Audio Sentiment Analysis after a single-channel Multiple Source Separation

Shivani Firodiya, Arpit Shah
School of Informatics, Computing and Engineering, Indiana University

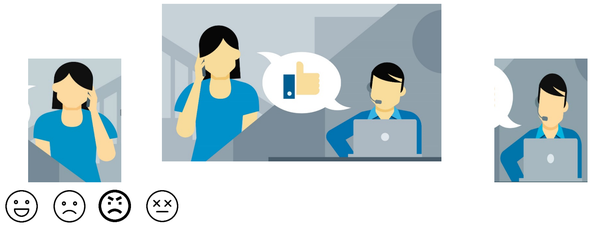
ABSTRACT

Call Centers or Support Centers in different companies aggregate huge amount of data everyday. From all the conversations, few conversations are not customer satisfactory i.e sometimes the customer is not satisfied with the support. Finding the sentiment of the customer helps in determining whether the customer was satisfied with the service.
We in this project aim at separating the sources from the conversation and classifying the sentiment of every chunk (speaker 1 and speaker 2) of the audio. We approach this problem in three stages:
Stage 1 - We perform source separation on the audio conversation, by performing VAD detection on the conversation and dividing the audio conversation into different chunks, on each chunk we apply GMM and a global GMM (UBM) on the whole conversation, using BIC and through spectral clustering we cluster every chunk into different speakers.
Stage 2 –We then apply sentiment analysis on supervised speech emotion dataset (RAVDESS) using Deep Neural Networks.
Stage 3 - We used the trained model from Stage 2 to classify the sentiment of the speaker chunks.

Keywords - Speaker Diarization, Deep Neural Network, GMM, Spectral Clustering, Emotion Detection.

MOTIVATION

- Call centers aggregate a large amount of data through calls. There is a need to build a model which can help us analyze how did the conversation between the agent and the customer go. The customer might have various emotions throughout the conversation.



- Since the agent's emotion might be neutral it might affect the overall sentiment of the conversation.
- We aim to separate these conversations into chunks to better understand the sentiment of the customer.

DATASETS

- Data for source separation was taken from **EXOTEL** which consists of 300 audio files. Each of these files contains conversation between the customer and agent on various topics.
- These audio files are labelled into four categories viz. Angry, Sad, Happy, Neutral.

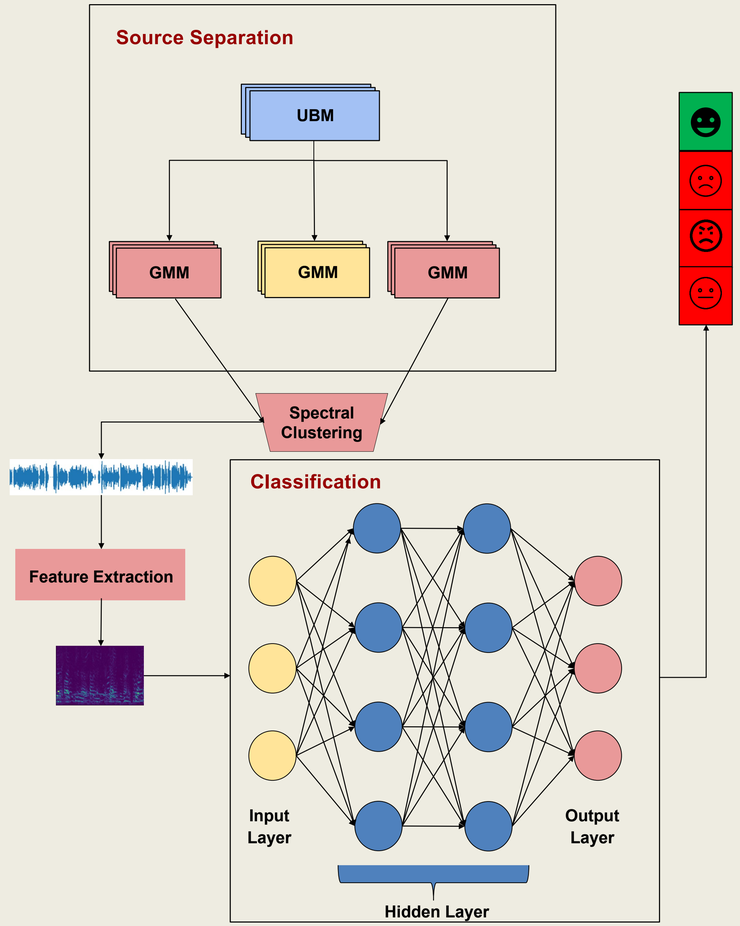
- **RAVDESS** dataset was used for classification purpose, it consists of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent
- Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions.

ARCHITECTURE

The architecture consists of two parts:

- The audio file is divided into chunks using VAD detection, a global GMM (UBM) is trained on the complete file. Separate GMM is trained on every chunk using the means and priors of the UBM for training. The weights of these GMM are compared using BIC criterion. Spectral clustering is applied on the nearest stacked GMM weights to cluster the chunks into two different speakers.

- From the **RAVDESS** data, we extracted 6 features: **MFCC, STFT, Contrast, Mel Spectrum, Chroma and Tonnetz**. These features are passed to a Deep Neural Network composing of 4 hidden layers, and this model is used to predict the sentiment of above chunks obtained using GMM_UBM Spectral Clustering.



DISCUSSION

- Audio files of conversation between customer and agent were first divided into separate chunks. Chunks of customer audio were then given to the classification model.
- Depending on the features of the chunks, a sentiment was classified which tell us about the sentiment of the customer over the whole conversation.
- The experiment was performed using two different algorithms.
  - XGBoost
  - Deep Neural Networks(DNN)
- We experimented with two different DNNs, for the first one we used Tanh and Sigmoid activation function for the hidden layer and softmax for the output layer. For our second DNN, we used relu for all the hidden layer and softmax for the output layer.

MODEL	TRAINING ACCURACY	TESTING ACCURACY
XGBOOST	80.00%	56.00%
DNN Activation – Tanh, Sigmoid, Softmax Optimizer - adadelta	97.00%	63.00%
DNN Activation – Relu,Softmax Optimizer - adadelta	91.94%	70.00%

CONCLUSIONS

- In this project ,we implemented a Deep Neural Network with different hyper parameter tuning on acoustic features from the RAVDESS dataset. This saved model was used for classifying sentiment on the source separated files. From the results, we came to the conclusion that the sentiment of customer varies with the overall sentiment of the whole conversation.
- Since the data we used for classification only involved 1440 samples, if trained on more data, the classification model will perform with good results. Also for VAD detection, if we use Unsupervised machine learning techniques, the chunking can be smoothed and the GMM will give more distinct clustering results.

REFERENCES

1. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) - https://zenodo.org/record/1188976#.XA1fK2hKhPb
2. GMM-UBM based open-set online speaker diarization - https://www.academia.edu/19904434/GMM-UBM_based_open-set_online_speaker_diarization