INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Feature Extraction of Assamese Speech Based on Emotion Analysis

Parismita Gogoi[1]*, Debashree Sharma[1], Rosy Bordoloi[1], Snigdha Sarma[1], Ananya Goswami[1]

1 Department of Electronics and Communication Engineering, DUIET, Dibrugarh University, Assam, India

Abstract

Objectives: The present work aims to investigate the recognition of emotion from Assamese speech. **Methods:** This work presents a method based on the Gaussian Mixture Model (GMM) classifier and Mel-frequency cepstral coefficients (MFCC) as feature extraction technique for emotion recognition from Assamese speeches. **Findings:** We have conducted experiments considering different emotions: Angry, Happy, Neutral and Sad. The speech emotion recognition system database is the emotional speech samples collected manually from 20 speakers and some standard samples available on the internet. The speakers are from different districts of Assam and use different dialects of the Assamese language, such as Western (Kamrupi), Central, and Eastern. They fall under the age group of 18-26 years. The field survey consists of recordings done at Dibrugarh University and outside the campus. After the GMM training and testing process, the accuracy we obtained is 51.25%. The experiments confirmed that angry and happy emotions have high energy in the higher frequency region. In contrast, neutral and sad emotions have low energy in the higher frequency region. **Novelty:** This work will help predict the attitudes and actions of different speakers based on their manner of speaking. In addition, the present work will also help in other aspects of human-machine interaction in our daily life. The Assamese emotional speech database used in the work is self-collected from different dialect groups to understand the variability of emotions in dialectal perspective.

Keywords: Assamese; GMM; emotion; speech; MFCC

1 Introduction

The goal of the branch of research known as Speech Emotion Recognition (SER) is to recognize and categorize the emotions that are communicated in speech signals by human speakers. Speech emotion can be considered as a similar type of stress on every sound event during the speech. An emotive speech describes specific prosody in speech[1]. A language's prosodic rules change throughout time as a community's culture does. Speakers also have their speaker-dependent style, which includes an unprecedented pace of articulation, habit of intonation, and loudness trait. As a result,

the speaker's community, culture, language, gender, age, education, social standing, health, physical activities, etc., affect how much emotion is communicated and implied in a speech. SER has many uses, including improving human-computer interaction, diagnosing mental illness better, and giving feedback for speech therapy. Feature extraction, feature selection, and emotion classification are the three fundamental parts of SER systems. The process of extracting features from a raw voice signal that can be used to identify emotions is known as feature extraction.

The process of choosing a subset of characteristics that is most informative and discriminative for emotion recognition is known as feature selection. Assigning an emotion label to a speech segment based on the chosen features is the process of emotion classification.

Speech emotion recognition (SER) is the task of identifying and classifying human emotions from speech signals. SER has many potential applications in fields such as human-computer interaction, affective computing, social robotics, and mental health diagnosis. However, SER is still a challenging problem due to the complexity and variability of human emotions and speech signals. Some of the research gaps in SER are:

- Lack of large-scale, high-quality, and diverse datasets that cover different languages, cultures, and emotional states.
- Lack of robust and generalizable models that can handle noisy, spontaneous, and natural speech in real-world scenarios.
- Lack of explainable and interpretable methods that can provide insights into the underlying mechanisms and features of emotion recognition.
- Lack of evaluation metrics and benchmarks that can measure the performance and reliability of SER systems across different domains and tasks.
- Lack of ethical and social considerations that can address the potential risks and challenges of SER systems in terms of privacy, security, fairness, and accountability.

These research gaps pose significant obstacles for the advancement and adoption of SER systems. Therefore, more efforts are needed to address these challenges and to explore new directions and opportunities for SER research. Acoustic, linguistic, and prosodic aspects are only a few examples of the various features that can be employed for SER. Pitch, intensity, and spectral characteristics of the voice signal are examples of physical qualities from which acoustic features are formed. Language elements like words, phrases, and grammar are derived from the content and organization of speech. Stress, intonation, and rhythm are only a few examples of the changes and patterns seen in speech signals that give rise to prosodic elements.

