Advances and Applications in Mathematical Sciences
Volume 21, Issue 7, May 2022, Pages 4133-4140

PERFORMANCE EVALUATION OF SPEECH EMOTION RECOGNITION BASED ON PROSODIC FEATURES USING VARIOUS MACHINE LEARNING TECHNIQUES

AKSHAT AGRAWAL and ANURAG JAIN

University School of Information
Communication and Technology
Guru Gobind Singh Indraprastha University
New Delhi 110078, India
E-mail: akshatag20@gmail.com
anurag@ipu.ac.in
akshatag20@gmail.com

Abstract

In this paper, the effect of three different attributes i.e., energy, pitch and Zero crossing rate (ZCR) has been studied. Various local features are gathered, and tests have been performed to validate the effect of above-mentioned attributes in speech. It was found that each of the three attributes individually contribute to identification of various emotions in any speech. The performance of male and female voices for five different emotions viz. angry, happy, sad, surprise and neutral have been evaluated. Further, five different machine learning techniques have been employed for three different data sets viz. RAVDESS, IEMOCAP and Hindi data sets for Male and Female both. It is found that different machine learning techniques employed in the present studies give an accuracy of 85% and 83% for male and female category respectively.

1. Introduction

Natural Speech is very effective and well-organized communication method in human being. It motivates research community to precisely work on speech signal and find out different speech features [1]. While a person communicates via speech with another person, one can use many emotions. Therefore, it is found that emotions are the major and impactful factors in speech communication. On the other hand, while a person conveys to another

2020 Mathematics Subject Classification: 68T20, 93C85.

Keywords: Speech features, local speech features of speech, energy of speech, pitch of speech, duration of speech.

Received October 5, 2021; Accepted December 12, 2021

one through speech, it is observed that mostly both can understand the emotions of each other whereas when one communicates with machines, the machine can easily understand the wording of speaker instead of recognizing their emotions. It is required to analyze the features of different emotions in speech before assessing the emotional state of person. However, speech is having two kind of features viz. local and global features. Primarily, basic feature of speeches like pitch, energy and ZCR are implemented for experimenting [3]. Speech also offers some normal acoustics and articulatory properties for a short interim of time. Since the speaker desires to deliver a sound arrangement comparing to the message to be delivered, most significant vocal tract developments have a deliberate premise. In this area, an exertion has been made to clarify the speech and acknowledgment process which looks like the system of speech creation [4].

In this field, ample amount of research work has been carried out by number of researchers to explore emotion in voice [5]. It is further reviewed that the work on fine pitch contour's structure in an emotion's cue pitch of expressional speech is also investigated in the literature. The important variations in pitch i.e., pitch mean, and range have also been studied [6]. The comparison of paralinguistic and linguistic features of pitch are also reviewed based on pitch shape association of linguistics structure of speech [5]. It is further reviewed in that literature that research on anger specifically much acoustic has been done along with the basic emotions [7]. The effect of angry emotions, including a word long tone content which is also explored and investigated the frequency (F0) contours that appear to remain steadily or decline slightly, however, average duration is smaller for an 'angry' word. [8-13]. In this paper, the importance of speech features in speech communication has been studied. Further, five different machine learning techniques have been employed for three different data sets viz. RAVDESS, IEMOCAP and Hindi data sets for Male and Female both [14]. Section 2 covers speech database and methodology followed by analysis and discussion of results in Section 3. The conclusion and future scope of the study has been covered in Section 4.

2. Speech Database and Methodology

In this paper, three datasets are employed to perform the comparative study of emotion recognition rate based on prosodic features. Two datasets viz

RAVDESS, IECOMAP belong to English language while one dataset belongs to Hindi Language to evaluate the performance of independent language barrier.

2.1 Speech Corpora

First dataset i.e., Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This is one of the most popular datasets among all the available datasets where open-source dataset in English language. This dataset contains 12 female and 12 male voices uttered by drama actors in English language with eight emotions viz: happy, calm, surprised, neutral, angry, sad, disgust and fearful. The total number of sentences are 1440.

The second dataset i.e., IEMOCAP is bit different from previous one. This dataset contains the combination of male and female actors' conversation in natural environment where total ten emotions viz: neutral, disgust, happy, frustrated, surprised, fearful, angry, sad, exited are recorded.

Third dataset is collected by authors due to necessity of Hindi language and sentences. It is reviewed that none of the openly accessible database gives the facility in this corpus were five exclusive fundamental feeling viz neutral, anger, sad, happy and surprise rediscovered. In proposed methodology, five essential feelings are implemented viz neutral, anger, sad, happy and surprise style to make correlation to find out proper correlation. It is found that exaggerated feeling nonprofessional speakers are articulated for 7 sentences in Hindi dialects in five distinct feelings. These articulations are recorded in quite condition in *.wav configuration having an inspecting pace of 16 kHz with an accuracy of ~6 bits per test. The speech corpora can be delivered via exact 4 audiences autonomously find out the expressions. Finally, dismissed expressions are re-recorded and the aforementioned strategy repeats until the rectification of database takes place.

