Speech Emotion Recognition

Deepesh Bhageria
CSE
IIIT Sri City
Andhra Pradesh, India
Email: deepesh.b17@iiits.in

Debapriya Tula
CSE
IIIT Sri City
Andhra Pradesh, India
Email: debapriya.t17@iiits.in

Suraj Agarwal
CSE
IIIT Sri City
Andhra Pradesh, India
Email: suraj.a17@iiits.in

Sravan Kumar Vinakota
CSE
IIIT Sri City
Andhra Pradesh, India
Email: sravankumar.v17@iiits.in

***Abstract*—Speech is a complex signal consisting of various information, such as information about the speaker, message being communicated, its language, region, emotions etc. Speech emotion recognition is important to have a natural interaction between human beings and machines. In speech emotion recognition, the emotional state of a speaker is extracted from his or her speech. The acoustic characteristic of the speech signal is an important Feature. Feature extraction is the process that extracts a small amount of data from the speech signal that can later be used to represent each speaker. Many feature extraction methods are available and Mel Frequency Cepstral Coefficient (MFCC) is the commonly used method. In this paper, speaker emotions are recognized using the data extracted from the speaker voice signal. Mel Frequency Cepstral Coefficient (MFCC) technique is used to recognize emotion of a speaker from their voice. This paper proposes a vocal-based emotion recognition method using random forests, where Mel Frequency Cepstral Coefficient (MFCC) technique is used in order to recognize the emotional state of a speaker. The proposed technique adopts random forests to represent the speech signals, along with the decision-trees approach, in order to classify them into different categories. The emotions are broadly categorised into the seven groups, which are calm, happy, sad, angry,fearful, surprise, and disgust. The Toronto Emotional Speech Set (TESS) and audio files from Ryerson Audio-Visual Database of Emotional Speech and Song(RAVEDESS) database are used. The proposed method has an average recognition rate of (73.55%)(on RAVEDESS) and (98.61%)(on TESS).**

1. Introduction

Emotion recognition from a human speech is an attractive field of speech signal processing. It is drawing more attention in the applications where emotion recognition eases the speaker identification and mental status, such as in criminal investigation, intelligent assistance [1], detecting frustration, disappointment, surprise/amusement [2], health and medicine [3] and a better Human Computer Interface [4]. Extracting the emotional state of a speaker from his/her speech is called speech emotion recognition. Speech emotion recognition involves analysis of the speech signal to identify the appropriate emotion based on training its features such as pitch, formant and phoneme. For feature extraction and testing of a speech signal a good number of algorithms have been formulated. Few of them are Artificial neural networks (ANN),linear prediction cepstral coefficients (LPCC), Mel Frequency Cepstral coefficients (MFCC), combination of Linear Prediction coefficients and Mel Cepstrum coefficients (LPCMCC),the Support Vector Machine (SVM); combination of HMM [5] and SVM etc [6]. Our proposed work is based on feature extraction using MFCC and decision making using standard deviation. Organization of this paper is as follows: Section I describes the various data augmentations applied to the raw signal. Section II describes some literature we read before experimentation. Section III describes how the Random Forest algorithm is used for classification of the emotions. Section IV discusses the results obtained on the datasets used.

1.1. Motivation

Computers are now much more than just number-crunching machines. They are able to analyse even the slightest details about a human being, including a person's face, fingerprints, gait, and speech patterns. Recognising the speaker's emotion is but a drop in the ocean.

With the advent of human machine interaction, apart of 'what is said' and 'who said it', 'how it is said' plays a key role for effective human-machine communication and for the machine to react properly. There are people who lack essential brain functions responsible for processing or handling human emotion, due to which are unable to take critical decisions. This is our motivation to design a system that can analyse a person's speech and detect the emotion they are feeling. It can help people understand the other person better.

A system like this can also help in fields such as psychiatric diagnosis, customer service call centers and lie detection.

2. State of the art/Background

The current state of the art papers and work you can discuss in this section(at least 10 papers). In this section,

we summarised the research papers we have read before settling down on our current model.

