# Audio Based Customer Satisfaction Measurement using Deep Learning.

Sunil Judhistira Gauda

July 2022

Abstract: The consumer serving industry face massive problems regarding the quality of the service provided, as the satisfaction of customers depends on varying sentiments expressed in the call, Using AI and Audio Processing we can create a quality analysis pipeline that will serve as a basis for customer satisfaction measurement, using audio's of a customer having a conversation with an operator, we will extract the sentiments and provide results as the sentiments of the conversation. This will help us rate the scale of satisfaction from fear, anger, disgust, neutral, pleasant surprise, or happy.

Index Terms/Keywords: Audio Processing, Mel-frequency cepstral coefficients (MFCC), Long Short-Term Memory, Deep Learning.

## 1 Introduction.

Customer cares have audio-based support systems, Even with the improvement in chat-bots adding a human voice makes a customer comfortable providing accurate descriptions of the issues they face, so call centers are more preferred than text-based solutions like chat-bots, email, and SMS. It is difficult to manage the quality of the service when calls are involved, even when the entire it is recorded, quality assurance requires manual work, and is labor intensive. When the quality is poor, the time it takes to get feedback can be questionable. To automate this process we can introduce a system that identifies a sentiment of a conversation and provides values that can measure the quality of a call. Automated quality management applications would reduce the need for continuous management of employee output and provide a low-stress environment for the company, call center operatives often face strict monitoring and regulation issues due to poor performance in stressful environments like this. Especially when we look into statistics where most of the customer support industries are situated are developing countries, where access to mental health care is poor due to lack of money or infrastructure support. To solve this problem we will develop an Audio Processing Pipeline with a deep learning model to analyze the sentiment of a call using which we can rate a call quality.

The Idea to create the process derives inspiration from [1] to establish the basic layout of the project. Author [2] provides insights on Mel-Frequency Cepstral Coefficient (MFCC), and Long Term Short Term Memory (LSTM) which was used to derive sentiments from the audio of variable length. Author [2] has then developed a CNN-based solution to produce results. The idea for choosing the data was derived from Author [3] paper, Author has

used the TESS Data set to train models for children with ASD. The TESS data-set when balanced provided near 100 percent on test, train, and validation with great recall, accuracy, and precision.

Author [4] did Opinion Mining, which is a similar approach to customer satisfaction measurement, but without technology like LSTM in it. Due to advancements in deep learning, we can build a better model of Opinion Mining with better accuracy.

## 2 State Of the Art.

Currently, Audio-based sentiment analysis is used in chat-bots for automation using IVR mediums and google dialogue flow, extraction of keywords and topic modeling is done. Audio based text extraction is used widely to generate captions in streaming platforms. Sentiment analysis is used by author [3] to help communicate children with ASD (Autism Spectrum Disorder). Author [4] did audio based opinion mining which does not use modern deep learning technology like LSTM and MFCC.

Field of Customer Satisfaction Measurement, Opinion Mining has models that depend on learning long-term dependence with sequence prediction problems like quality management and review analysis, When destructed the problems faced are usually one sequence of keywords after another, which can be extracted with the use of Topic Modelling, but the customer satisfaction is the primary goal of the entire system, for extracting sentiments there are no long-term learning models working on it.

## 3 Methodology.

Proposed method uses TESS Audio data set to train an LSTM based RNN model, using MFCC to select the features of the data, this will

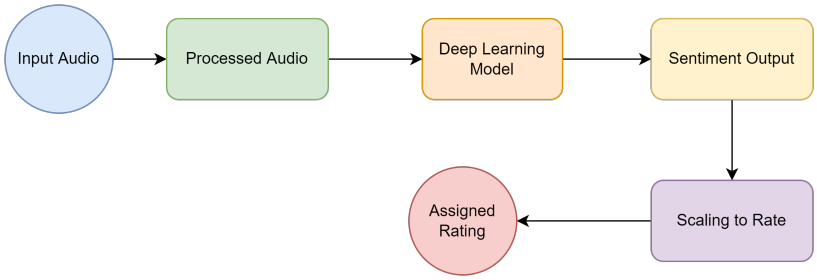


Fig 1. Work-Flow Block Diagram.

### 3.1 Preparing Data-set Toronto emotional speech set (TESS).

TESS Data-set contains 7 set of audio for training 7 sets of emotions i.e, fear, angry, disgust, neutral, sad, pleasant surprise, happy. To make the data-set accessible for exploratory analysis and prepare it for our model we need to follow the steps below :

1. Read all the folders in the directory's and extract paths and labels from location and names of the files, the data-frame should contain the following things.

| Path | Labels |
|---|---|
| ../../../ Data- set Location | fear |
| ../../../ Data- set Location | pleasant surprise |
| ../../../ Data- set Location | happy |

2. Exploring the Data: To Understand more of the data we can visualise the data, a wave diagram can provide a visual output of the Label.

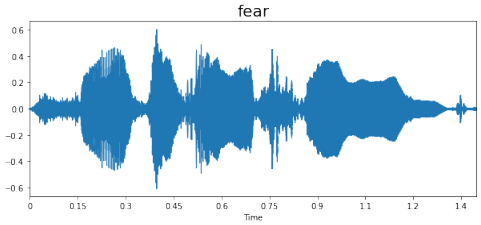


Fig 2. Wave Diagram of a sample audio with "Fear" as a label.

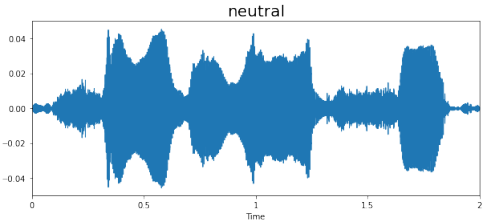


Fig 3. Wave Diagram of a sample audio with "Neutral" as a label.