Additionally, there are other sorts of emotion categorization techniques that can be used to SER, including rule-based techniques, statistical techniques, and deep learning techniques. Rule-based techniques rely on predetermined rules that, using domain or expert knowledge, connect features to emotions. Statistical techniques rely on probabilistic models that use machine learning algorithms on labelled data to learn the association between attributes and emotions. Deep learning techniques are used to train neural networks to learn intricate and nonlinear mappings between features and emotions from massive amounts of data. Speech data gathered from actual scenarios is far more applicable than speech data from acted situations. The recordings of radio news broadcasts of significant events are a well-known example. These recordings include utterances that portray emotions in a very genuine way. Unfortunately, specific ethical or legal concerns can limit their use in the study. As in most available databases, sound laboratories can also induce emotional sentences. The idea that performed emotions is not the same as real ones have long been challenged. Most databases include the following emotions: surprise, neutral, joy, sadness, and rage.

1.1 Basics of emotion recognition

Emotion is a vague, arbitrary concept by nature. Various people have used the word "emotion" in various contexts. Since emotion is a unique mental state that develops unconsciously rather than through conscious effort, it is challenging to define emotion objectively. As a result, there is no universally accepted definition of what constitutes an emotion. This is the main obstacle to moving forward with a scientific approach to study.

No common speech corpora can be used to compare the effectiveness of research methods for emotion recognition. Real-life emotions are widespread and underlying, yet most emotional speech systems are produced utilising fully expressed emotions. While some databases are created by expert artists, others by semi- or inexperienced subjects. Since most databases do not provide a large variety of emotions, the study of emotion recognition is only focused on 5–6 emotions. Speaker and language-dependent information may impact emotion identification systems created utilising various features. In theory, speech emotion recognition systems should be independent of speaker and language.

Finding appropriate features that accurately define various emotions is a crucial problem in developing spoken emotion detection systems. In addition to characteristics, appropriate models must be found in order to extract emotion-specific data from extracted speech features. Systems for recognizing emotions in speech should be capable of processing noisy and real-world speech.

This has given rise to a brand-new area of study called automated emotion recognition, whose fundamental objectives are to comprehend and retrieve desirable feelings. Speech signals serve as a valuable source for affective computing due to a number of intrinsic benefits. For instance, speech signals may typically be obtained more quickly and affordably than many other biological signals (such as the EKG). Because of this, most researchers are drawn to speech-emotion recognition (SER)[2]. For the SER system to be successful, the following three challenges must be addressed:

- Selecting an appropriate emotional speech database.
- Identifying useful features, and
- Creating trustworthy classifiers using machine learning methods.

In actuality, the SER system's biggest problem is extracting emotional features. Important speech components that convey emotion have been proposed by numerous researchers, including energy, pitch, formant frequency, linear prediction cepstrum coefficients (LPCC), and mel-frequency cepstrum coefficients (MFCC)[3]. As a result, most researchers favour merging feature sets, which are made up of a variety of features and contain more emotional information. Speech normalization removes dc and noise components before performing feature extraction and selection. The extraction and selection of features from the speech are the most crucial steps in the additional processing of the input speech signal to detect emotions. Typically, the speech signal analysis in the temporal and frequency domains yields speech features. The database is then created to train and test the speech features retrieved from the input speech signal. The classifiers pick up on emotions in the last stage. The classifier uses various pattern recognition algorithms (HMM, GMM) to identify the emotion[4].

1.2 Review of previous works on SER

We give an overview of the key issues, approaches, and datasets in SER research in this literature review. To define and depict emotions in SER in a consistent and meaningful manner is one of the primary problems. Emotions are dynamic and complicated phenomena that vary depending on a variety of elements including environment, culture, personality, and individual characteristics. As a result, getting trustworthy and accurate speech emotion annotations from human raters is challenging.

