SPEECH EMOTION RECOGNITION SYSTEM

Dileep Sai Ellanki
Data Science, Indiana University Bloomington

Abstract

In recent years we have witnessed an increase in the amount of focus placed on the exploration of emotional speech signals in human-machine interfaces as a direct result of the availability of high computation capabilities. The research on this topic has presented a great number of different methods for determining an individual's emotional state based on their speech. The most important challenges that speech emotion identification systems must overcome are the selection of adequate feature sets, the creation of effective classification algorithms, and the preparation of an appropriate dataset. In this study, a comprehensive analysis of the currently existing approaches to speech emotion identification systems based on the three criteria for judging was carried out (feature set, classification of features, accurate usage). In addition to this, it sheds light on the current potential direction for the development of improvements to speech-emotion recognition systems with the help of machine learning and deep learning models. Nowadays many companies are working on sentimental analysis where they mainly deal with how customers are reacting towards them. This study can draw insights that help the growth of the company. Keywords- Emotional speech signal, Machine learning, Deep learning, Sentimental analysis.

1 Introduction

Speech Emotion Recognition, or SER for short, is the process of identifying the feelings sent by a speaker's words, regardless of the semantic content of the words. Research is still being done to figure out how to make programmable devices capable of performing this work automatically, despite the fact that people are capable of carrying it out in an effective manner as a natural component of spoken communication.[10]

The goal of research into automatic emotion recognition systems is to develop effective and real-time methods for determining the feelings of people who use human-machine communication devices such as mobile phones, call center operators and customers, car drivers, and pilots, amongst a wide variety of other individuals. According to André et al. (2004), one of the most important aspects of giving robots the ability to seem and behave in a human-like manner is giving them the ability to feel emotions.

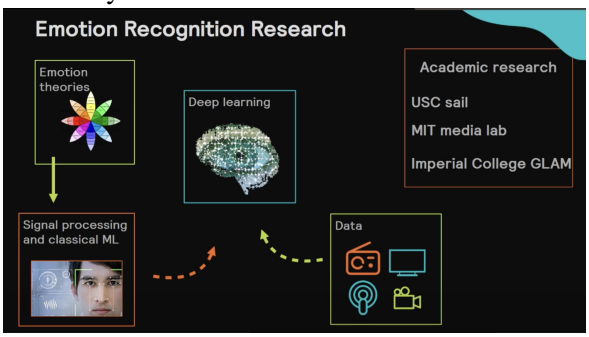


Fig 1.1: Intro to SER

Robots that are capable of comprehending human emotions could display emotional personalities and deliver reactions that are suited to those feelings. There are some situations in which people could be replaced by computer-generated characters who are capable of holding dialogues that are very genuine and convincing by appealing to human emotions. It is necessary for machines to comprehend the feelings that can be transmitted through speech. Only with this capability is it possible to develop a completely meaningful discourse between humans and machines that is founded on mutual trust and comprehension.

Traditionally, machine learning (ML) entails deriving feature parameters from raw data. These

feature parameters are then used in an algorithm (e.g., speech, images, video, ECG, EEG). A model is trained with the features so that it can learn to create the output labels that are wanted. The selection of features is frequently one of the challenges that this method must overcome. In general, it is unknown which features, when combined, can result in the most effective clustering of data into its respective groups (or classes). By evaluating a large number of distinct features, integrating distinct features into a single feature vector, or employing a variety of strategies for feature selection, one can acquire some insights. It is possible that the performance of the categorization system will be significantly impacted by the quality of the hand-crafted features that are produced[10].

2 Literature Review

The development of deep neural network (DNN) classifiers has provided a clever workaround that sidesteps the issue of selecting the best possible features in a way that is both elegant and effective. The plan is to make use of a network that is complete from beginning to finish and has as its input raw data and produces a class label as its output. It is not necessary to create features by hand or to compute them, nor is it necessary to identify which parameters produce the best results from a classification standpoint. Everything is handled by the network on its own. Specifically, during the training process, the network parameters, which include the weights and bias values that are assigned to the network nodes, are optimized so that they may effectively function as features that divide the data into the categories that are required. In comparison to more traditional approaches to classification, this solution's use of labeled data samples necessitates a substantially higher number of them. This is the trade-off for the otherwise relatively convenient technique.