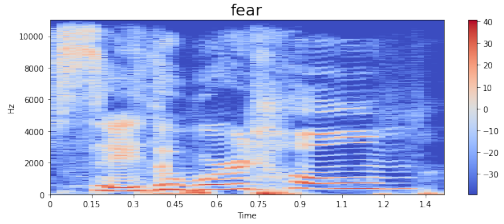


Fig 4. Spectogram of a sample audio with "Fear" as a label.

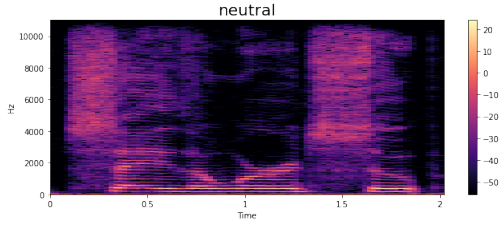


Fig 5. Spectogram of a sample audio with "Neutral" as a label.

From the above figure we observe the differences between each sample emotions, the spectogram especially shows a stark difference between the audio in colour density.

3. Feature Extraction: We use MFCC to convert an audio signal transformation to extract features from each frame of audio, below is the path to MFCC.

(a) A/D Conversion - Converting to digital for model to learn.

(b) Preemphasis - Adjusting sounding Vowels.

(c) Windowing - Chopping the audio signal.

(d) DFT - Taking Discrete Fourier to make analysis easy as audio is time dependent.

(e) Mel-Filter Bank - Mapping actual frequency to human perspective.

$$mel(f) = 1127 \ln(1 + \frac{f}{700})$$

Fig 6. Mel-filter for human audio.

(f) Log Transformation - To match the input and output gradient.

(g) IDFT - Inverse Transform to Accommodate Human audio Frequency.

(h) Dynamic Feature - Considering First order or second order Derivatives.

## 3.2 Modelling LSTM

Modelling LSTM is done with tensorflow keras, The LSTM Model will be added to keras module as follows.

1. Create a Sequential Block
2. Add an LSTM Block
3. Add Dense Layers with activation functions

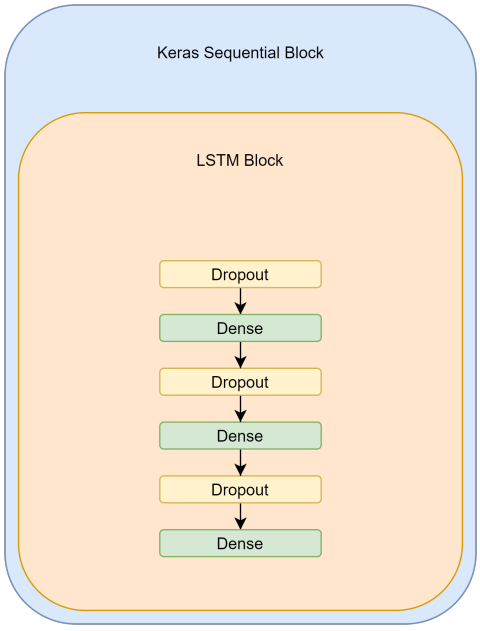


Fig 7. Pseudo Layout.

The outputs for this CNN network will not all be in the same format. Each output will be in a separate format required for the RNN Network in the song prediction to follow. The outputs of the CNN layer will be as:

- Emotion: A String variable that will be selected from a fixed dictionary of emotions like Sad, Angry, Happy or Excited.
- Age: This will be a Boolean with 0 if the individual is older than roughly 20 and 1 if younger than 20.
- Gender: Boolean 0 for Male, 1 for Female.
- Race: This too will be a string of Racial Background.
- Surrounding: Boolean 0 for Indoor and 1 for Outdoor.
- Alone: Boolean 0 for Alone and 1 for Group/Party.

### 3.3 Comparing Sentiment Output and Converting it to Rating

With Our model providing us the output in terms of fear, angry, disgust, neutral, sad, pleasant surprise, happy, A web application could match the emotions to rating's from 1-5, where each point will determine the quality of conversation and performance of the employee, the application will facilitate can accomodate following points .

- Capture/Upload Audio: In this stage an audio clip can be captured or uploaded to provide source for the model
- Process : The Audio will go through the process pipeline to make it ready for the model to consume
- Predict : Model will predict the outcome
- Result : Application will provide result of the prediction within scale 1-5

## 4 Expected Outcome.

The Final Result will be a AI powered application which can take an audio clip and derive sentiments from it, then convert it to numerical rating so that it could be used for evaluation in real world, the rating will be a result of quality of conversation that a customer care representative has and can later justify how well the employee performs


Fig 8. High level Overview.

## 5 Conclusion.

The use of such cleaver systems can provide an automated process of quality management, in this case it is implemented over a call center data, with few minor adjustments it can be used in critical places like emergency services where quality of the conversation is of serious matter, places like police hot-lines and suicide prevention hot-lines are prime places to implement such measures.

## References


[1] A. Rao, A. Ahuja, S. Kansara, and V. Patel, "Sentiment analysis on user-generated video, audio and text," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. IEEE, 2021, pp. 24–28.

[2] S. Suganya and E. Charles, "Speech emotion recognition using deep learning on audio recordings," in *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, vol. 250. IEEE, 2019, pp. 1–6.

[3] D. Valles and R. Matin, "An audio processing approach using ensemble learning for speech-emotion recognition for children with asd," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0055–0061.

[4] A. L. Rane and A. R. Kshatriya, "Audio opinion mining and sentiment analysis of customer product or services reviews," in *ICDSMLA 2019*. Springer, 2020, pp. 282–293.