The ability to collect and choose pertinent features from speech signals that can effectively capture emotional information presents another barrier for SER. A variety of features, including acoustic ones (such as pitch, intensity, duration, and spectral features), linguistic ones (such as words, phrases, and prosody), and paralinguistic ones (such as laughter, sighs, and hesitations), have been employed in SER. Some elements may be redundant or useless for SER, while some features are more instructive for specific emotions than others. As a result, strategies for feature selection and dimensionality reduction are frequently used to enhance the functionality and effectiveness of SER systems.

Designing and training efficient classifiers that can generalize to different speakers, languages, domains, and scenarios is the third difficulty in SER. Support vector machines, decision trees, hidden Markov models, Gaussian mixture models, artificial neural networks, convolutional neural networks, recurrent neural networks, and attention mechanisms are just a few of the machine learning and deep learning techniques that have been applied in SER. Certain features or activities lend themselves to particular methodologies more than others, and some methodologies may call for significant amounts of labelled data. Authors have proposed two methods: (a) based on a statistical-based parameterization framework for representing the speech through a fixed-length vector and (b) a deep learning approach that combines three convolutional neural networks architectures[1]. Their results achieved 87.08% and 83.90% for RAVDESS and EMOVO datasets respectively.

Some syllable structured Assamese word dataset were made using recorded emotional speech of both male and female of same age group in Assamese language and established the significant consequence of emotions on intonational features such as F_0 contour using PRAAT and MATLAB[5]. The outcomes of their analysis resulted that, emotions have effect on different types of pitch level for various syllable structured word and are different in both male and female speech.

The authors proposed a deep learning model by combining a two-dimensional Convolutional Neural Network (CNN) and a long short-term memory (LSTM) network[6]. They conducted experiments on a customized dataset developed as a combination of RAVDESS, SAVEE, and TESS datasets. Eight states of emotions (happy, sad, angry, surprise, disgust, calm, fearful, and neutral) were considered. The result of this achieved an average test accuracy rate of 90%. Some studies are carried out for the development of an automatic SER system for Indo-Aryan and Dravidian languages[2]. This paper presents a brief study of the prominent databases available for SER experiments. Authors have identified recent relevant literature related to the SER systems' varied design components/methodologies[3]. Some researchers performed experiments for emotion recognition using two models, one with CNN and one without. The accuracy obtained from these two experiments is 61.2% and 62% respectively[4].

1.3 Categories of emotions

Emotions are described in different classes. These classes are (a) Categorical and (b) Dimensional.

- In the Categorical class, Ekman proposed a list of 7 basic emotions: Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral.
- In the Dimensional class, the basic emotions are again classified into three classes: a) Valence: Usually, Happiness has positive valence and anger, and sadness has a negative valence. b) Activation: Sadness has low activation energy, whereas happiness, and anger has high activation energy. c) Dominance: Anger is dominant, whereas fear is dominant.

In this work, we are mainly focusing on the four basic emotions, namely angry, happy, neutral, and sad.
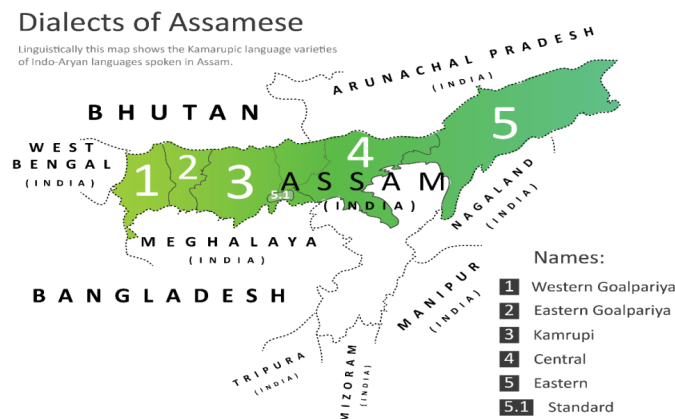

