Audio Sentiment Analysis after a Single-Channel Multiple Source Separation

Arpit Shah
Indiana University Bloomington
arpishah@iu.edu

Shivani Firodiya
Indiana University Bloomington
shfirodi@iu.edu

Abstract— **This paper aims at speaker diarization on an audio clip and sentiment analysis. Audio segmentation is done on the audio file using Voice Activity Detection (VAD). Once the audio is segmented, speaker identification is done using MAP estimation on every chunk using Universal Background Model (UBM) and Gaussian Mixture Model (GMM). Every chunk represents a different GMM while UBM represents a GMM on the whole audio file. Speech clustering is done using spectral clustering on every audio segment. Audio sentiment analysis is performed on a supervised emotion dataset (The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)) using Deep Neural Networks (DNN). The trained model was used to classify the sentiment of every chunk of the audio clip.**

Keywords: Speaker Diarization, Sentiment Analysis, GMM, UBM, DNN

I. INTRODUCTION AND BACKGROUND

Analyzing the sentiment of Audio can be done by transcribing the audio into text. However, these transcripts are difficult to read and most of the times do not capture all the information contained in the audio. To build such a text to speech system (TTS) which can be used to analyze sentiment, is a time consuming process and it requires a lot of pre-processing since we need to eliminate a lot of text from it which is not reasonable. This leads to data loss too. This technique is essentially speaker-dependent since based on the utterance of the speaker , the quality of text transcribed will also vary. A lot of techniques have been employed to analyze the sentiment of the TTS system [3]. It used Fuzzy Neural Network to perform sentiment analysis on the Text derived from the speech. The dataset used was Doordarshan Tamil News Dataset.

However relying on speaker-dependent data requires large amount of data to perform sentiment analysis. Therefore, we need a lot of text data labelled with emotion for each speaker. This is time consuming and unreliable since gathering huge amount of text data of a speaker is cumbersome. Therefore, [1] proposed that using acoustic features gives better performance and is speaker-independent. All the acoustic features do not contribute to the sentiment variation. According to [2], prosodic features and acoustic features such as power and pitch contribute to the sentiment variation. SVM was used for performing emotion classification. Power, pitch and its second-order derivatives were used as features for performing classification.

However in some cases, the audio may contain multiple sources in a single channel. The audio sources could be music segments, noise, clapping or multiple speakers. In cases when there are multiple speakers like a telephonic conversation, the sentiment of the audio may get dominated with the sentiment of the speaker speaking for longer duration. Separating these sources is essential in these cases. Separating the speakers from an audio conversation where the speakers are speaking on a turn-by-turn basis with little overlap is Speaker Diarization. It involves removing the non-speech data like noise, pause-in between the conversation to separate the segments into different chunks with voiced and unvoiced data. Since audio many involve different number of speakers, [5] used agglomerative clustering. It does segmenting of data into many small pieces. Each piece is assigned to a cluster. The system then iterates and merges clusters based on modified version of BIC. In [4] a modified version of [5] used acoustic fusion of all available channels into a single enhanced channel via delay and sum beam-forming algorithm. Later speaker diarization was performed by using un-biased estimator for variance along with minimum variance thresholding. A purification algorithm was used to clean the clusters with non-homogeneous data. However this is not the scope of this paper. [8] uses acoustic beam$-$forming based on weighted-delay and sum microphone array theory, which is a generalization of the well-known weighted-delay and sum beam-forming techniques [6],[7]. [9] used Deep neural network to learn embeddings on segmented data and a scoring metric was used to discriminate between pairs of embeddings (same-speaker vs different-speaker pairs classification). [10] is a model based approach and it uses GMMs. It creates 2 models, one for male, one for female and one for garbage. The models for the occurring speakers are created during run time as the corresponding speakers occur. It uses GMM-UBM for MAP adaptation.

In this paper, we have used [10] MAP adaptation method for Speaker Verification. For Speaker Segmentation, we have used Google's VAD detection, webrtc VAD. While for clustering, we used Kmeans, Spectral Clustering with cosine similarity between the neighbouring segments to cluster the segments into different speakers.

For Sentiment Analysis, [11] used SVM on the acoustic features extracted from the corpus containing emotional speeches with 721 short utterances expressing four emotions. In this paper, we used Deep neural network for sentiment classification on 6 features extracted from RAVDESS speech

dataset.
The trained classification model was used to predict emotion of the chunks generated from Speaker Diarization.

