Anjani Reddy J, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 11, Issue 6, (Series-IV) June 2021, pp. 01-05

RESEARCH ARTICLE OPEN ACCESS

Emotion Recognition from Speech Using MLP and KNN

Anjani Reddy J, Dr. Shiva Murthy G
PG Scholar Artificial Intelligence and Machine Learning
Visvesvaraya Technological University For Post Graduate Studies Bangalore, India
Professor Artificial Intelligence and Machine Learning Visvesvaraya Technological University For Post Graduate Studies Bangalore, India

ABSTRACT:
Quantitative human nature research relies heavily on emotion recognition. When the number of applications based on emotion recognition grows, so does the need for a more accurate emotion recognition system. In speech, there are two types of content. Think as long as the utterances are rendered according to the language's rules of pronunciation, the semantic part of the speech carries linguistic knowledge. Paralinguistic knowledge, on the other hand, refers to hidden messages like the speaker's emotional state. The recognition of paralinguistic features that describe the speaker's emotional state is an essential first step in speech emotion recognition. In human communication, there are over 25 different forms of emotions. In this project, the seven basic emotions in the speech signal: angry, calm, disgust, fear, happy, sad, surprise, and neutral for no emotions are considered for the recognition and classification. Experiments were carried out for automatic classification. The procedure consisted of applying KNN classifier and MLP classifier for speech emotion recognition. The results of application of each algorithm for various parameters are tabulated and the best settings of the variables in these algorithms to get the maximum performance are inferred. For emotion recognition, the KNN algorithm proved to be more efficient than the MLP Classifier.
Keywords: Speech Emotion Recognition, KNN Classifier, MLP Classifier

---

Date of Submission: 15-06-2021 Date of Acceptance: 30-06-2021

I. INTRODUCTION

Humans communicate using speech signals because it is the fastest and most popular method. There are a variety of reasons for determining a speaker's emotional state. This reality has prompted researchers to evaluate speech as a fast and efficient form of human-machine interaction. If the state of emotion is identified during human–computer interaction, the machine can be programmed to generate more suitable responses. To increase the accuracy of recognition of spoken words, most modern automatic speech recognition systems use natural language comprehension. Such language processing will be enhanced even further if the speaker's emotional state can be identified, which would increase the system's accuracy. If the speaker's emotional state could be recognised and presented, it would provide additional useful information to the negotiating parties, particularly in non-face-to-face contexts. An automated emotion recognizer's output may naturally consist of emotion labels. It is critical to choose a suitable set of labels. Linguists have a large vocabulary of emotional state terms. Williams and Stevens (1981) stated that the sympathetic nervous system is aroused by the emotions of Anger, Fear, or Joy in the articulation of speech. As a result, the heart rate and blood pressure rise, the mouth becomes dry, and muscle tremors occur on occasion. Speech is loud, fast, and enunciated with a lot of high frequency energy. But in the other hand, when the parasympathetic nervous system is stimulated, as with Sadness, heart rate and blood pressure fall and salivation rises, resulting in slow, low-frequency speech. The effects of such physiological changes on speech can thus be seen in the overall energy, energy distribution across the frequency spectrum, and the frequency and duration of pauses in the speech signal. For the reasons listed below, the task of speech emotion recognition is extremely difficult. [7] First, it is unclear which aspects of speech are most effective at differentiating between emotions. The acoustic heterogeneity introduced by the presence of various sentences, speakers, speaking styles, and speaking speeds introduces another stumbling block since these properties have a direct impact on the majority of commonly extracted speech features like pitch and energy contours. Furthermore, a single utterance can include several perceived emotions, each of which corresponds to a separate component of the spoken utterance. Furthermore, determining the differences between these sections is extremely difficult. Another