Fig 1. Dialects of Assamese Language on the Assam state map

1.4 The Assamese Language

Eastern Indo Aryan Assamese language is primarily spoken in the Indian state of Assam and is the official language of that state. Although the precise circumstances of its birth and development are still unclear, Assamese has its roots in old Indo-Aryan dialects. While some scholars dispute a strong relationship between Assamese and Magadhi Prakrit, it is generally accepted that Assamese and the Kamatapurilects descended from the Kamrupi dialect of Eastern Magadhi Prakrit that remained to the north of the Ganges. Assamese originated from the Indo-Aryan towns of Kamarupa, surrounded by Tibeto-Burman and Austroasiatic people in urban areas and along the Brahmaputra River. The Brahmaputra Valley, including western and eastern Assam, is home to the Assamese. In the states of Arunachal Pradesh and Nagaland, it is also spoken. Rakhine state of current Myanmar contains Assamese script. Figure 1 shows the map of Assam with the dialects marked across the region.

1.5 Motivation and contribution

The difficult task of speech emotion recognition (SER) involves determining a speaker's emotional state from voice data. Human-computer interaction applications for SER include discussion systems, e-learning, and health care. However, the majority of the SER research that has been done thus so far has concentrated on widely used languages (high resourced) like English, Mandarin, and German, and studies on SER for less resourced languages of Northeast of India like Assamese are scarce.

We examine the prior work on SER for Assamese speech in this article and identify the difficulties and potential directions for further investigation. We start by outlining the fundamental ideas and techniques of SER, including feature extraction, classification, and assessment. Then, we examine the traits and limitations of the Assamese speech emotion recognition databases and corpora that are currently accessible. The most recent approaches and models for SER for Assamese speech, including the Gaussian mixture model (GMM), recurrent neural network (RNN), and i-vector, are then reviewed. Finally, we explore several unresolved challenges and future possibilities for research on SER for Assamese speech, including domain adaptation, multimodal fusion, and cross-lingual and cross-cultural adaptation.

The following are the main contributions of the paper:

- Emotion recognition in speech is one of the most versatile fields for human interaction.

- In this paper, the Assamese language is taken as a special case. The study was done based on a self-prepared database, which emphasizes an important fact towards contributing to this paper.

2 Methodology

2.1 Database preparation

The field survey consists of recordings done on and outside the university campus. A total of 20 persons are recorded till now, among whom ten persons are male, and ten are female. The recordings were taken in two sessions, giving rise to variability in their manner of speaking. They fall under the age group of 18-26 years. Their native locations are Tinsukia, Dibrugarh, Sivasagar, Jorhat, Tezpur, Guwahati, and Nalbari. The mother tongue of all the persons is Assamese. We have collected 10 dialogues from the famous Assamese novel "অসীমতযাৰহেৰালসীমা" (Oximot Jar Heral Xeema) and asked them to read the dialogues in four different emotions namely angry, happy, neutral, and sad. The Assamese dialogues that we have collected are:

1. "মইভাবিছিলো, তইহয়তোআগতেগলিগৈ।"
2. "ব'ল,সৌপুখুৰিপাৰৰঘাঁহনিডৰাতবহোগৈ।"
3. " মোৰক'বলৈএকোনাই, কালিকা- ফালিকাৰকথামইভবাওনাই|"
4. " আইদেৰেৰাৰলগততযোযা, মোৰকিবাদৰকাৰহলেজয়ৰামেইদিবপাৰিব।"
5. " তইমোৰলগতআজিনগলেওহ'বতিলক।"
6. "বেছিৰাতিনকৰিবি,সাঁজলগাৰআগতেউভতিবি।"
7. " আশাকৰোএইবাৰহয়তোলাহেলাহেসকলোচিনিবি।"
8. " তোৰদেখিছোনিজৰবুদ্ধিৰওপৰতঅগাধবিশ্বাস।"
9. "দেখিলোআধলিটোৰদুয়োটাপিঠিয়েইমোৰপৰাসমানআঁৰত।"
10. "লিখাস্থানৰবৰঅভাৱ,গতিকেএকোনাই, চিৰশূণ্য।"

