SRM

by Srm Srm

Submission date: 14-Apr-2021 09:20PM (UTC+0530)
Submission ID: 1559110787
File name: Publication_4.pdf (314.61K)
Word count: 2294
Character count: 12243

EMOTION WITH GENDER CLASSIFICATION THROUGH SPEECH USING MACHINE LEARNING

Shivam Singh
Final Year B.Tech.(CSE)
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Kattankulathur, India
Email: ss9532@srmist.edu.in

Prashant Kumar
Final Year B.Tech.(CSE)
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Kattankulathur, India
Email: pr1485@srmist.edu.in

Mrs. G. Malarselvi
Assistant Professor
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Kattankulathur, India
Email: malarseg@srmist.edu.in

Abstract*— Past few years, there are several experiments done on human speech, in which emotion and gender classification of a human through their voice is a challenging research area in voice technology. In this report we discussed the prediction/classification of gender and emotion group through human voice and tried different architecture. Features like log-mel, chroma values, Mel Frequency Cepstral Coefficients (MFCC), pitch, formants etc. are extracted from every voice samples and then selected through PCA (Principal Component Analysis) and RFE (Redundant Feature Elimination) techniques. The given network obtains the emotion information through given data and classifies the gender category. The proposed algorithm is worked on two public datasets with two architectural systems. The dataset used for training and testing are from the TESS and RAVDESS. The proposed model gives us high accuracy. We are going to explain the approaches used in our solutions and their drawbacks.

***Keywords*—** Emotion Prediction, Gender Prediction, PCA, MLP, Deep Learning, REF, RAVDESS, MFCCs, TESS

I. INTRODUCTION

The Emotion and Gender group prediction finds a variety of uses in modern human voice communication systems. A few uses of this model include focusing on feedback and advertising by voice assistants and personalized editing using speech details to help search the criminal examination. Various algorithms have been used to predict the Emotion and Gender of users using image processing, but little progress has been made in predicting Emotion and Gender using only voice analysis. Emotional and gender speculation using speech elements is a major challenge due to the lack of training labels for training. Also, the variation of the pitch internal speaker, formants, anger, emotions, context, etc. enhances the quality in predicting Emotion and Gender groups accurately. Our paper provides information on key speech features from a person's voice, as well as in-depth learning how to use those features to predict the user's Emotion and Gender. The whole report is shown as follows; Phase II covers previous/related work, Phase III describes the prediction methods of the Emotion and Gender team accurately; Section IV contains the concluding part of the report.

II. RELATED WORK

[1] The R-CNN and the gender information block are included in proposed algorithm. The R-CNN block removes the required emotional data from the speech details and separates the emotions. The three data used for the different language systems are used to test the proposed algorithm. The test accuracy achieved by proposed model is 5.6% in Mandarin, 7.3% in English and 1.5% in German compared to the most accurate algorithms available. To prove the algorithm's release, we tend to use the FAU and information data in these different repositories, the given algorithm can achieve 85.8% accuracy and 71.1%, respectively.

[2] As a function extractor, the CNN model is integrated with the RF classifier. The Chinese speech recognition device is based on this, and is used in the NAO robot. The first audio box is upgraded during the application process, and the new recording bo was created. Speech recognition acquired by the Advanced Recorder box not only meets the requirements of the speech recognition format, but in conjunction with the requirements of the speech recognition format.

[3] The proposed convolutional recurrent algorithm operates on audio to perform spontaneous emotion classification task from speech datasets. We also suggest direct optimization of the concordance correlation coefficient, which is used to assess the rate of agreement between forecasts and the gold-standard. On the RECOLA database, the given approach performs substantially higher than conventional built features, showing the effectiveness of learning features that are better suited to the task. Finally, we investigate the activations of the recurrent layers and discover cells that strongly associate with prosodic features that were previously thought the cause of delay.

[4] Studies of this idea have shown that this approach plays an important role in the medical and technical fields. This certification principle is additionally a rigorous supply of authentication processes. The idea is to reclaim its origins by aggregating data that can be traced back to a vast amount of data from computer science and other related disciplines and keep its parent company as inevitable in the world of the future. The reason for the development of these methods is not only to make people progress and to reduce time.

[5] This paper proposes a language independent emotional classification systems for identifying the emotional state of human in speech code. Emotional speech with different disciplines, completely different languages speaking, is used to develop and test the possibilities of system. First identify the potential speech features which

are collected from the datasets. Then a systematic feature selection process was introduced, in which sequential forward selection (SFS) was used in combination with a consistency-based selection approach based on a common regression neural network (GRNN). To understand the identification of emotions, selected features are fed into a modular neural network (MNN). The proposed system provides very satisfactory sense detection accuracy, although its propensity for language freedom reflects a huge increase in proficiency.

II. PROPOSED WORK

The overall principles used for proposed emotion with gender classification framework is discussed in this section. The solution to this problem can be addressed in three ways.

[A] Datasets

A total of 5252 samples were used to construct the dataset which are:

1.Ryerson Audio-Visual Database of Emotional Speech and Song dataset.

2.Toronto Emotional Speech Set dataset.