In many instances, and SER is one of them, just a little amount of data is available for use in the training process. According to the findings of this research project, the challenge of insufficient training data can be largely circumvented by employing a methodology known as transfer learning. In order to tackle a general problem involving categorization, it makes advantage of an existing network that has already been trained on substantial data. This network is then trained (fine-tuned) using a smaller number of accessible data to accomplish a more specific task in order to improve its performance.

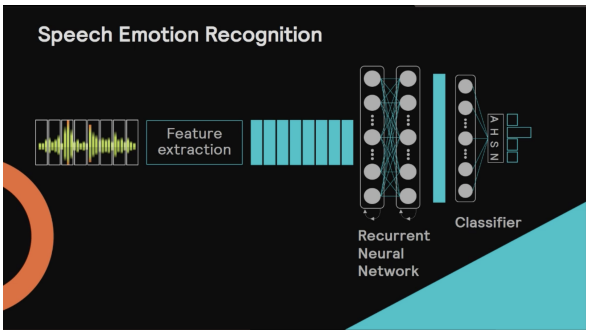


Fig 2.1: SER Flow

Given that, at the moment, the most powerful pre-trained neural networks were built for image classification, in order to apply these networks to the problem of SER, the speech signal needs to be translated into an image format. This can be done by using a specialized program (Stolar et al., 2017). This study explains the training and testing procedures, as well as the conditions that need to be met, in order to achieve real-time emotion recognition from continuously streaming speech. The steps involved in the transition from speech to image are described, as are the conditions that need to be met. The impacts of speech companding and bandwidth reduction on the real-time SER are examined. This is because many programmable speech communication platforms use speech companding and compress speech bandwidth to a small range of 4 kHz.

The necessity to establish a collection of key emotions that can be classified by an automatic emotion recognizer is a significant challenge that must be overcome in the field of speech emotion recognition. Linguists have created inventories of the emotional states that are most common in our lives and have categorized them. The authors Schubiger, O'Connor, and Arnold present a typical set, which includes a total of 300 different emotional states. Having said that, it is quite challenging to categorize such a big number of feelings.

The 'palette theory,' which asserts that each feeling can be broken down into primary feelings in a manner analogous to the way that any color is a combination of a few fundamental hues, has garnered the support of a significant number of researchers. Anger, disgust, fear, joy, and sadness, as well as surprise, are some of the primary emotions. These are the feelings that stand out as the clearest and most distinguishable examples of

human experience. These feelings are known as archetypal.[1]

3 Methodology

In the path of automatic speech recognition, there were various implementations based on SVM, GMM, and HMMs. In early attempts to recognize emotions from the speech signal, almost all of the implementations were based on machine learning and signal processing approaches. They were implementing frameworks that were based on machine learning algorithms, which required extensive feature engineering and a deep understanding of the subject matter in order to be able to infer the features that were helping them the most to bring them into the calculations. In addition to this understanding, the frameworks also required extensive feature engineering. In this section, we will examine two of the approaches that are based on HMM and SVM in order to construct a foundation for the subsequent generation of algorithms.[2]

3.1 Data Collection

We used a ravdess data set which has the audio of 20 different actors, and they are distributed between various emotions. These emotions include surprise, neutrality, disgust, fear, sadness, calm, happiness, and anger[6][7].

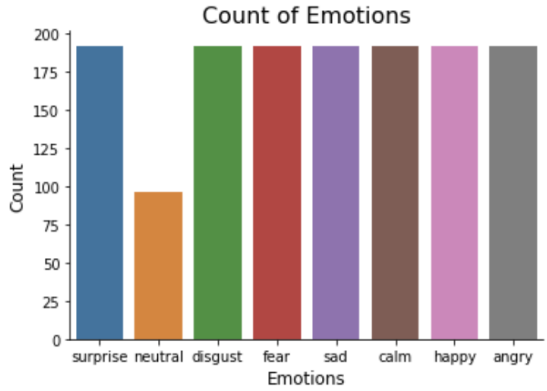


Fig 3.1.1: Count of emotions

Then we created the functions to generate the waveforms and spectrograms. From these, we can identify the emotion of a person on various factors like the pitch of the person speaking, and the tone of the person. We have generated the spectrogram of various emotions and waveforms Waveplots are useful for obtaining information about the relative volume of the sound at a particular instant in time.