2.1. Speech emotion recognition based on rough set and SVM [7]

Here, the concept of rough set was used where all available features that could be obtained for a given speech signal were first chosen as the conditional attributes and attribute reduction is applied to select the optimal subset of condition attribute set to express the target conceptions of the speech recogntion system. A tree based SVM was then applied for classification. This method was applied to the CDLC dataset.

2.2. Multimodal speech emotion recognition and ambiguity resolution [8]

A manual feature engineering approach is taken where attributes such as pitch, harmonics, speech energy, pause and central moments are fetched from the speech signal. An LSTM based model fed with these features as input is used as a classifier. This method was applied to the IEMOCAP dataset provided by USC.

2.3. Emotion recognition using a hierarchical binary decision tree approach [9]

This paper proposed a framework that uses multiple binary classifiers instead of a single multi-class classifier. The binary classifiers are arranged in a tree-like structure with the number of classifiers at the bottom level equal to the number of emotions. The binary classifiers used are Bayesian Logistic Regression classifiers. They conducted experiments on the AIBO and USC IEMOCAP datasets with a total of 384 acoustic features extracted.

2.4. Speech emotion recognition using hidden Markov models [10]

This paper proposed to use an ensemble of Hidden Markov models [5]. HMMs (Hidden Markov models) are a type of probabilitic models generally used for speech recognition. If there are N emotions, then N HMMs will be built on the dataset. When given an input audio sequence, the HMM that gives it the highest score will decide the emotion label.

2.5. Vocal-based emotion recognition using random forests [11]

This paper extracts features on the whole speech signal, namely, pitch, intensity, the first four formants, the first four formants bandwidths, mean autocorrelation, mean noise-to-harmonics ratio and standard deviation in order to recognize the emotional state of a speaker. The random forests along with the decision-trees approach is used, in order to classify them into different categories. The proposed method has an average recognition rate of 66.28% on Surrey Audio-Visual Expressed Emotion database with 13.78% higher average recognition rate as compared to the linear discriminant analysis as well as 6.58% higher average recognition rate than the deep neural networks results.

2.6. Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest [12]

Firstly,Convolution Neural Network is used as the feature extractor to extract the speech emotion feature from the normalized spectrogram, then for classifying the speech emotion features Random Forest classification algorithm is used with 200 Decision Trees. The recognition accuracy of CNN-RF is 84.68% which is 3.25% higher than that of CNN model.

2.7. Emotion Recognition using Bidirectional LSTM [13]

An LSTM based model that contains three layers namely the input layer, the representation layer, and the classifier layer. The input to the model is 40-dimensional MFCC features. The models worked very well for the intent classification task.

2.8. Emotion Recognition using Transformers [14]

The model uses transformer encoder blocks with the convolution layer. The input to the model is 83-dimensional filter bank features, which works very well even on the noisy dataset. The transformer-based model outperforms the LSTM based model.

3. Proposed System

Our system consists majorly of two parts: a Random Forest Classifier and a simple translator. The Random Forest Classifier predicts the emotion of the input speech signal and the translator translates that emotion into other languages. In the sections below, we will describe how we approached to making this system.

3.1. Data Augmentation

Data Augmentation is the process of slightly altering the training examples to generate new ones. For audio signals, there are four major augmentation techniques.

3.1.1. Noise Injection. It simply adds some random value or random background music into data.

3.1.2. Changing Pitch. Pitch is the fundamental period of the speech signal.

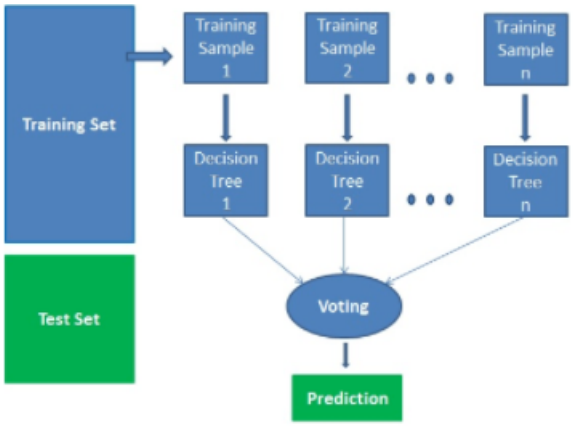

Figure 1. General framework of Random Forest Classifier

3.1.3. Adding Reverb. It is the reflection of sound waves created by the superposition of echoes.

3.1.4. Changing Speed. The audio signal is squished or stretched to give the effect of speeding up or slowing down.

3.2. Feature Extraction

There are many types of features that can be extracted from audio signals, but the most widely used features are the MFCC features.