RAVDESS and TESS dataset, both consist of around five thousand two hundred speeches labels with emotion and gender (male and female) with various geographical areas, and different pronunciation and accents. This dataset is generally used for ASR (Automatic Speech Recognition) systems.

[B] Principles

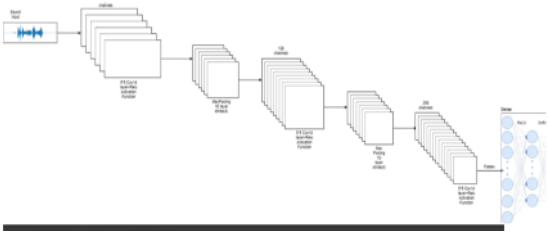
The emotion with gender classification model is based on deep learning architecture, those are CNN, SVM Classifier, MLP Classifier. The key thought is study about the MFCC regularly referenced in light of the fact that the "spectrum of a spectrum", on the grounds that the solitary element to train and test the model. Mel-frequency cepstrum may be a different point of view of work shown to the newest technology of sound standardize for ASR tasks, and it could be a different understanding of the MFC. The MFC coefficient have been commonly used to describe the amplitude of spectrum of an acoustic wave.

To obtain statistically stationary waves, the audio of datasets is divided into different fixtures. The amplitude range is normalized by decreasing the frequency scale of the “mel”. This procedure is carried out in order to empathize with the frequency in order to reconstruct the wave as accurately as the human auditory system would interpret it. The feature extracted from each audio file is 18. After feature extraction the audio file is converted into floating-point time series. The time series was then translated into an MFCC sequence.

[C] Models

CNN

Figure 1 shows the operational results of the convolutional neural networks developed for the prediction. For every audio dataset supplied as an input, the network will operate with 40 feature vectors. The 40 values reflect the two-second audio frame's compact numerical form. We gave input X on which execute 1 cycle of 1Dimension CNN with an activation function (rectified linear unit) ReLu, with 30% dropouts. The ReLu (rectified linear unit) helps us to get greater value in the case of activation, by using this function to represent hidden units. In this case, pooling will assist the model in focusing only on the most important characteristics of each piece of data, making them position invariant. We repeated the procedure, this time adjusting kernel size. After that, we used different dropout and flattened the result to make model compliant using following layers. Finally, we used a Dense layer with a soft max activation function to estimate the probability distribution of each of the encoded classes.



