ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 6|| Jun – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Comparative Study Of Various Classifiers Using Assamese Speech Utterances

G.R.Michael[1], N.K.Kaphungkui[2], Piyush Kr. Thakur[3]
1 Department of ECE, Dibrugarh University,
2 Department of ECE, Dibrugarh University,
3 Department of ECE,Dibrugarh University.
Corresponding Author: roberteld008@gmail.com[1]

ABSTRACT:
Emotion Recognition from speech is a recent 'state of the art' topic of research in the Human Computer Interaction (HCI) field. The need has risen for a more robust communication interface between humans and computer, as computers have become an integral part of our lives. A lot of work is currently going on to improve the interaction between humans and computers. To achieve this goal, a computer should be able to differentiate its present situation and respond differently corresponding to the observation. To make the human computer interaction more robust, the objective is that computer should be able to recognize emotional states in the same way as we humans do. The efficiency of the emotion recognition system depends on type of features extracted and the classifiers used for detection of emotions. The proposed system aims at identification of basic emotional states such as anger, surprise, neutral, fear, happy and sadness from human speech. While classifying different emotions, features like MFCC (Mel Frequency Cepstral Coefficient) is used [3]. For training and testing, data is collected from various Assamese movies of one short emotionally biased sentence. The emotion recognition process consists of an audio feature extraction module followed by implementation of different binary classifiers for emotion (represented by a binary string) classification. This paper compares the performances of different classifiers.
KEYWORDS: speech recognition, speech Emotion Recognition, MFCC, Gradient boosting, Logistic regression, SVM, Random Forest.

Date of Submission: 15-06-2018 Date of acceptance: 30-06-2018

I. INTRODUCTION:

Emotion recognition through speech is an area which increasingly attracting attention between the engineers in the field of pattern recognition and speech signal processing in recent years. Emotion recognition plays an important role in identifying emotional state of speaker from speech signal. Emotional speech recognition aims at automatically identifying the emotional state of a human being from his or her voice as input. The emotional states of a speaker are known as emotional aspects of speech and are included in the so-called paralinguistic aspects. Accurate detection of emotion from speech has clear benefits for the design of more smooth and natural human- machine speech interfaces or for the extraction of useful information from large quantities of speech data. It is also becoming more and more important in computer application fields as health care, children education, etc. In speech-based communications, emotion plays an important role [3].

II. EMOTION RECOGNITION PROBLEM DESCRIPTION

This paper is dedicated to speech signal processing for detection of emotion. The main goal is to learn how the classification of speech by the emotional state of the speaker. Unfortunately, there is no strict definition to what is emotion. Moreover, it will be shown further, different people classify emotion of speech differently. The second difficulty is about time properties of speech signals. Often the utterances have no emotion (i.e. speaker is in neutral state), but emotionality is contained in a few words or phonemes of the utterances.
Emotion recognition problem can be reformulated in mathematical terms as a classification task. In brief a function from the utterances space to the set of emotional states has to be constructed. In this space decision rule separates utterances with one emotion from the others. But what if people evaluates utterances differently? If we

assume that the utterances emotional states are picked from an unknown probability distribution, then the model should predict and process these probabilities in the most sensible way [1].

Another problem is the availability of a proper labelled speech corpus. The thinly populated communities fail to keep the purity of their mother tongue. India has thousands of ethnic communities with distinct languages and dialects. Assam is also inhabited by a number of ethnic group with distinct languages e.g. Assamese, Bishnupriya Manipuri, Bodo, Deori, Dimasa, Garo, Karbi, Koch, Kuki, Mishing, Rabha, Tai-Ahom, Tiwa, etc. The mother tongues of some of the communities of Assam, have already become extinct, e.g. Lalung, Chutia, Sonowal, etc [4].

The present work aims to investigate the recognition of emotion from Assamese speech. In addition, the present work will also help in other aspects of human-machine interaction, in our daily lives such as healthcare, law, business, etc.

III. DATA DESCRIPTION

Regardless of model type, training procedure requires a labelled emotional corpus. There are quite a few databases available in the internet nowadays. We carried out all experiments with audio data from emotion biased dialogues of Assamese movies.

It consists of about 2 hours of audio data from 4 Assamese movies. All recordings have a structure of dialogue between a man and a woman that are scripted or improvised on the given topic. After collecting the audio signals we divided the dialogues into small utterances of length mainly from 2 to 10 seconds and then the signals are evaluated. Then we had to evaluate each utterance based on both audio streams. The evaluation consisted of 10 options (neutral, happiness, sadness, anger, surprise, fear, disgust frustration, excited, other). Here we take for analysis only 6 of them — anger, fear, neutral, happy, surprise and sadness.