    Inside the university campus, we have recorded in our Electronics and communication Department as well as in Gyanmalini Studio. Other than that, we have also recorded in Guwahati, Jorhat, Dibrugarh and Tinsukia districts. One person has sent us his recordings from Rochester. We have used devices like Zoom h1N and mobile recorders for recording purposes. The detailed list of female and male speakers is tabulated in Table 1 and Table 2.

Table 1. List of female speakers

Speaker	Age	Languages Known	Dialect
AG	23	Assamese, Hindi, English	Eastern Assamese
DS	26	Assamese, Hindi, English	Eastern Assamese
KB	23	Assamese, Hindi, English	Central Assamese
MB	23	Assamese, Hindi, English	Western Assamese (Kamrupi)
ND	23	Assamese, Hindi, English	Eastern Assamese
PD	24	Assamese, Hindi, English	Central Assamese
RB	23	Assamese, Hindi, English	Central Assamese
RB	19	Assamese, Hindi, English	Eastern Assamese
SS	23	Assamese, Hindi, English	Western Assamese (Kamrupi)
TR	23	Assamese, Hindi, English	Western Assamese (Kamrupi)

2.2 Data collection

ZOOM H1n device and mobile recorders were used for single channel recording of 4 emotionally biased utterances of different lengths in each emotion from 10 male and ten female speakers of the Assamese language in a closed-room noise-free environment. For digitization, 16000 Hz of sampling frequency is used. Speech samples were collected for 3 archetypal (full-blown) emotions and also for neutral moods. Each speaker was asked to utter 2 times fixed sets of 10 short sentences with four different emotions. This required emotional acting by the speakers and the sentences' meaning was narrated to sufficiently arouse the same emotion in them. The set of utterances was recorded in two different sessions. In both sessions, the utterances corresponding to angry, happy, and sad were recorded in the same order. The neutral utterances were recorded at the beginning of any of the above sessions.

Table 2. List of Male speakers

| Speaker	Age	Languages Known	Dialect
APK	23	Assamese, Hindi, English	Eastern Assamese
AS	26	Assamese, Hindi, English	Eastern Assamese
BS	23	Assamese, Hindi, English	Western Assamese (Kamrupi)
BB	23	Assamese, Hindi, English	Eastern Assamese
HM	23	Assamese, Hindi, English	Eastern Assamese
KC	24	Assamese, Hindi, English	Western Assamese (Kamrupi)
MPB	23	Assamese, Hindi, English	Eastern Assamese
RB	19	Assamese, Hindi, English	Eastern Assamese
US	23	Assamese, Hindi, English	Eastern Assamese
WRK	23	Assamese, Hindi, English	Eastern Assamese

2.3 Experimental details

Feature Extraction and Selection: The many characteristics that make up a speech indicate every emotion the speaker intends to convey, and any changes to these parameters will cause a corresponding shift in the speaker's emotions[7]. As a result, a key component of a speech emotion detection system is the extraction of these speech components that indicate emotions. The two primary categories of speech features are long-term features and short-term features. The short-term properties, such as formants, pitch, and energy, are what we call short-term features. Additionally, the statistical method for digitising voice signals uses long-term properties. The mean and standard deviation are a couple of the often-utilised long-term features. The categorization procedure will be enhanced the more substantial the feature that is utilised. An important factor that must be taken into consideration when extracting features from voice signals is the area of analysis used.