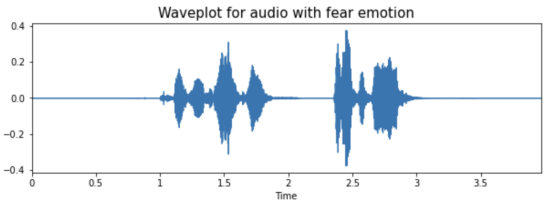


Fig 3.1.2: Wave Form Output

A spectrogram is a graphical representation of the spectrum of frequencies of a sound or other signal as it changes over the course of time. The analysis of data and sounds can be done with the help of a spectrogram. It is a representation of the way in which the frequencies of various audio or music signals vary over the course of a particular period.

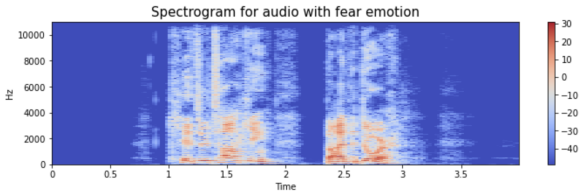


Fig 3.1.3: Spectro gram output

3.2 Data Agumentation

The process of generating brand-new synthetic data samples by making relatively insignificant alterations to the initial training set is referred to as the data augmentation technique. The approach that we utilize is known as data augmentation. The process of producing syntactic data for audio can involve the addition of noise, the shifting of time, the modification of pitch, and the acceleration of the sound. The objective here is to render our model immune to the effects of those kinds of disturbances while also increasing its ability for generalization as much as is humanly achievable. The addition of the perturbations while retaining the same label as the primary training sample is necessary for this to be successful. It is possible to improve the quality of the data on images by rotating the image, zooming in or out on the image, or shifting the image.

3.3 Feature Extraction

Using the sample rate and the sample data, one can conduct a variety of transformations on it to extract useful characteristics from it. The pace at which the signal's sign changes during the time that a given frame is being shown is referred to as the Zero Crossing Rate. Energy is defined as the sum of squares of the signal values after they

have been normalized by the length of the respective frame. The entropy of the normalized energies of the sub-frames is referred to as the entropy of energy. It is possible to understand it as a measurement of sudden shifts.[3] The spectral centroid is also known as the spectrum's gravitational center. The second central moment of the spectrum is referred to as the spectral spread. The entropy of the normalized spectral energy for a given set of sub-frames is referred to as spectral entropy.

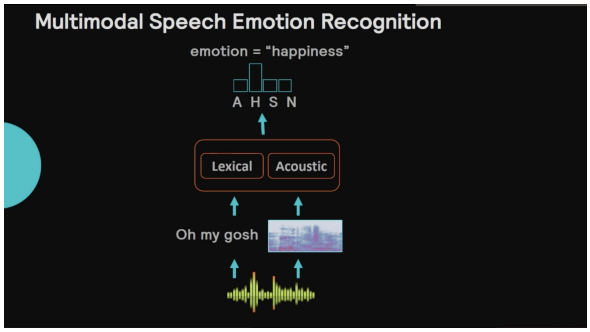


Fig 3.3.1: Feature extraction

Spectral Flux is defined as the squared difference between the normalized magnitudes of the spectra in the two frames that immediately follow each other. The frequency below which 90 percent of the spectrum's magnitude distribution is concentrated is referred to as the spectral roll- off frequency. MFCCs Mel Frequency Cepstral Coefficients are responsible for the formation of a cepstral representation, in which the frequency bands are not linear but rather scattered according to the Mel-scale.