difficult problem is that the way a person expresses an emotion is largely determined by the speaker, his or her background, and the atmosphere in which he or she lives. The majority of research has concentrated on monolingual emotion classification, assuming that there are no cultural differences between speakers. [3] Another issue is that an emotional condition, such as depression, will last for days, weeks, or even months. Other emotions will be temporary in this case, lasting no over than a few minutes. As a result, it's unclear if the automated emotion recognizer can pick up on the long-term or fleeting emotion. There is no universally accepted scientific concept of emotion. [8] There's a need to specify a collection of significant emotions to be identified by an automated emotion recognizer is a major problem in speech emotion recognition. Scholars have developed warehouses of the most common emotional states we experience. A typical set is given by Schubiger and O'Connor and Arnold, which contains 300 emotional states. . However, classifying such a large number of emotions is very difficult. Schubiger and O'Connor and Arnold have an unique link, which includes 300 emotional states. It is, nevertheless, quite hard to characterize such a vast range of emotions. Many experts agree with the 'palette principle,' which says that any emotion can be broken down into primary emotions in the same manner that any colour is made up of a few simple colours. Anger, disgust, fear, joy, sadness, and surprise are the primary emotions. These are the most visible and distinct feelings we experience in our lives. The archetypal feelings are what they're called. [1] [2]

II. LITERATURE SURVEY

ASSESS (McGilloway et al., 2000) is a framework for identifying four archetypal feelings, Fear, Rage, Sadness, and Joy, using a few landmarks—peaks and troughs in the profiles of fundamental frequency, duration, and limits of pauses and fricative bursts. A classification rate of 55% was achieved using discriminant analysis to separate samples that correspond to various groups.[4]

For classification, Dellaert et al. (1996) worked on the F0 information. Happy, sad, anger, and fear were the four emotions studied. The maximum, minimum, and median of the fundamental frequency, as well as the mean positive derivative of the regions where the F0 curve is increasing, are said to be the most important features that describe the acoustic correlates of emotion. [5]

III. IMPLEMENTATION

A. DATABASE

In this contribution, RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) database was used. The RAVDESS is a multimodal collection of emotional speech and song that has been verified. This database consisted of 7356 files of 24 actors (12 female actors and 12 male actors), vocalizing two lexically-matched statements in a neutral North American accent. [6] This database consists of calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each statement or expression was provided in 2 intensity levels (strong, normal) with a neutral expression. Each file has a unique name which is divided into 7 parts. Each part of the name gives characteristics of that particular file as given below.

Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
Vocal channel (01 = speech, 02 = song).
Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
Repetition (01 = 1st repetition, 02 = 2nd repetition).
Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).
example: Filename - 03-02-08-01-02-02-14.mp4 has audio-only(03), Song(02), Surprised(03), normal(01), "Dogs are sitting by the door"(02), second repetition(02) and actor is female, as actor number is even.

B. SOFTWARE (TOOLS AND TECHNIQUES)

For this project, Google Colaboratory, or "Colab" for short was the platform used to write the python code. Colab is a web-based Python editor that allows developers to write and run arbitrary Python code. It's particularly useful for machine learning, data processing, and education. Colab is a hosted Jupyter notebook programme that doesn't need any configuration and offers open access to computing services, including GPUs.

C. ALGORITHMS

MLP Classifier

A neural network called multilayer perceptron (MLP) is a kind of feedforward artificial neural network (ANN). The term MLP is fuzzy, referring to networks consisting of several layers of

perceptrons (with threshold activation) in some cases and in other cases it is seen as feedforward ANN. An MLP has minimum three layers of nodes: an input layer, a hidden layer, and an output layer. Each node, with the exception of input nodes, is a neuron with a nonlinear activation function. Backpropagation is a supervised learning method used by MLP for teaching. MLP is distinguished from a linear perceptron by its many layers and non-linear activation. It will tell the difference between data that isn't linearly separable and data that is. Multilayer perceptrons, particularly those with a single hidden layer, are often pointed to as "vanilla" neural networks. In the 1980s, MLPs were a common machine learning solution, with uses in speech recognition, image recognition, and machine translation tools.

KNN Classifier

The K-Nearest Neighbour algorithm is built on the Supervised Learning methodology and is one of the most basic Machine Learning algorithms. When K-NN algorithm believes that the new case/data and existing cases are identical, it places the new case in the group that is more similar to existing categories. The K-NN algorithm collects all relevant information and classifies a new data point based on its similarities to the existing data. This means that new data can be quickly sorted into a well-defined group using the K-NN algorithm. The K-NN algorithm could be used for both regression and classification, but it is most often used for classification tasks. It's also known as a lazy learner algorithm because it doesn't learn from the training set right away; instead, it saves the dataset and performs an operation on it when it comes time to classify it.