The speech signal is separated into frames, which are discrete intervals. This experiment captures the speech signal in brief frames with a frameshift of 20ms. The audio barely changes during this time. The features in each frame are identical. Our experiment makes use of the 13-Dimension MFCC. The Mel Frequency Cepstrum Coefficient (MFCC), which has a very high identification rate, is widely utilized in speech recognition and voice emotion recognition systems. As opposed to the high-frequency zone, the low-frequency region could benefit from higher frequency resolution and noise robustness with the aid of MFCC[8]. The Mel Frequency Cepstrum Coefficient demonstrates the short-term power spectrum of sound. Additionally, the Shifted Delta Coefficients (SDC) feature is employed along with the MFCC function. SDC is purchased from the MFCC. SDC features are better suitable for long-term speech signals because they distinguish the speaker's dynamic behaviour and the speech signal's prosodic characteristics[9]. The resilience of channel mismatch, speaking style, and session variability are also determined using this method. A more effective way to identify people's many emotional states is provided by combining the two MFCC-SDC features.

Classifier Selection: After the features are calculated, the best features are given to the classifier in the speech emotion identification system. A classifier can identify the speaker's emotion in their voice utterance. For the job of recognising speech emotions, various classifier types have been suggested. In the literature, the speech emotion identification system uses a variety of classifiers, including the Gaussian Mixtures Model (GMM), K-nearest neighbours (KNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Artificial Neural Network (ANN) etc. Each classifier outperforms and falls short of the others in different ways. The Gaussian Mixture Model is better suited for speech emotion recognition only when the global characteristics are derived from the training utterances, as there are fewer training and testing requirements for GMM. In our experiment, 256-DGMM was employed[10].

Emotional Database: The performance of the emotional speaker recognition depends heavily on the database's quality. Emo-DB (consisting of 535 audio files), RAVDESS (consisting of 1440 audio files), and SAVEE (consisting of 120 audio files) were chosen as the emotional speech corpora for this work[11]. It consists of recordings of dyadic mixed-gender actor pairings in voice, video, and motion capture. The audios have a broader emotional theme. The basic objective is to have an expression that most closely resembles how emotions are expressed naturally. Then, these expressions were broken down into utterances that were manually annotated with categorical labels such as "angry," "happy," "sad," "neutral," "excited," "frightened," "surprised," and "disgusted," as well as in terms of three-dimensional axes such as "valence," "activation," and "dominance." In addition to this normative data, we manually gathered data sets from ten male and ten female speakers who were instructed to speak the ten Assamese sentences in four different emotional tones. So, a total of 1600 datasets are manually collected, of which 1280 datasets are utilized for training and 320 datasets are used for testing. We now have 3695 universal data in total.
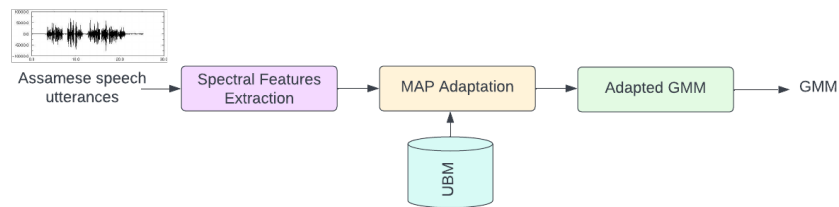
Fig 2. Construction of the GMM model from Assamese emotional utterances

Design of Universal Background Model (UBM): A model of person-specific feature characteristics is compared against a model of general, person-independent feature characteristics using the UBM. It is frequently employed with the discriminative Gaussian mixture model (GMM), where the validity of the sample is assessed by comparing its user-dependent properties to a sample of all other users[12]. The UBM is taught to use large amounts of data derived from a wide variety of speakers, and it is useful in speaker-independent models. For various common databases, including EmoDB, SAVEE, and RAVDESS, we have developed UBM. All phonemes are grouped together by UBM into a cluster, and the characteristics are then extracted to yield the matching mean (m), variance (v), and weight (w).