II. System Overview

Our system is mainly divided into two parts, source separation and classification. Once we have the audio file of conversation of customer and call center agent, we pass it through first phase of our model. It requires us to first segment the audio into parts and once segmentation is done using Google's webrtc VAD. These segments are passed to the Adaptive MAP estimation for Speaker Verification. The generated super-vectors are passed to Clustering algorithm to cluster every chunk with the speaker. The second part involves Audio Sentiment Classification. The Audio Sentiment classification is performed on RAVDESS dataset and the generated model is used to predict sentiment on the chunks generated in first phase of the system.

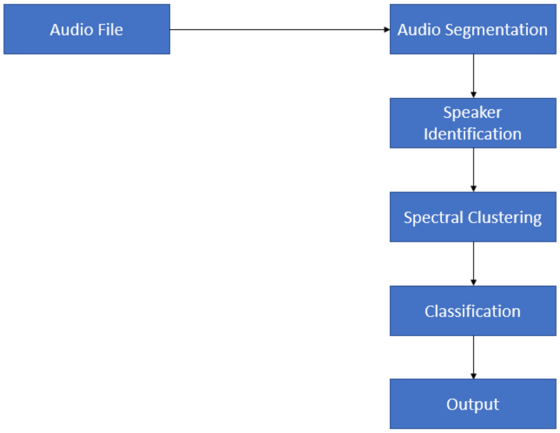


Fig. 1. System Workflow

A. *Speaker diarization*

*1) **Audio Segmentation**:* Input audio file is first segmented using VAD. VAD involves separating the voiced part from the unvoiced part. This is done by passing a sliding window over the whole audio with some overlap. Whenever there is a pause in the audio, segmentation process starts. We used window size of 32 ms and hop size of 10 ms. We used Google's webrtc vad for VAD detection.

*2) **Feature Extraction**:* After the audio file is segmented, feature extraction is performed on the segmented clips. Acoustic features such as MFCC, its first order derivative along with its second order derivative are used as features for the adaptive MAP estimation for speaker identification.

$$MFCC(f) = 1125 ln(1 + f/700)$$

*3) **Speaker Verification**:* We need to find the inconsistencies in channel and better represent the variations in two speaker. By the factor analysis approach, we convert the features into a super vector, a vector consisting of means trained using GMM. Since there are similar speech utterances in the whole audio, we train a Universal Background model (UBM) which is a GMM on the whole audio file. The UBM is speaker independent model. It represents features that are not dependent on the speaker. These feature vectors comprising of mean and covariance are used to train the individual models. The means and prior probabilities from the UBM are used as initial values for the means and priors of the individual GMMs on each chunks. The means of UBM are trained using EM algorithm, in the M step the means are updated using Adaptive Co-efficient.

$$\mu_c^s = \alpha_c \mu_c + (1 - \alpha_c)\mu_c^{ubm} \tag{1}$$

Here,

$$\mu_c = \text{Mean spectral vector of individual GMMs}$$
$$\alpha_c = \frac{n_c}{(n_c + r)}$$
$$r = relevance factor$$

*4) **Speaker Clustering**:* The supervector obtained from the Speaker Identification is passed to different Clustering algorithms with affinity as cosine similarity between each chunk to cluster the Gaussian mixtures into two speakers.

$$cosine\_score(w_1, w_2) = \frac{(w_1)^t (w_2)}{||w_1|| \cdot ||w_2||}$$

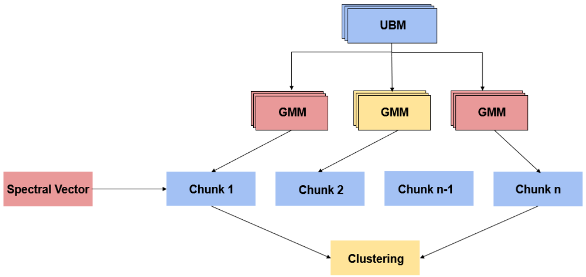


Fig. 2. Model for Source Separation

B. *Classification*

*1) **Feature Extraction**:* RAVDESS speech data is used for classification. Acoustic features such as MFCC, STFT, Contrast, Mel Spectrum, Chroma and Tonnetz are extracted from the audio clips of the dataset. We stack all the features to get a feature vector of 193 dimensions.

*2) **Training**:* The extracted features are passed to a DNN composed of 3 hidden layers with number of perceptrons to be 200, 400, 200 with relu activation for all 3 hidden layers, 2 dropout layers of 0.2. Last layer being the softmax layer and adadelta optimizer.