MFCC stands for Mel-Frequency Cepstral Coefficients. Simply put, it is the coefficients of the Short-time Fourier Transforms over windows of the input audio signal.

3.3. Random Forest Classifier

Random Forest Classifier is a flexible, easy to use machine learning algorithm that produces, even without hyperparameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). It is a supervised learning algorithm. The ”forest” it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. A general framework of a Random Forest Classifier is shown in Figure 1.

4. Results

We conducted our experiments on two benchmark datasets: **RAVDESS** [15] and **TESS** [16].

4.1. RAVDESS

RAVDESS [15] stands for ”The Ryerson Audio-Visual Database of Emotional Speech and Song”. The RAVDESS dataset contains both audio and video samples. However, for the purposes of our project, only the audio samples were considered.

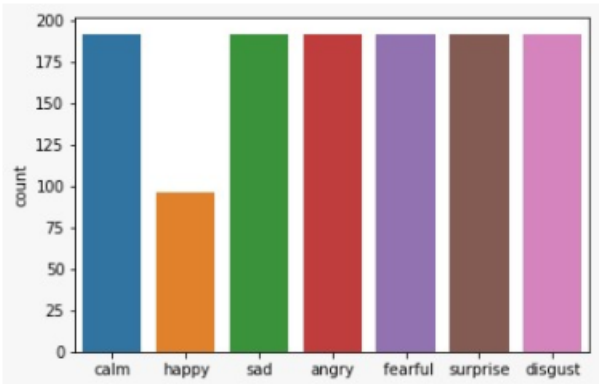

Figure 2. Distribution of emotion classes of the RAVDESS dataset

TABLE 1. PERFORMANCE SUMMARY ON RAW RAVDESS

Class	Precision	Recall	F1-Score
Calm	0.47	0.66	0.55
Happy	0.67	0.43	0.52
Sad	0.84	0.75	0.79
Angry	0.57	0.51	0.54
Fearful	0.48	0.76	0.56
Surprise	0.76	0.66	0.70
Disgust	0.59	0.50	0.54

TABLE 2. PERFORMANCE SUMMARY ON AUGMENTED RAVDESS

Class	Precision	Recall	F1-Score
Calm	0.64	0.73	0.69
Happy	0.91	0.53	0.67
Sad	0.77	0.81	0.79
Angry	0.71	0.71	0.71
Fearful	0.67	0.75	0.71
Surprise	0.77	0.82	0.79
Disgust	0.82	0.71	0.76

4.1.1. About RAVDESS.

- Gender balanced consisting of 24 professional actors.
- 7 emotions in total: calm, happy, sad, angry,fearful, surprise, and disgust expressions
- Each expression is produced at two levels of emotional intensity, with an additional neutral expression.
- 1250 samples in total.

A distribution of the emotion classes is shown in Figure 2.

4.1.2. Performance on RAVDESS. The net average accuracy obtained by our model on the raw and augmented data are 61.2% and 73.5% respectively. Detailed summaries of our system on RAVDESS are shown in Table 1 and Table 2.

4.2. TESS

TESS [16] stands for ”Toronto Emotional Speech Set”.