GMM Training and Testing: The training and testing are done using our own database which was manually collected. The main aim of the training phase is to compute the best parameters to match the distribution of the feature vectors[13]. Training is done using the parameters obtained from the UBM. 80% of the total Assamese data is taken as Training data and the rest 20% is taken as Testing data. The block diagram showing the overall procedure is presented in Figure 2.

During the cluster formation stage, thereissomemean mean values for all the clusters. When training is done, the means of all the clusters are updated. The adaptation of the means of UBM leads to GMM. For four different emotions (Angry, Happy, neutral, and Sad), four GMMs are created, and their corresponding means are recorded. Using the means of four GMMs, Testing is done. The testing phase comprises the log likelihood function, which is used to derive the maximum likelihood estimator of the parameter. A classification problem's prediction outcomes are compiled in a confusion matrix. Count values describe the number of accurate and inaccurate predictions for each class. It demonstrates the ways in which the predictions made by our categorization model are erroneous. It provides us within formation about the mistakes a classifier is making. The percentage of correctly classified data instances over all the data instances is known as accuracy.

Table 3. Confusion Matrix

Prediction→ Ground-truth	Angry	Happy	Neutral	Sad
Angry	**50%**	25%	18.75%	6.25%
Happy	26.25%	**36.25%**	21.25%	16.25%
Neutral	1.25%	25%	**55%**	17.50%
Sad	1.25%	12.50%	22.50%	**63.75%**

3 Results and Discussion

After all the experimental works done by using 13 Dimensional MFCC and SDC and 256 Dimensional GMM, we have obtained the following confusion matrix. The confusion matrix is gender independent. In the confusion matrix shown in Table 3, the columns show the emotions that the speakers tried to induce, and the rows are the percentages of output recognized emotions. The accuracy that we have obtained is 51.25%. From Table 3, it is observed that each of the four emotions is confused with the other three emotions at a less or more amount. For example, angry emotion is highly confused with happy emotion by 25%. Similarly, happy is confused with angry, neutral is confused with happy and sad is confused with neutral at a rate of 26.25%, 25% and 22.5% respectively. This is because test persons have difficulties with feigning certain emotions. Furthermore, prosodic and voice quality features also create confusion among the emotions. Our work has been compared with a previous work reported in [10] modelled using Assamese language with the classifier designed by GMM and MFCCs feature vectors. In that study, it was reported that the utilisation of 39 Mel-Frequency Cepstral Coefficients (MFCC) features and Gaussian Mixture Model (GMM) classifier facilitated the development of diverse emotion recognition systems for distinct speakers, encompassing four emotions

namely, anger, happiness, neutrality, and sadness. Their study yielded a recognition rate of 43.385%. In our study, the accuracy obtained is 51.25% which is better than the previous work reported.

4 Conclusion

The study of speech recognition and machine learning has advanced significantly. However, an incredibly accurate system has not yet been created. This research aims to use speech signals to analyse speech for emotion states (angry, sad, neutral, and happy). The use of various classifiers in speech-emotion recognition systems is illustrated here. The signal processing unit, which extracts pertinent features from the available speech signal, and a classifier, which recognizes emotions from the speech signal, are the two critical components of a speech emotion recognition system. The confusion matrix tables show that some emotions are frequently confused with others. Additionally, some emotions appear to be easier to recognise than others. This can be because all the test phrases played out emotions, and test subjects had trouble acting out emotions. Other features, like prosodic and voice quality elements, are also responsible for the frequently misunderstood emotional states. In this paper, the Assamese language is taken as a special case. The study was done based on a self-prepared database, which emphasizes an important fact towards contributing to this paper. The obtained accuracy was 51.25%. In this experiment, a correct classification rate of less than 60% was attained, demonstrating that while MFCC coefficients are standard features in speech recognition, they are not appropriate for speech emotion recognition alone. We will now attempt to increase performance accuracy in the future by combining several classifier models, such as Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), etc. Additionally, we will work to expand the speech database by gathering a lot of Assamese data from more speakers of all the dialect groups and employ deep learning models to achieve better performance. Additionally, the accuracy of the speech emotion identification system can be improved by extracting additional functional speech elements.