IV. RESULTS AND DISCUSSION

The comparison of accuracy for both the classifiers is given in the figure 1. MLP classifier has highest accuracy of 51.11% and lowest accuracy of 39.7%, whereas KNN classifier has highest accuracy of 63.61% and lowest accuracy of 35.56%. The accuracy of KNN algorithm depends on various parameters of the algorithm such as number of neighbors, weights, power parameter, etc. Various experiments were performed by varying the values of these parameters which resulted in different accuracies.

The first parameter that was studied, was to find the optimum number of neighbors to get the best accuracy. This experiment showed that as the number of neighbors was increased, the accuracy decreased. This is shown in figure 2. The accuracy was high when the number of neighbors was set to 2.

The second parameter that was studied, was to find which weight function to be used. There are two types of weight functions: “uniform” weight function and “distance” weight function. In uniform weight function, all the points in each neighborhood will be weighted equally. In distance weight function, the points are weighted inversely to their distance, that is the points closer to the query point will be having higher weight values compared to the points which are further. An experiment was conducted to find which weight function performs better for the problem stated. This experiment results are shown in figure 3, where the accuracy is better for distance weight function compared to the uniform weight function.

Classifier	MLP	KNN
Accuracy	51.11%	63.61%
Precision	0.5866	0.6501
F-score	0.5189	0.6372
Matthews correlation coefficient	0.4672	0.5839
Recall	0.5111	0.6361

Figure 1: Comparison of algorithms

n-neighbors	2	3	5	21
Accuracy	47.50%	46.67%	45.00%	35.56%
f1_score	0.45992714	0.45704299	0.44071487	0.32664600
Recall	0.475	0.46666666	0.45	0.35555555
Precision	0.47404959	0.47120143	0.44622367	0.33457527
Matthews correlation coefficient (MCC)	0.45685000	0.39027321	0.37054544	0.26478683

Figure 2: Finding optimum number of number of neighbors

Weights	Uniform	distance
Accuracy	47.50%	54.17%
f1_score	0.45992714	0.54209030
Recall	0.475	0.54166666
Precision	0.47404959	0.55216548
Matthews correlation coefficient (MCC)	0.40060861	0.47579639

Figure 3: Finding optimum weight function

p	1	2
Accuracy	63.61%	54.17%
f1_score	0.63722742	0.54209030
Recall	0.63611111	0.54166666
Precision	0.65017200	0.55216548
Matthews correlation coefficient (MCC)	0.58391783	0.47579639

Figure 4: Minkowshi metric

The third parameter that was studied is power parameter for the Minkowski metric. Two values can be assigned to this parameter (p1 and p2). When this parameter is set to 1, Manhattan distance metric is used and when it is set to 2, Euclidean distance metric is used. The results of this experiment are shown in the figure 4. The results showed that Manhattan distance metric gave better accuracy compared to Euclidean distance metric.

V. CONCLUSION

As discussed, the quantitative human nature research relies heavily on emotion recognition. In this paper, two algorithms were presented for speech emotion recognition. MLP algorithm gave a highest accuracy of 51.11% as discussed above and KNN gave an accuracy of 63.61%. The best settings of the variables in the algorithms to get the maximum performance was inferred.

REFERENCES

[1]. T. Athanaselis, S. Bakamidis, R. Cowiea, E. Douglas Cowie, I. Dologlou and C. Cox, "ASR for emotional speech: Clarifying the issues and enhancing performance," Neural Netw., vol. 18, pp. 437–444, 2005.

[2]. F. Burkhardt, A. Paeschke, M. Rolfes, and W. S. et al., "A database of German Emotional Speech," in Proc. Interspeech, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[3]. S. W. Foo, T. L. Nwe, and L. C. D. Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, 2003.
[4]. Sinad McGilloway, Roddy Cowie, Ellen Douglas-Cowie, Stan Gielen, Machiel Westerdijk, and Sybert Stroeve, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark", ISCA, Sept 5-7, 2000.
[5]. Frank Dellaert, Thomas Polzin and Alex Waibel, "RECOGNIZING EMOTION IN SPEECH", ISCA, Oct 3-6, 1996.
[6]. Steven R. Livingstone, 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): e0196391, May 2018.
[7]. Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014), "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks", IEEE transactions on multimedia, vol. 16, no. 8, Dec 2014.

Anjani Reddy J, et. al. "Emotion Recognition from Speech Using MLP and KNN." *International Journal of Engineering Research and Applications (IJERA)*, vol.11 (6), 2021, pp 01-05.