Emotion was assigned to the utterance only if at least half of our team were consistent in our choice. And it is not always like that. There is a 30% of utterances in which more than a half of us gave different labels and emotion was not assigned at all. This fact illustrates that emotion is a subjective notion and there is no way to classify emotions precisely even if humans can't do that. In other words, any model cannot learn to recognize emotions, but can learn how experts label emotional utterances [1].

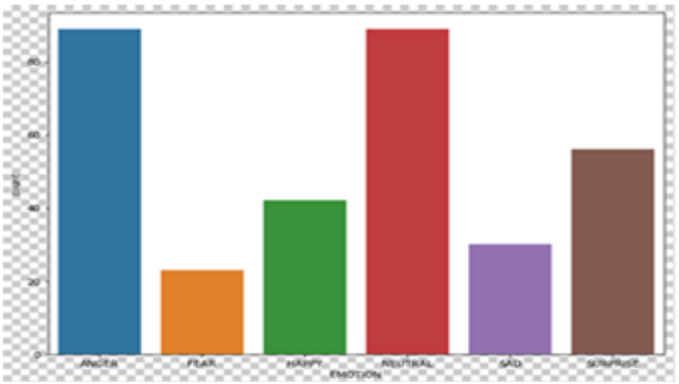

Figure1: Emotional label distribution

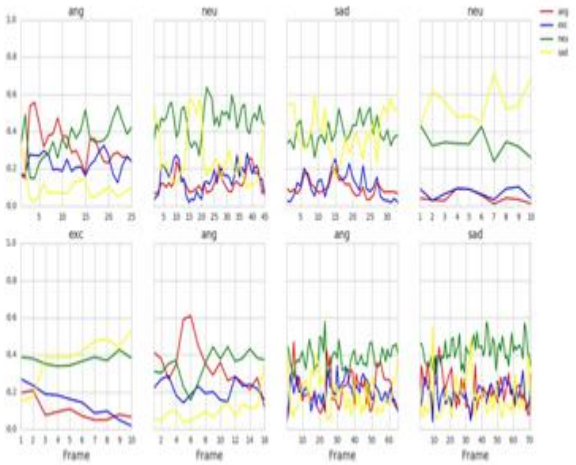

Figure 2: Frame wise Classification

IV. EMOTION RECOGNITION EXPERIMENTS

In series of experiments we had to investigate different models and approaches to the emotion recognition task. All the code can be found in GitHub repository [9].

As it was mentioned in section 2, data structure is the following — dialogues are broken into utterances by human assessors. Utterances are the least structural unit with emotion tag in original labelling. But utterances considerably vary in length. Thus we decided to split them into overlapped frames of 0.2 seconds duration with overlapping of 0.1 second. The problem is that frames have no labels. While it is obvious that not all frames of angry utterance also can be referred as an angry frame.

Further in this paper, unless otherwise specified, the test set consists of 20% points randomly picked from overall dataset.

4.1Frame wise classification

The first method implemented is the frame wise classification. The goal of this method is to try to classify each frame separately. Under certain specific conditions we came up with the following workflow:

•We take two of the loudest frames from each utterance. Loudness is the synonymous for spectral power.
•We assign these frames with the emotion label of the utterance.
•We then train the frame classification model on the derived dataset.

Here we make some assumptions that the emotions of each of the utterance is contained not in all frames but only in the loudest frames. Experiment shows that 2 frames is the optimal number. Regarding the classification model we used various classifiers such as Logistic Regression, Random Forest Classifier, Gradient Boosting classifier and Linear SVM from scikit-learn package [10].

In the figure 2. We can observe the results of this method for random utterances from the test data. It looks somewhat reasonable with the short utterances. On the longer ones it becomes saw-tooth and unstable.

The next step is the classification of utterances based on its artificial labelling. A simple majority voting algorithm gives an accuracy of about 44%. Taking into account that the neutral class is about 20% of a dataset it does not look very good.

Moreover, the error distribution in this case looks unnatural to that in comparison with the human one got by us. 70% of the answers were neutral, which implicitly confirm our assumptions that most of the frames in the utterances don't have any emotion

4.2Utterance-level classification

The goal of the following experiment is to use frame features itself as an input to the classifier model. The key idea in utterance level classification is that RNN can learn sufficient features from input features stream itself and made the final classification based on them [2].