5 Declaration

Presented in Fourth Industrial Revolution and Higher Education (FIRHE 2023) during 23rd-25th Feb 2023, organized by DUIET, Dibrugarh University, India. The Organizers claim the peer review responsibility.

6 Acknowledgement

References

1) Sekkate S, Khalil M, Adib A. A statistical feature extraction for deep speech emotion recognition in a bilingual scenario. *Multimedia Tools and Applications*. 2023;82:11443–11460. Available from: https://doi.org/10.1007/s11042-022-14051-z.
2) Monisha STA, Sultana S. A Review of the Advancement in Speech Emotion Recognition for Indo-Aryan and Dravidian Languages. *Advances in Human-Computer Interaction*. 2022;2022:1–11. Available from: https://doi.org/10.1155/2022/9602429.
3) Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*. 2021;9:47795–47814. Available from: https://doi.org/10.1109/ACCESS.2021.3068045.
4) Horii D, Ito A, Nose T. Analysis of Feature Extraction by Convolutional Neural Network for Speech Emotion Recognition. In: 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), 12-15 October 2021, Kyoto, Japan. IEEE. 2021. Available from: https://doi.org/10.1109/GCCE53005.2021.9621964.
5) Kolita S, Acharjee PB. Analysis on Syllable-Based Intonational Features of Assamese Speech Signals. In: Mathematical and Computational Intelligence to Socio-scientific Analytics and Applications ;vol. 518 of Lecture Notes in Networks and Systems. Springer Nature Singapore. 2022;p. 231–242. Available from: https://doi.org/10.1007/978-981-19-5181-7_18.
6) Singh J, Saheer LB, Faust O. Speech Emotion Recognition Using Attention Model. *International Journal of Environmental Research and Public Health*. 2023;20(6):1–21. Available from: https://doi.org/10.3390/ijerph20065140.
7) Ayadi ME, Kamel MS, Karray F. Survey on speech emotion recog-nition: Features, classification schemes, and databases. *Pattern recognition*. 2011;44(3):572–587. Available from: https://doi.org/10.1016/j.patcog.2010.09.020.
8) Sudhakar RS, Anil MC. Analysis of speech features foremotion detection: a review. In: 2015 International Conference on Computing Communication Control and Automation, 26-27 February 2015, Pune, India. IEEE. 2015. Available from: https://doi.org/10.1109/ICCUBEA.2015.135.
9) Mansour A, Lachiri Z. SVM based Emotional Speaker Recognition using MFCC-SDC Features. *International Journal of Advanced Computer Science and Applications*. 2017;8(4):538–544. Available from: https://pdfs.semanticscholar.org/fd5d/b3ca0d157866259af2c9cfed8a77a6e9bb88.pdf.
10) Kandali AB, Routray A, Basu TK. Emotion recognition fromAssamese speeches using MFCC features and GMM classifier. In: TENCON 2008 - 2008 IEEE Region 10 Conference, 19-21 November 2008, Hyderabad, India. IEEE. 2008. Available from: https://doi.org/10.1109/TENCON.2008.4766487.
11) Ververidis D, Kotropoulos C. A review of emotional speech databases. In: Proceedings of the Panhellenic Conference on Informatics (PCI). 2003;p. 560–574. Available from: http://delab.csd.auth.gr/bci1/Panhellenic/560ververidis.pdf.

12) Hu H, Xu MX, Wu W. GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 15-20 April 2007, Honolulu, HI, USA. IEEE. 2007. Available from: https://doi.org/10.1109/ICASSP.2007.366937.

13) Kaushik R, Sharma M, Sarma KK, Kaplun DI. I-vector based emotion recognition in Assamese speech. *International Journal of Engineering and Future Technology*. 2016;1(1):111–124. Available from: http://www.ceser.in/ceserp/index.php/IJEFT/article/view/4423.