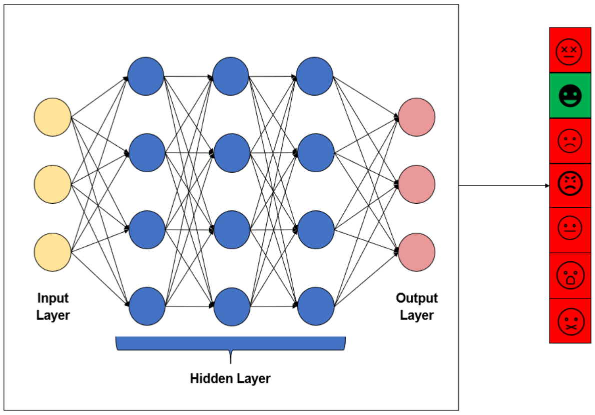


Fig. 3. Model for Classification

*3) **Testing**:* The model trained in the Training phase is used to test the sentiment of the chunks generated after source separation. Every chunk will yield sentiment which can help us understand the sentiment of the customer throughout the conversation.

III. Experiments and Results

*1) **Speaker Diarization**:* An audio file with telephonic conversation from Exotel was used. The dataset composes of 300 audio files with sentiment labelled for whole conversation. The files are labelled with four emotions such as sad, angry, happy and neutral. After performing segmentation on the audio file through VAD detection, chunks are obtained. These chunks contain voiced data which are passed to the Adaptive MAP estimation to get super vector. This super vector when passed to the clustering gives the chunk numbers that belong to Speaker 0 and the chunks that belong to Speaker 1. Label 0 or 1 is determined on the fact that which person starts to speaks first, first person to speak is assigned label 0 and other person is assigned label 1.

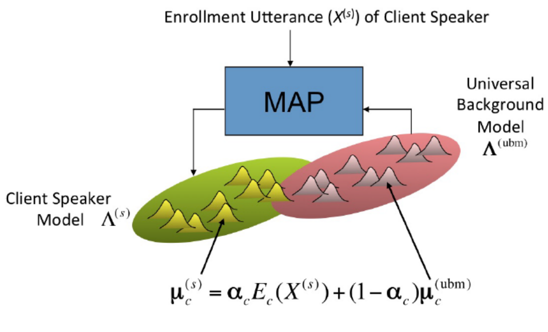

Enrollment Utterance ($X^{(s)}$) of Client Speaker
MAP
Universal Background Model $\Lambda^{(\mathrm{ubm})}$
Client Speaker Model $\Lambda^{(s)}$
$\mu_c^{(s)} = \alpha_c E_c(X^{(s)}) + (1 - \alpha_c)\mu_c^{(\mathrm{ubm})}$

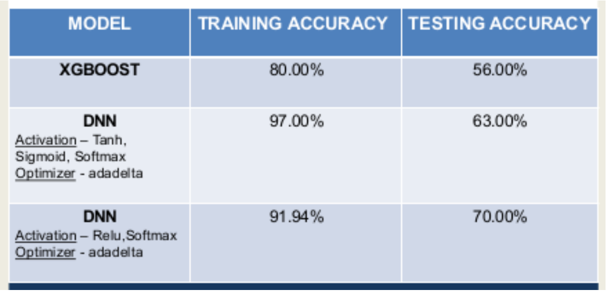
Fig. 4. MAP Adaptation

*2) **Clustering**:* Kmeans and Spectral clustering were used for clustering on the super vector obtained after MAP estimation. The labels obtained after clustering were verified manually. Even though most of the separation was done properly, there were few chunks in which overlap between the two people having conversation could be observed. The main reason behind this overlap was because there was no pause in the conversation which is why different chunks were not created during speaker diarization process. Performance of both the clustering algorithms were affected to a certain extent by overlap. From the above mentioned two clustering algorithms, spectral clustering gave good results compared to Kmeans. For both the algorithms, we take two clusters since we only have conversation of two people. For spectral clustering, affinity used was cosine since we are looking for the cosine similarity between the chunks to cluster them. After clustering is done, we then pass the chunks of customer audio to our classification model.

*3) **Classification**:* Classification model was trained on the RAVDESS speech dataset. This dataset comprises of audio files from 24 actors. These actor utter a sentence in 8 emotions, happy , angry, calm, sad, surprised, neutral, disgust, fearful. We train our model by extracting acoustic features of the audio clip. We use a total of 193 features of all the audio clips from our RAVDESS dataset for classification purpose. We use two different algorithms for classifying the sentiment of our input chunk. We first train our model on XGBOOST for which we directly pass our 193 features and labels of sentiment of that audio clip. XGBOOST gives us a training accuracy of around 80% and testing accuracy of only 56%. Due to low testing accuracy, we train our model using Deep Neural Networks. Two different types of DNNs were used for training the classification model. Both the DNNS were fully connected and consisted of one input layer, three hidden layers and output layer. For the first DNN, we take 200 perceptrons and a tanh activation function for the first hidden layer, 400 perceptrons and a tanh activation function for the second hidden layer and for our final hidden layer we used 200 perceptrons along with an activation function of sigmoid. For our second DNN, we use the same number of perceptrons as the previous one but the activation functions of the hidden layers change. We used relu activation function for all three hidden layers and softmax activation function for the output layer. The labels used were one-hot encoded to assign similar weight to each class. After our testing procedure, we get the sentiment of the chunk audio file that was given as input to the model. Training accuracy of 97% and testing accuracy of 63% was achieved from our fist DNN model while training accuracy of 91% and testing accuracy of 70% was achieved from our second DNN model.

IV. Conclusion and Future Work

We constructed a system in which an audio file of conversation between two people from the EXOTEL dataset was taken as the basic input which was then separated into different segments using VAD. Once the audio is segmented, we then cluster the chunks of customer together



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%

Fig. 5. Results

and then carry out sentiment analysis on all the chunks of customer that were created. During VAD, we use of sliding window of size 32 ms and hop size of 10 ms. Audio is segmented depending on the pauses that are present in the conversation, but if there is an overlap between two people's conversation then audio separation does not take place properly which in turn affects the clustering process and furthermore miss-classifies the sentiments of the chunks as well. So instead of using a sliding video and only relying on the pauses in the conversation, future scope would be to build an unsupervised model (Auto-encoders) that performs VAD detection and further cleans out the segments with non-homogeneous data.

Additionally, since most of the times, the customer is different and we only have 1440 audio files from RAVDESS dataset, classifier doesn't yield a very good performance on such a limited data. On increasing the size of dataset, the accuracy and classification results could be improvised. Also, on increasing dataset with a number of other languages, the model could be language independent. Currently it supports or performs better on English language.

Furthermore, in this system, sentiments were analyzed for chunks from the audio clip and variation was observed in each sentiment of the clip. However for a conversation, the sentiment of speaker 1 in a chunk and next consecutive chunk should not have much variation since its the same speaker speaking. Hidden Markov Models or Monte Carlo Markov Chain could be used to solve this problem since it accounts for the sentiment of the previous chunk and its transition to the next chunk.

References

[1] Poria, Soujanya, et al. "Fusing audio, visual and textual clues for sentiment analysis from multimodal content." Neurocomputing 174 (2016): 50-59.

[2] Navas, Eva, Inma Hernez, and Iker Luengo. "An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS." IEEE transactions on audio, speech, and language processing 14.4 (2006): 1117-1127..

[3] Sudhakar, B., and R. Bensraj. "An efficient sentence-based sentiment analysis for expressive text-to-speech using fuzzy neural network." Research Journal of Applied Sciences, Engineering and Technology 8.3 (2014): 378-386.

[4] Anguera, X., Wooters, C., Peskin, B., Aguil, M. (2005, July). Robust speaker segmentation for meetings: The ICSI−SRI spring 2005 diarization system. In International Workshop on Machine Learning for Multimodal Interaction (pp. 402-414). Springer, Berlin, Heidelberg.

[5] C. Wooters, J. Fung, B. Peskin, and X. Anguera, Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system, in Rich Transcription Workshop, New Jersey, USA, 2004.

[6] J. Flanagan, J. Johnson, R. Kahn, and G. Elko,Computer-steered microphone arrays for sound transduction in large rooms,J. Acoust. Soc. Amer, vol.78, pp. 1508 1518, Nov. 1994.

[7] D. Johnson and D. Dudgeon, "Array Signal Processing," Englewood Cliffs: Prentice-Hall, 1993.

[8] Anguera, Xavier, Chuck Wooters, and Javier Hernando. "Acoustic beamforming for speaker diarization of meetings." IEEE Transactions on Audio, Speech, and Language Processing 15.7 (2007): 2011-2022.

[9] Garcia-Romero, Daniel, et al. "Speaker diarization using deep neural network embeddings." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[10] Geiger, Jrgen, Frank Wallhoff, and Gerhard Rigoll. "GMM-UBM based open-set online speaker diarization." Eleventh Annual Conference of the International Speech Communication Association. 2010.

[11] Yu, Feng, et al. "Emotion detection from speech to enrich multimedia content." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001.

[12] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44.3 (2011): 572-587.

[13] Code available at: https://github.com/thumblas/MLSPAudio/