Model Flow Chart

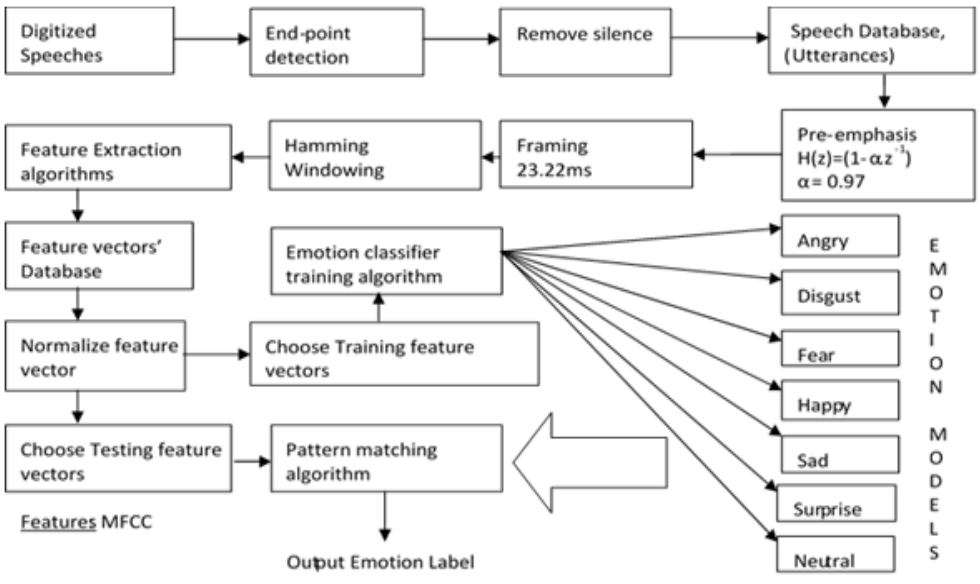

Figure 3.Pre-processing and emotion classifier training and testing flow diagram [4]

Validation Classification Report			
Class	Precision	F1-score	Recall
Angry	0.58	0.42	0.33
Fear	0.33	0.33	0.33
Happy	0.14	0.12	0.10
Neutral	0.62	0.57	0.53
Sad	0.31	0.43	0.71
Surprise	0.17	0.22	0.33
Avg/Total	0.45	0.40	0.39

Fig.4: Mean Classification Score (for Logistic Regression)

	Pred: angry	*Pred: fear*	*Pred: happy*	*Pred: neutral*	*Pred: sad*	*Pred: surprise*
Actual: angry	15	0	0	5	0	1
Actual: fear	1	1	0	1	0	0
Actual: happy	2	0	2	3	0	3
Actual: neutral	0	0	0	19	0	0
Actual: sad	1	0	0	1	1	4
Actual: surprise	1	0	0	2	1	2

Fig. 5: Confusion Matrix by Logistic Regression

Validation Classification Report			
Class	Precision	F1-score	Recall
Angry	0.75	0.73	0.71
Fear	1.00	0.50	0.33
Happy	1.00	0.33	0.20
Neutral	0.61	0.76	1.00
Sad	0.50	0.22	0.14
Surprise	0.20	0.25	0.33
Avg/Total	0.68	0.57	0.61

Fig.6: Mean Classification Score (Random Forest classifier)

	Pred: angry	*Pred: fear*	*Pred: happy*	*Pred: neutral*	*Pred: sad*	*Pred: surprise*
Actual: angry	15	0	0	5	0	1
Actual: fear	1	1	0	1	0	0
Actual: happy	2	0	2	3	0	3
Actual: neutral	0	0	0	19	0	0
Actual: sad	1	0	0	1	1	4
Actual: surprise	1	0	0	2	1	2

Fig. 7: Confusion Matrix by Random Forest classifier

Validation Classification Report			
Class	Precision	F1-score	Recall
Angry	0.70	0.68	0.67
Fear	0.67	0.67	0.67
Happy	0.00	0.00	0.00
Neutral	0.60	0.73	0.95
Sad	0.50	0.22	0.14
Surprise	0.18	0.24	0.33
Avg/Total	0.50	0.50	0.56

Fig.6: Mean Classification Score (Gradient boosting classifier)

	Pred: angry	*Pred: fear*	*Pred: happy*	*Pred: neutral*	*Pred: sad*	*Pred: surprise*
Actual: angry	14	1	0	5	0	1
Actual: fear	0	2	0	1	0	0
Actual: happy	1	0	0	3	1	5
Actual: neutral	1	0	0	10	0	0
Actual: sad	2	0	0	1	1	3
Actual: surprise	2	0	0	2	0	2

Fig. 7: Confusion Matrix Gradient boosting classifier

Validation Classification Report			
Class	Precision	F1-score	Recall
Angry	1.00	0.09	0.50
Fear	0.00	0.00	0.00
Happy	0.00	0.00	0.00
Neutral	0.29	0.45	1.00
Sad	0.00	0.00	0.00
Surprise	0.00	0.00	0.00
Avg/Total	0.40	0.16	0.30

Fig.6: Mean Classification Score (Linear SVM)

	Pred: angry	*Pred: fear*	*Pred: happy*	*Pred: neutral*	*Pred: sad*	*Pred: surprise*
Actual: angry	1	0	0	20	0	0
Actual: fear	0	0	0	3	0	0
Actual: happy	0	0	0	10	0	0
Actual: neutral	0	0	0	19	0	0
Actual: sad	0	0	0	7	0	0
Actual: surprise	0	0	0	6	0	0

Fig. 7: Confusion Matrix by Linear SVM

V. RESULTS AND DISCUSSION:

For After exhaustive trials of experiment with various binary classifiers such as Logistic Regression, Random Forest, Gradient Boosting and Linear SVM, the highest mean classification score rose to 0.45 using Logistic Regression, 0.68 using Random Forest classifier, 0.50 with Gradient Boosting classifier and 0.40 using Linear SVM with 23 MFCC vectors. We did observe that from the experiment that the Random Forest classifier gives the highest mean classification score. In this model, all emotional models had the same number of states. There is no much change of effect in pre-emphasis filter before framing is done. Hence, pre-emphasis filtering is done before framing to reduce computational burden. The flow diagram, used in our experiments, is given in figure 5. The results of experiments by different binary classifier are listed below and their confusion matrix is plotted respectively. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. It shows us the amount of samples that were misclassified with respect to their actual values and the predicted values. A detailed presentation of the outputs obtained by using different binary classifiers has been given in the figures (6 to 13).

VI. CONCLUSION

From observations of the results presented in the figures we find that the surprise emotion is the most difficult one to disambiguate from other emotions, since surprise may be expressed along with any other emotion such as angry-surprise, fear-surprise, happy-surprise, etc.

Also we showed that the results are comparable with the state-of-the art ones in this field. Moreover, we analyzed model answers and error distribution along with human performance and came to the conclusion that emotion is a very subjective notion and even if humans outperform computer the difference is not so significant.

In this study, the overview of different SER methods are discussed for extracting audio features from speech sample, various classifier algorithms are explained briefly. Speech Emotion Recognition has a promising future and its accuracy depends upon

The combination of features extracted, type of classification algorithm used and the correct of emotional speech database. This study aims to provide a simple guide to the researcher for those carried out their research study in the speech emotion recognition systems.

REFERENCES

[1]. Vladimir Chernykh, Pavel Prihodko and Grigoriy Sterling, MIPT IITP, Skoltech. Emotion Recognition from Speech with Recurrent Neural Networks,2017.

[2]. Jinkyu Lee and Ivan Tashev. Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea. High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition, In Interspeech 2015, 2015.

[3]. Miss. Aparna P. Wanare, Prof. Shankar N. Dandare, Department of Electronics & Telecommunication Engineering, Sant Gadge Baba Amravati University, Amravati. Human Emotion Recognition from Speech, 2014.

[4]. Aditya Bihar Kandali, Student Member, IEEE, Aurobinda Routray, Member, IEEE, and Tapan Kumar Basu. Emotion recognition from Assamese speeches using MFCC features and GMM classifier, 2009.

[5]. C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation, 42(4), 2008.

[6]. The Association for the Advancement of Affective Computing. http://emotion-research. net/wiki/Databases, 2014.

[7]. A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning, 2006.

[8]. L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 1989.

[9]. V.Chernykh, G. Sterling, and P. Prihodko. https://github.com/vladimir-chernykh/ emotion_recognition, 2016.

[10]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12, 2011.

[11]. K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In ICCV 2015, 2015.

[12]. R. Altrov and H. Pajupuu. The influence of language and culture on the understanding of vocal emotions. Journal of Estonian and Finno-Ugric Linguistics, 6(3), 2015.

[13]. A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. ArXiv e-prints, 2016.

[14]. AL Sweigart: Automate the Boring Stuff with Python Programming. https://automatetheboringstuff.com, 2015.

[15]. Mark Gales and Steve Young. The Application of Hidden Markov Models in Speech Recognition, Cambridge University Engineering Department, Trumping ton Street, Cambridge, CB2 1PZ, UK.

[16]. Suriyadeepan Ram. Linear Regression: The Probabilistic Perspective. http://suriyadeepan.github.io/2017-01-22-mle-linear-regression/

G.R.Michael "Comparative Study of Various Classifiers Using Assamese Speech Utterances " International Journal of Computational Engineering Research (IJCER), vol. 08, no. 06, 2018, pp. 33-38.