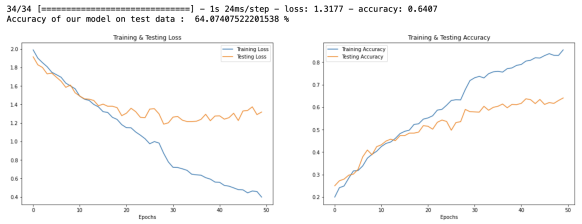
The Chroma Vector is a representation of the spectral energy that consists of 12 elements, each of which corresponds to one of the 12 equal-tempered pitch classes used in western- style music (semitone spacing). The Chroma Vector is a representation of the spectral energy that consists of 12 elements, each of which corresponds to one of the 12 equal-tempered pitch classes used in western- style music (semitone spacing). The standard deviation of the 12 chroma coefficients is referred to as the chroma deviation.[3]

3.4 Proposed method

The proposed method is taking the audio samples and extracting the features from them. After extracting features then classifying the data into emotions including surprise, neutrality, disgust, fear, sadness, calm, happiness, and anger. The convolution neural network is used to classify emotions.

4 Experimental results

Finally, after executing all the steps and processes, the classification results are acquired after running the convolution neural network.



Img 4.1: Model Output

	Predicted Labels	Actual Labels
0	fear	fear
1	neutral	neutral
2	sad	sad
3	neutral	neutral
4	fear	fear
5	angry	angry
6	fear	fear
7	fear	fear
8	disgust	disgust
9	calm	calm

Img 4.2: Predicted output

After running the model we got an accuracy of 64 percent and tested the model with sample inputs we got an accurate prediction of outputs.

5 Conclusion

By using spectrogram and waveform we can see how speech signals are. After analyzing the signals, we extracted the features from them. Data augmentation is used to see how these signals vary when noise is there, and how signals vary when it lags. After extracting features, we used CNN to train the model for the classification of the emotions and predicting them.

6 References

[1]El Ayadi, Moataz, et al. "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases." Pattern Recognition,

vol. 44, no. 3, Mar. 2011, pp. 572–587, 10.1016/j.patcog.2010.09.020. Accessed 14 Oct. 2019.

[2]Abbaschian, Babak Joze, et al. "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models." Sensors, vol. 21, no. 4, 10 Feb. 2021, p. 1249, 10.3390/s21041249.

[3]Medium. (n.d.). Medium. Retrieved December 14, 2022, from https://medium.com/heuristics/audio-

[4]Ramakrishnan, S., and Ibrahiem M. M. El Emary. "Speech Emotion Recognition Approaches in Human Computer Interaction." Telecommunication Systems, vol. 52, no. 3, 2 Sept. 2011, pp. 1467–1478, link.springer.com/article/10.1007

[5]Amazon re:MARS. (2019). Speech Emotion Detection. In YouTube. https://www.youtube.com/watch?v=26$_q iXEa8lw$

[6]"RAVDESS Emotional Speech Audio." Www.kaggle.com, www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio.

[7] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)." Zenodo, Zenodo, 5 Apr. 2018, zenodo.org/record/1188976.Y5lSbezMKw1. Accessed 14 Dec. 2022.

[8]Dey, Lily, et al. "Emotion Extraction from Real Time Chat Messenger." IEEE Xplore, 1 May 2014, ieeexplore.ieee.org/abstract/document/6850785?casa$_token =$
$E6SJmQWKFt0AAAAA$:
$MJ0ELRh2RuEUrC8SlMkbdoqSlxOU$ –
$GrHwEWR3hSN0q$ –
$GODNM6TbXBqtbCX9iA2rULGpah60o_g.Accessed14Dec.$2022.

[9]Ververidis, Dimitrios, and Constantine Kotropoulos. "Emotional Speech Recognition: Resources, Features, and Methods." Speech Communication, vol. 48, no. 9, Sept. 2006, pp. 1162–1181, 10.1016/j.specom.2006.04.003. Accessed 16 Oct. 2019.

[10] Lech, Margaret, et al. "Real-Time Speech Emotion Recognition Using a Pre-Trained Image Classification Network: Effects of Bandwidth Reduction and Companding." Frontiers in Computer Science, vol. 2, 26 May 2020, 10.3389/fcomp.2020.00014.

https://tinyurl.com/2hk4ejm5