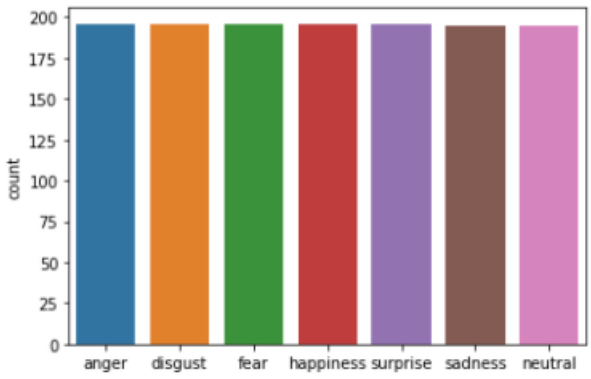

Figure 3. Distribution of emotion classes of the TESS dataset

TABLE 3. PERFORMANCE SUMMARY ON RAW TESS

Class	Precision	Recall	F1-Score
Anger	0.97	0.95	0.96
Disgust	0.97	1.00	0.99
Fear	0.96	0.90	0.93
Happiness	0.87	0.97	0.92
Surprise	1.00	0.97	0.99
Sadness	1.00	1.00	1.00
Neutral	1.00	1.00	1.00

TABLE 4. PERFORMANCE SUMMARY ON AUGMENTED TESS

Class	Precision	Recall	F1-Score
Anger	0.98	0.99	0.98
Disgust	1.00	0.98	0.99
Fear	0.97	0.97	0.97
Happiness	0.97	0.97	0.97
Surprise	0.99	1.00	0.99
Sadness	0.99	0.99	0.99
Neutral	1.00	0.99	0.99

4.2.1. About TESS.

- Gender balanced consisting of 2 professional actresses (aged 24 and 64 years).
- 7 emotions in total: anger, disgust, fear, happiness, surprise, sadness, neutral
- Audiometric testing indicated that both actresses have thresholds within the normal range.
- 1370 samples in total.

A distribution of the emotion classes is shown in Figure 3.

4.2.2. Performance on TESS. The net average accuracy obtained by our model on the raw and augmented data are 96.7% and 98.6% respectively. Detailed summaries of our system on TESS are shown in Table 3 and Table 4.

5. Conclusion

End-to-end emotion recognition approaches provide a new perspective for various applications since the speech directly maps to emotion (Emotion in different languages using random vector and context vector approach). In this paper, we proposed an ASR free end-to-end machine learning based approach for the emotion classification task. The experiment results show that our proposed approach performs very well on the ravdess dataset. We will train and evaluate the architecture on other emotion related datasets in future work.

References

[1] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on affective computing*, vol. 6, no. 1, pp. 69–75, 2015.

[2] D. D. Joshi and M. Zalte, "Recognition of emotion from marathi speech using mfcc and dwt algorithms," *International Journal of Advanced Computer Engineering and Communication Technology (IJACECT)*, vol. 2, no. 2, pp. 59–63, 2013.

[3] R. Subhashree and G. Rathna, "Speech emotion recognition: Performance analysis based on fused algorithms and gmm modelling," *Indian Journal of Science and Technology*, vol. 9, no. 11, pp. 1–8, 2016.

[4] A. Milton, S. S. Roy, and S. T. Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[5] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[6] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[7] J. Zhou, Y. Yang, and P. Chen, "Speech emotion recognition based on rough set and svm," vol. 1, 08 2006, pp. 53–61.

[8] G. Sahu, "Multimodal speech emotion recognition and ambiguity resolution," 04 2019.

[9] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162 – 1171, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639311000884

[10] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603 – 623, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639303000992

[11] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 239–246, 2017.

[12] L. Zheng, Q. Li, H. Ban, and S. Liu, "Speech emotion recognition based on convolution neural network combined with random forest," in *2018 Chinese Control And Decision Conference (CCDC)*. IEEE, 2018, pp. 4143–4147.

[13] E. Palogiannidi, I. Gkinis, G. Mastrapas, P. Mizera, and T. Stafylakis, "End-to-end architectures for asr-free spoken language understanding," 2020.

[14] C. Wang, Y. Wu, Y. Du, J. Li, S. Liu, L. Lu, S. Ren, G. Ye, S. Zhao, and M. Zhou, "Semantic mask for transformer based end-to-end speech recognition," 2020.

[15] S. R. Livingstone and F. 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*, vol. 13, no. 5, pp. 1–35, 05 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391

[16] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF