. Felix Burkhardt and Walter F Sendlmeier. Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Tutorial and Research Workshop (ITRW) on speech and emotion*,2000.
25. Miriam Kienast and Walter F Sendlmeier. Acoustical analysis of spectral and temporal changes in emotional speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*,2000.
26. Ryohei Nakatsu, Joy Nicholson, and Naoko Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 343–351,1999.
27. Yasuhisa Niimi, Masanori Kasamatsu, Takuya Nishinoto, and Masahiro Araki. Synthesis of emotional speech using prosodically balanced vcv segments. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*,2001.
28. Akemi Iida, Nick Campbell, Soichiro Iga, Fumito Higuchi, and Michiaki Yasumura. A speech synthesis system with emotion for assisting communication. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*,2000.
29. MarcSchröder.Experimentalstudyofaffe ctbursts. *Speech communication*, 40(1-2):99–116, 2003.
30. SylviaJLMozziconacciandDikJHermes.As tudy ofintonationpatternsinspeechexpressin gemotion orattitude:productionandperception.*IP OAnnual Progress Report*, 32:154–160,1997.
31. JavierMMontero,JGutiérrez-Arriola,JoséColás, Emilia Enriquez, and José Manuel Pardo. Analysis andmodellingofemotionalspeechinsp anish.In *Proc. of ICPhS*, volume 2, pages 957–960,1999.
32. Inger Samso Engberg and Anya Varnich Hansen. Documentation of the danish emotional speech database des. *Internal AAU report, Center for Person Kommunikation, Denmark*, page 22,1996.
33. Åsa Abelin and Jens Allwood. Cross linguistic interpretation of emotional prosody. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*,2000.
34. Veronika Makarova and Valery A Petrushin. Ruslana: A database of russian emotional utterances. In

Seventh international conference on spoken language processing,2002.

35. Shashidhar G Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K Sreenivasa Rao. Iitkgp-sesc: speech database for emotion analysis. In *International conference on contemporary computing*, pages 485–492. Springer, 2009.
36. Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. Anewemotiondatabase:considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*,2000.
37. SumitraShukla.*SpectralAnalysisofStressedSpeech for Speech Recognition*. PhD thesis,2014.
38. Robert S Bolia and Raymond E Slyh. Perception of stress and speaking style for selected elements of the susas database. *Speech Communication*, 40(4):493– 501,2003.