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 40, 64)            384
_________________________________________________________________
activation_1 (Activation)    (None, 40, 64)            0
_________________________________________________________________
dropout_1 (Dropout)          (None, 40, 64)            0
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 10, 64)            0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 10, 128)           41088
_________________________________________________________________
activation_2 (Activation)    (None, 10, 128)           0
_________________________________________________________________
dropout_2 (Dropout)          (None, 10, 128)           0
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 2, 128)            0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 2, 256)            164096
_________________________________________________________________
activation_3 (Activation)    (None, 2, 256)            0
_________________________________________________________________
dropout_3 (Dropout)          (None, 2, 256)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 4104
_________________________________________________________________
activation_4 (Activation)    (None, 8)                 0
=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
```

CNN Layer Description

SVM

The Support Vector Machine (SVM) is a supervised machine learning algorithm for solving regression and classification problems. It is, however, primarily employed in classification problems.

We draw every audio data as an extent in n-dimensional with specific coordinate in the SVM algorithm.

Before applying the data are usually scaled to associate with SVM classifier to avoid the bigger numeric ranges whereas process it. Scaling conjointly serves the aim of preventing some numerical stuff throughout the calculation.

MLP

A **MLP** is a type of artificial neural network that uses feedforward learning (ANN). For preparation, MLP employs backpropagation, a supervised learning technique. The error in an output node in a data point can be represented by, the chosen standards and is the standards provided by the perceptron. The modification of node weight is depended on the correction of overall performance which reduce, as defined by

$$\mathcal{E}(n) = \frac{1}{2}\sum_{j} e_j^2(n)$$

The change in weight is carried out by using gradient descent,

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

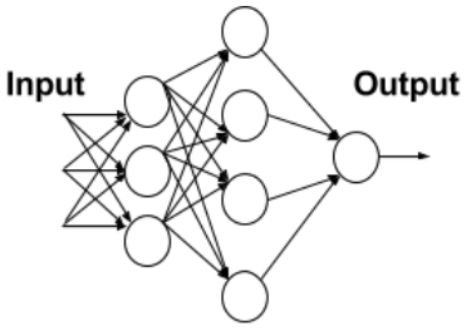
The induced local field, which varies, affects the derivative to be measured. In simple way we can show that the derivative can be reduced for output node as given

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = e_j(n)\phi'(v_j(n))$$

The observation of a shift in weights of a hidden node is more complicated, but it can be showed the correct derivation is

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = \phi'(v_j(n))\sum_{k} -\frac{\partial \mathcal{E}(n)}{\partial v_k(n)} w_{kj}(n)$$

The output layer showed by the change in weights of the Kth nodes, determines this. As a result, to adjust the weights of hidden layer, weights of the output layer is changed in accordance with the activation function's derivative, and thus the algorithm is a backpropagation of the activation function.



The MLP model has been tested in terms of predicting a speaker's gender and emotion based on speech features. MLP outperforms other classification algorithms such as CNN, SVM, R-CNN, and LSTM in terms of feature selection and prediction accuracy with the proposed single model approaches. With eight emotion cats, very good performance metric scores were obtained.

RESULT

SVM classifier achieved 0.82.

CNN Classifier achieved 0.85.

MLP Classifier achieved 0.99 over the 8 classes.

Features extracted: 180

Accuracy: 99.44%

	precision	recall	f1 score	support
angry_female	0.99	1.00	0.99	89
angry_male	1.00	0.99	1.00	113
disgust_female	0.99	1.00	0.99	96
disgust_male	1.00	1.00	1.00	104
fearful_female	0.99	1.00	1.00	109
fearful_male	1.00	0.99	1.00	110
happy_female	1.00	1.00	1.00	97
happy_male	0.98	1.00	0.99	104
neutral_female	1.00	1.00	1.00	91
neutral_male	1.00	1.00	1.00	86
sad_female	1.00	0.99	0.99	96
sad_male	1.00	1.00	1.00	108
surprised_female	1.00	1.00	1.00	95
surprised_male	1.00	0.98	0.99	102

	precision	recall	f1 score	support
accuracy			0.99	1400
macro avg	1.00	0.98	0.98	1400
weighted avg	1.00	0.99	0.99	1400

II. CONCLUSION

After all these research and experiments in this survey report, we conclude that we trying with three different models to identify the emotion with gender, which gives different accuracy at last we continue with the model whose accuracy is high using TESS dataset and we assure that after combining the emotion with gender no one get that much accuracy.

The limitation of this experiment is we still don't test our model on multiple speaker, because of that we use single audio voice of each male, female with their different emotions.

REFERENCES

[1] TING-WEI SUN, (Graduate Student Member, IEEE),” End-to-End Speech Emotion Recognition with Gender Information”, Received July 4, 2020, accepted August 12,

2020, date of publication August 18, 2020, date of current version August 28, 2020.

[2] Li Zheng, Qiao Li, Hua Ban, Shuhua Liu," Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest", Received July 9, 2018, date of conference June 11, 2018, IEEE, Shenyang, China.

[3] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi Mihalis, A. Nicolaou, Björn Schuller, Stefanos Zafeiriou," END-TO-END SPEECH EMOTION RECOGNITION USING A DEEP CONVOLUTIONAL RECURRENT NETWORK", Received May 19, 2016, date of conference March 20-25, 2016, IEEE, Shenyang, China.

[4] K. Tarunika, R. B Pradeeba, P.Aruna," Applying Machine Learning Techniques for Speech Emotion Recognition", Received 18 October 2018, date of conference 10-12 July 2018, IEEE, Bengaluru, India.

[5] Muhammad Waqas Bhatti, Yongjin Wang and Ling Guan," A NEURAL NETWORK APPROACH FOR HUMAN EMOTION RECOGNITION IN SPEECH", Received 03 September 2004, date of conference 23-26 May 2004, IEEE, Vancouver, BC, Canada.

[6] R.B Pradeeba, K.Tarunika, Dr.P.Aruna, "Accuracy of speech emotion recognition through deep neural network and k-nearest", International Journal of Engineering Research in Computer Science and Engineering, Vol 5, Issue 2, February 2018.

[7] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, ''Semi supervised autoencoders for speech emotion recognition,'' IEEE/ACM Trans. Audio, Speech, Lang., Process., vol. 26, no. 1, pp. 31–43, Jan. 2018.

[8] Lin Yilin, Wei Gang. Speech Emotion Recognition Based on HMM and SVM // Proc of the 4th International Conference on Machine Learning and Cybernetics. Guangzhou, China, 2005, VIII:4898-4901.

[9] K. Wang, N. An, B. Nan Li, Y. Zhang, and L. Li, ''Speech emotion recognition using Fourier parameters,'' IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 69–75, Jan. 2015.

[10] W. Lim, D. Jang, and T. Lee, ''Speech emotion recognition using convolutional and recurrent neural networks,'' in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA), Jeju, South Korea, Dec. 2016, pp. 1–4.

SRM

ORIGINALITY REPORT

8%	4%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijrte.org Internet Source	3%
2	Ting-Wei Sun. "End-to-End Speech Emotion Recognition With Gender Information", IEEE Access, 2020 Publication	1%
3	Submitted to Middlesex University Student Paper	1%
4	www.mdpi.com Internet Source	1%
5	M.W. Bhatti, Yongjin Wang, Ling Guan. "A neural network approach for human emotion recognition in speech", 2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512), 2004 Publication	1%
6	Li Zheng, Qiao Li, Hua Ban, Shuhua Liu. "Speech emotion recognition based on convolution neural network combined with	1%

random forest", 2018 Chinese Control And Decision Conference (CCDC), 2018
Publication

7 Sung-Woo Byun, Seok-Pil Lee. "Human emotion recognition based on the weighted integration method using image sequences and acoustic features", Multimedia Tools and Applications, 2020 — 1%
Publication

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		