2.2. Feature Selection

Feature selection and extraction is most important process of finding the emotions from the speech. There are 2 different types of speech features which are commonly used to identity the emotion from speech first one is prosodic feature also called as local features of speech and second is spectral features it's called a log power spectrum feature. Prosodic features viz: Pitch, Energy, and ZCR are investigated in the present study.

2.2.1 Pitch

Pitch is one of the import features to find out the emotion from speech. It varies on the sound frequency wave. Usually pitch of male voice is lower than female voice and when a person is in shouting mode viz: Angry and happy then also one's pitch is high and vice-versa.

2.2.2 Energy

Energy belongs to strength of a voice received by the physical ear of human beings. This can be determined by wave amplitude.

2.2.3 Zero Crossing Rate (ZCR)

In ZCR, signal varies from positive bias to negative bias through zero or vice versa.

It identifies loud and small sound and its changes. This can be used to figure out human speech sample availability in sound. ZCR is majorly used in speech recognition and speech synthesis.

2.2.4 Classification for Speech emotion Recognition

Emotion identification is tedious task when identifying directly from the speech. In this study, various machine learning algorithms are used to classify different type of emotions. Machine learning algorithms take input data under training dataset head usually 70% to 80 % data are taken as input and works based upon their own algorithm, further, 20% to 30% dataset are kept for testing purpose to evaluate the performance.

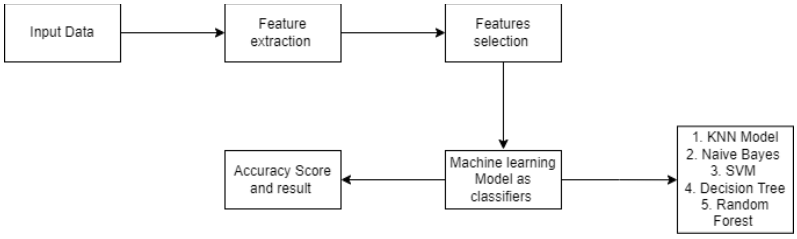

Figure 1. Baseline model for Classification of Emotion.

Figure 1 shows baseline model for classification. First preprocessed speech data should be taken as input, after that feature extraction can be done based on various features viz: prosodic and spectral features. In the

present study, prosodic features are considered due to its simplicity, which is basic feature of speech after feature selection process machine learning (ML) algorithm applied on the datasets in the section of training and testing purpose. Further, accuracies can be compared among different datasets and algorithm.

3. Results and Discussion

Performance evaluation has been done with similar emotion from all the available datasets where 5 basic emotions are considered for all the speech corpora viz: Happy, Angry, Sad, Neutral, Surprise. However, all the available datasets have more than 7 emotions recorded but for maintaining similarity 5 basic emotions are kept in this study.

3.1 Various datasets utilized in the study

Machine learning models are exploited to RAVDESS datasets, which is in English. Additionally, classification based on Male and Female voices are investigated to find more clarity in the results. Furthermore, three features viz: Pitch, Energy, ZCR have been exploited in the present study.

Table 1. Performance of Male and female voice for emotion classification on different ML model and 5 basic emotions for RAVDESS dataset.

RAVDESS Dataset ENGLISH Language										
	Male Voice					Female Voice				
	Naïve Bayes	KNN	SVM	Decision Tree	Random Forest	Naïve Bayes	KNN	SVM	Decision Tree	Random Forest
Angry	39	42	78	89	88	31	43	81	85	86
Happy	32	48	74	82	76	28	52	79	73	82
Sad	34	42	79	77	82	39	76	88	91	92
Surprise	44	56	79	83	77	45	67	89	93	92
Neutral	28	67	87	88	95	56	67	76	87	92

Table 2. Performance of Male and female voice for emotion classification on different ML model and 5 basic emotions for IEMOCAP dataset.

IEMOCAP Dataset ENGLISH Language	
Male Voice	Female Voice

	Naïve Bayes	KNN	SVM	Decision Tree	Random Forest	Naïve Bayes	KNN	SVM	Decision Tree	Random Forest
Angry	31	44	79	85	87	37	47	84	89	91
Happy	29	51	81	89	81	27	58	83	91	93
Sad	39	48	81	82	91	26	66	89	93	87
Surprise	38	51	81	89	84	40	69	91	82	92
Neutral	21	71	91	83	85	35	76	89	91	95

Table 3. Performance of Male and female voice for emotion classification on different ML model and 5 basic emotions for Hindi dataset

HINDI Language Dataset										
Male Voice						Female Voice				
	Naïve Bayes	KNN	SVM	Decision Tree	Random Forest	Naïve Bayes	KNN	SVM	Decision Tree	Random Forest
Angry	35	49	75	89	92	22	41	78	85	95
Happy	27	58	74	89	88	21	47	86	92	89
Sad	29	51	89	88	89	27	61	78	85	95
Surprise	38	67	75	82	90	37	58	92	94	96
Neutral	20	57	85	97	97	25	44	75	87	95

Table 1-3 shows performance of male and female voice for five different emotions viz. angry, happy, sad, surprise and neutral. Further, five different machine learning techniques have been employed for three different data sets. The data sets employed in the study are RAVDESS, IEMOCAP and Hindi data sets. Furthermore, the aforementioned techniques are implemented for Male and Female both.

It is concluded from Table 1 that sad emotions are relatively difficult to determine in male category compared with Female category with decision tree and random forest techniques. Similar pattern of variations is observed with IEMOCAP dataset as can be seen from Table 2, whereas these types of emotions are exploited clearly for male and female in Hindi data set. Exhaustive comparison among five decision making techniques for three data sets have been carried out and it is found that Decision tree and Random Forest ML techniques give comparatively good results for all the three data sets employed in the study for Male and Female Category. It can be concluded that Decision tree and Random Forest techniques give good result for multi class classification, therefore, these two techniques can further be

exploited for other datasets of multi class category. In this study, features of uttered sentences are considered instead of semantics of different data sets. The above-mentioned ML techniques give good result in uttered sentences of three different datasets.

4. Conclusions and Future Scope

In this paper, the effect of three different attributes i.e., the pitch, energy and duration in speech recognition has been studied. Different local features are collected, and tests have been performed to validate the effect of above-mentioned attributes in speech. The major outcomes of the study are summarized as below:

It is found that Decision tree and Random Forest ML techniques give comparatively good results for all the three data sets viz. RAVDESS, IEMOCAP and Hindi data sets exploited in the study for Male and Female Category.

It can be concluded that Decision tree and Random Forest techniques give good result for multi class classification, therefore, these two techniques can further be employed for other datasets of multi class category.

The study is useful to establish clearly primary emotions among 5 classes and helpful for evaluating the emotion identification performance.

The different machine learning techniques employed in the present studies give an accuracy of 85% and 83% for Male and Female category respectively.

The proposed machine learning techniques can further be exploited for other data sets. The performance of chosen data sets can further be evaluated with different deep learning techniques with spectral features of emotion recognitions. Further, the combination of spectral and prosodic features can be assessed based on different machine learning techniques for several data sets.

References

[1] A. Agrawal and A. Jain, Speech emotion recognition of Hindi speech using statistical and machine learning techniques, Journal of Interdisciplinary Mathematics 23(1) (2020), 311-319.

[2] L. S. Chen and T. S. Huang, Emotional expressions in audiovisual human computer interaction, in 2000 IEEE International Conference on Multimedia and Expo, ICME2000, Proceedings, Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532) 1 (2000), 423-426.

[3] G. Shashidhar, K. Koolagudi and R. Sreenivasa, Emotion recognition from speech: a review, Springer Sci. Bus. Media 15 (2012), 99-117.

[4] A. Jain, Acoustical analysis of emotions in Hindi speech and their transformations using prosodic information, (PhD. Thesis), 2014.

[5] P. Lieberman and S. B. Michaels, Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech, J. Acoust. Soc. Am. 34(7) (1962), 922-927.

[6] T. Bänziger and K. R. Scherer, The role of intonation in emotional expressions, Speech Commun. 46(3-4) (2005), 252-267.

[7] S. Chuenwattanapranithi, Y. Xu, B. Thipakorn and S. Maneewongvatana, The roles of pitch contour in differentiating anger and joy in speech, Int. J. signal Process. 3 (2006), 129-134.

[8] J. W. Mullennix, T. Bihon, J. Bricklemyer, J. Gaston and J. M. Keener, Effects of variation in emotional tone of voice on speech perception, Lang. Speech 45(3) (2002), 255-283.

[9] Livingstone R. Steven and Frank A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, PloS one 13(5) (2018), e0196391.

[10] Busso Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee and Shrikanth S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Language resources and evaluation 42(4) (2008), 335-359.

[11] D. Virmani, C. Gupta, P. Bamdev and P. Jain, iSeePlus: A cost effective smart assistance archetype based on deep learning model for visually impaired, Journal of Information and Optimization Sciences 2 41(7) (2020), 1741-56.

[12] Charu Gupta, Prateek Agrawal, Rohan Ahuja, Kunal Vats, Chirag Pahuja and Tanuj Ahuja, Pragmatic analysis of classification techniques based on hyperparameter tuning for sentiment analysis, International Semantic Intelligence Conference (ISIC'21), Delhi (2021), 453-459.

[13] Prateek Agrawal, Deepak Chaudhary, Vishu Madaan, Anatoliy Zabrovskiy, Radu Prodan, Dragi Kimovski and Christian Timmerer, Automated bank cheque verification using image processing and deep learning methods, Multimedia tools and applications (MTAP) 80(1) 5319-5350. https://doi.org/10.1007/s11042-020-09818-1

[14] L. Jain and P. Agrawal, English to Sanskrit transliteration: an effective approach to design natural language translation tool, International Journal of Advanced Research in Computer Science 8(1) (2017), 1-10.