ACOUSTIC FEATURE-BASED SENTIMENT ANALYSIS OF CALL CENTER DATA

A Thesis

Presented to

The Faculty of the Graduate School

At the University of Missouri-Columbia

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

By

ZESHAN PENG

Dr. Yi Shang, Thesis Supervisor

DECEMBER 2017

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

ACOUSTIC FEATURE-BASED SENTIMENT ANALYSIS OF CALL CENTER DATA

Presented by Zeshan Peng

A candidate for the degree of

Master of Science

And hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Yi Shang

Dr. Detelina Marinova

Dr. Dong Xu

ACKNOWLEDGEMENTS

I would like to thank my academic advisor Dr. Yi Shang for all of the helps and those great suggestions that I received from him. I will not be able to have this thesis work without his helps.

I would also like to thank all the people in our team for this project, especially Nickolas Wergeles and Wenbo Wang. They always provide me the best supports and critical thoughts in the whole process, which let me make this far for this work.

Finally, I would like to thank Dr. Detelina Marinova and her student Bitty Balducci for their great efforts to provide me datasets for this project as well as their business insights about my work that directed my initial work.

I would like to give my last thanks to Dr. Dong Xu for being my committee member and helping me defense my thesis work.

- Zeshan Peng

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

With the advancement of machine learning methods, audio sentiment analysis has become an active research area in recent years. For example, business organizations are interested in persuasion tactics from vocal cues and acoustic measures in speech. A typical approach is to find a set of acoustic features from audio data that can indicate or predict a customer's attitude, opinion, or emotion state. For audio signals, acoustic features have been widely used in many machine learning applications, such as music classification, language recognition, emotion recognition, and so on. For emotion recognition, previous work shows that pitch and speech rate features are important features. This thesis work focuses on determining sentiment from call center audio records, each containing a conversation between a sales representative and a customer. The sentiment of an audio record is considered positive if the conversation ended with an appointment being made, and is negative otherwise. In this project, a data processing and machine learning pipeline for this problem has been developed. It consists of three major steps: 1) an audio record is split into segments by speaker turns; 2) acoustic features are extracted from each segment; and 3) classification models are trained on the acoustic features to predict sentiment. Different set of features have been used and different machine learning methods, including classical machine learning algorithms and deep neural networks, have been implemented in the pipeline. In

our deep neural network method, the feature vectors of audio segments are stacked in temporal order into a feature matrix, which is fed into deep convolution neural networks as input. Experimental results based on real data shows that acoustic features, such as Mel frequency cepstral coefficients, timbre and Chroma features, are good indicators for sentiment. Temporal information in an audio record can be captured by deep convolutional neural networks for improved prediction accuracy.

1. INTRODUCTION

Speech analytics allows people to extract information from audio data. It has been used widely to gather business intelligence through the analysis of recorded calls in contact center for many business companies. Sentiment analysis is one of the speech analytics that tries to infer subject feelings about products or services in a conversation. This analysis can be also used to help agents who talk to customers over the phone build customer relationship and solve any issues that may emerge.

There are two major methods that have been used for audio sentiment analysis. One is acoustic modelling and the other is linguistic modeling. In linguistic modeling, it requires transcribing audio records into text files and conduct analyses based on text content. It assumes that some specific words or phrases are used in a higher probability in some certain environments. One can expect things to happen if some words or phrases appear frequently in the transcripts. Many researchers have showed strong evidences that linguistic modelling has a good performance on audio sentiment analysis. Among their work, lexical features, disfluencies and semantic features are mostly used when building their models. However, acquiring a good transcript for each audio record can be humanly tedious and financially costly. A good automatic transcribing system is also hard to

establish to accurately transcribe audio records into text. Even a small error in the text can result a big difference in a linguistic model. In acoustic modelling, it relies on acoustic features of audio data. These features include pitches, intensity, speech rate and so on that can be easily calculated by computer programs. They together can provide basic indicators of sentiment in some degree. Acoustic features have been used in many other business or psychological analyses as well, which is another reason that we should use them in sentiment analysis. However, the problem with acoustic modeling is that the quality of audio records has strong influence on the final result. Since we are using acoustic characteristic of audio data, the quality of audio can significantly impact the ability to get accurate values for those acoustic features, which will fail building good models in the end. The other problem is that in a real-world setting, there are always some random noises in the background. Model training data may not be able to capture these random noises, which would challenge the sentiment analysis result if we want to monitor it through a live conversation. But still, this method is worthy to explore more and research on because of its convenience and efficiency.

The remaining structure of this thesis is as follows: chapter 2 gives background and related work about sentiment analysis, chapter 3 introduces two methods that we propose and our sentiment recognition pipeline in detail, chapter 4 explains our selection of feature sets and classification models, chapter 5 shows our

experiment results, chapter 6 is the conclusion, chapter 7 talks about potential problems and future works, and the last chapter is the work references.

2. BACKGROUND AND RELATED WORK

Sentiment analysis can be generalized into three categories: text-based approach, acoustic feature-based approach and multimodal approach – a combination of previous two. Different features are used for different approach. Classification models are trained on different learning algorithms and performances are compared. Many sentiment classification techniques have been proposed in recent years [4]. However, the overall performance is not as good as we thought. Much more work could be done for this problem in the future.

2.1 Text-based approach

Text data are one of the most immense corpora that are generated daily. Researches are eager to take advantage of this and many data mining works have been done for this sake. For example, Twitter, as one of the most popular social media in the world, generates tons of text every day. Different mining methods based on Twitter text have been proposed for sentiment analysis [23] [24]. In 2010, it is shown that recurrent neural network can exploit temporal features in text contents for modeling language [13]. Even for music field, mining lyrics directly can create models for song sentiment classification problem [20] and music mood classification problem [21].

2.2 Acoustic feature-based approach

Many different acoustic features can be generated from an audio file. Pitches, intensities, speech rates, articulation rate and so on. Short-term acoustic features, such as Mel frequency cepstral coefficients(MFCCs), are proved to be useful for music modeling in [5]. MFCCs are good for musical genre classification job and evidences are shown in [6]. Authors in [9] take MFCCs as major acoustic features for emotion recognition problem. MFCCs features are also used on training single deep neural network for speaker and language recognition problem in [12]. Timbre and Chroma are other acoustic features that interest many researchers in the past years. They can be used to generate music from text in [7]. In their work, emotions are extracted from text, and then a combination of timber and Chroma features are used to composed music that expresses those emotions in the text. Chroma and timbre features can be used to classify music as well in [8]. They also claim that "Chroma features are less informative for classes such as artist, but contain information that is almost entirely independence of the spectral features." Therefore, they are equally important to be included when conducing sentiment analysis on acoustic features. Rhythm, timber and intensity of music are used for music mood classification problem as well [22]. Acoustic feature-based

approach can be extended to speech recognition in [10] and automatic language identification in [11] when modeling with deep neural network.

2.3 Multimodal approach

With the emerge of social media, except for text corpora and audio records, videos and images brings new opportunities for sentiment analysis. Facial and vocal features are extracted from these new streams and multimodal sentiment analysis seems to have much more potential in the future [14] [15]. YouTube, as one of the biggest video hosts, has a huge number of videos. Researches are trying to identify sentiments from those videos based on linguistic, audio and visual features all together [25] [26]. Fusion of modalities of audio, visual and text sentiment features as source of information is another way that is proposed in [27]. For linguistic analysis, language difference could be a problem. It is shown that language-independent audio-visual analysis is competitive with mere linguistic analysis [28]. Works in [16] [17] [18] show that systems that combine lyric features and audio features significantly outperform other systems that use lyrics or audios singularly for music mood classification problem. Fusion of lyrics features and melody features of music improves performance of music information retrieval system [19]. Fusion of multiple modalities is current trend since we can always find new information when we view an existing object from a new angle. Sometimes,

the access to visual and linguistic sources could be hard and costly. For example, in our data source, it is hard to get facial expressions from a customer over the phone. But in other circumstances when we have all the information, multimodal approach would be the first choice.

3. PROPOSED METHODS

The problem that this thesis work wants to solve is to identify whether a meeting has been scheduled between a customer and a representative in a phone call conversation. There are two methods that we propose to address this kind of problem. The first method is to split each audio record into segments by speaker turns before extracting feature vectors from each segment, and then use classic machine learning algorithms to classify those feature vectors according to the sentiments. All segments that come from the same audio record share the same sentiment label as the audio record label. The second method is to cut each audio record into short audio clips with same length. Then features vectors are extracted from each audio clips. After that, we stack up those feature vectors into feature matrixes in time order and then feed feature matrixes into deep convolutional neural networks to build classification models. All feature matrixes that come from the same audio record share the same sentiment label as the audio record label. The next two subsector describes these two methods in more detail.

Figure 1. Problem Illustration

3.1 Acoustic Feature-Based Sentiment Recognition Using Classic Machine Learning Algorithms

The whole sentiment recognition process consists of five steps. Firstly, we gather useful audio data from call center records. We select audio data that has label with it since we are more interested in using supervised machine learning algorithms. It is highly possible that unsupervised machine learning can provide interesting results for audio signal sentiment analysis, which is left to future exploring either by combining it with supervised machine learning or merely itself.

The second step is cleaning and pre-processing the chosen audio files. Any unrelated parts in audio data are removed, such as ringtones, receptionist talking, music, advertisement, etc. Several methods have been tried to remove those unrelated parts, however, none of them met our expectations. So, we decide to remove them manually since the dataset we currently have is pretty small. After that we apply speaker diarization algorithm to split each audio record into segments by speaker turns. The third step is feature extraction and selection. We consider different combination of feature sets and select different features to train on in the next stage. The fourth step is classification model training. We use different training algorithms, techniques and models, then compare their performance in terms of prediction accuracies. The last part is testing the trained classification model.

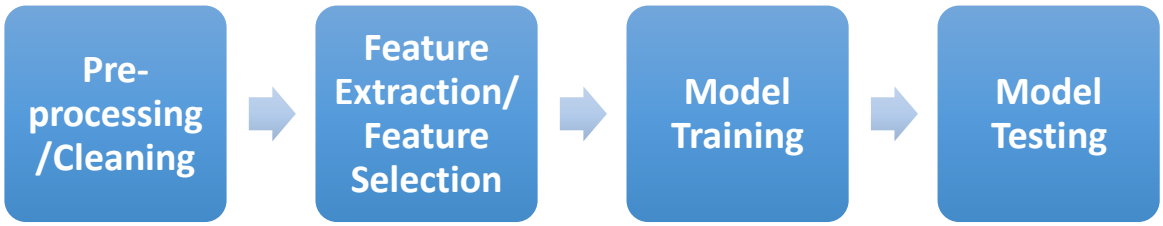

Figure 2. Sentiment Recognition Process Overview

3.1.1 Gathering Audio Data

Every training data impacts the final classification model. We do not want bad data to affect model's performance, and we want to have effective training datasets that can generalize well to all possible scenarios but not many unlikely events. As for audio datasets, the quality of the audio file could be one metric to be used when selecting audio files. Call center audio records are not always recorded good. Due to the internet latency and background noises, some files are very blurred and even hard for human to understand the conversation in it. We need audio data that are recorded with good quality. There are also some other files that do not have a label with it, and we want to get rid of those files as well for supervised learning purpose.

3.1.2 Data Cleaning and Pre-processing

In an audio file, it is usually a conversation between two speakers: a customer and a salesperson. In some scenarios, there is a receptionist answering the phone call and transferring the salesperson to whoever he or she wants to reach to. The features of receptionist are not our mainly interest here so we need to cut those conversations between salesperson and receptionist if it happens in the audio file. After that, we would like to split each audio file into segments by speaker turns.

Knowing changes of certain features between consecutive segments would help the whole sentiment analysis. That is why it is important to split audio files into segments by speaker turns before gathering features from them. We also want to study the temporal relations of features in consecutive segments in an audio file. That would help monitor live conversations in call centers. There are plenty of speaker diarization algorithms that are publicly available. As referred in [1], [2], and [3], a lot of researches has been conducted for this type of problem, and many of them have a good performance to split audio file by speaker turns automatically.

In our research, we have used IBM Watson and Google cloud computing services to split our audio files into segments by speaker turns. Both of these systems have the ability to identify who speaks at what time and transcribe the conversation into text. Even though the transcription part is not as accurate as we though, the timestamps of start and end time for each speaker turn are good enough for us to use to separate each audio file into segments.

Since we only have a small portion of audio data for now, and the transcripts of all audio files are available for us, we split each audio file into segments based on timestamps in the transcript accordingly. However, the speaker diarization procedure is still crucial when have more and more data coming in in the future and the access to all of the transcripts would be very costly. That is the time we may want to use automatic speaker diarization techniques.

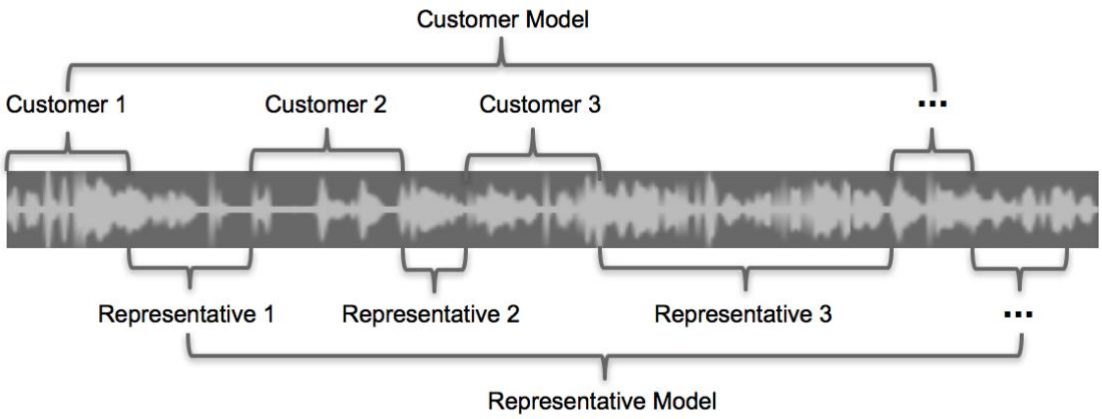

Figure 3. Speaker Diarization Process

Pseudo-code:

```
function speaker_diarization ():
    open audio file
    open transcription file
    for each timestamp pair (start_time, stop_time) in transcription file:
        segment = cut_audio_file (start_time, stop_time)
        segment_label = audio_file_label
        if segment belongs to customer:
            Add segment to customer_cluster
        If segment belongs to representative:
            Add segment to representative_cluster
```

3.1.3 Feature Extraction and Selection

After we have split each audio file into segments by speaker turns, we extract features from each segment. There are many features that can be extracted from an audio file. For our research purpose, we focus on acoustic features. There are mainly two group of features that interest us. One group is prosody features, such as pitch, intensity, speech rate, number of pauses, etc. The other group is short-term features. We will explain our features sets in detail in chapter 4.

Pseudo-code:

```
function feature_extraction ():
    for each segment in customer_cluster:
        prosody, short_term = calc_algorithm(segment)
        Add prosody to prosody_feature_cluster for customer
        Add shor_term to short_term_cluster for customer
    for each segment in representative_cluster:
        prosody, short_term = calc_algorithm(segment)
        Add prosody to prosody_feature_cluster for representative
        Add shor_term to short_term_cluster for representative
```

3.1.4 Train Classification Model

Once we have features for each segment, we split data into training set, validation set and testing set, then we train a classification model based on these features and their associated labels. Every segment in the same audio file has the same label as that audio file. We also partition segments into two groups: customer group and salesperson group. Since customers may have different behaviors than salesperson, we would like to train two different models to represent each of them. We take all features in customer group to be our training dataset for the customer model, and all features in salesperson group to be our training date for the salesperson model. Validation set is used to set up hyper parameters of learning algorithms, and testing set is used to measure the performance of the trained model. More details of classification models and learning algorithms are in chapter 4.

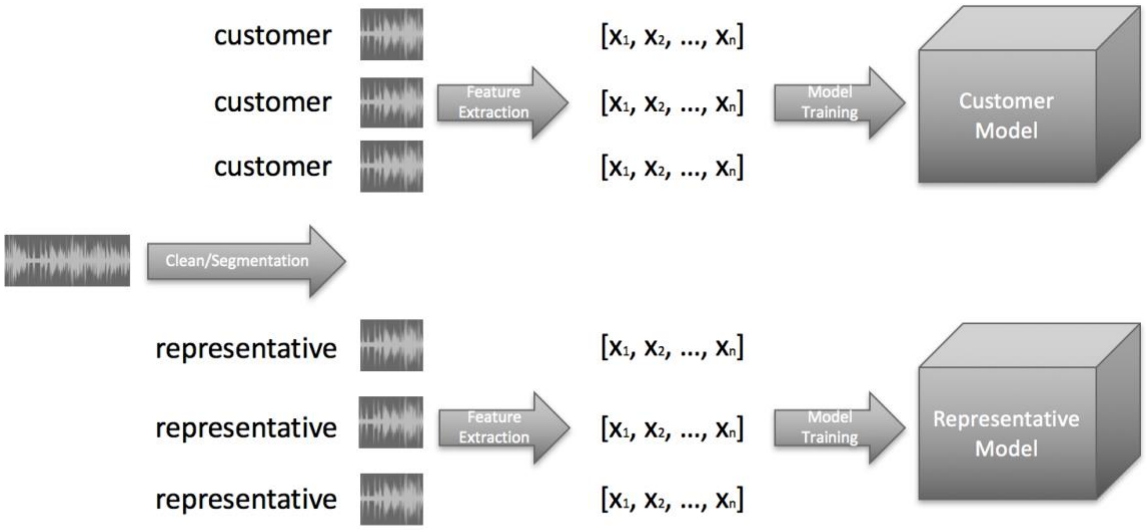

Figure 4. Data Flow Diagram for Model Training

3.1.5 Test Classification Model

After the model is built in the last procedure, we use it to predict labels for any audio file that we want to do sentiment analysis on. The audio file is still split into segments by timestamps in the transcript. The model will predict label for each segment in that file, and we take the majority vote of the labels as the final label for that test audio file.

$$y = \sum_i^n y_i / n \geq 0.5 \,?\, 1 : 0$$

where n is the number of segments in an audio record

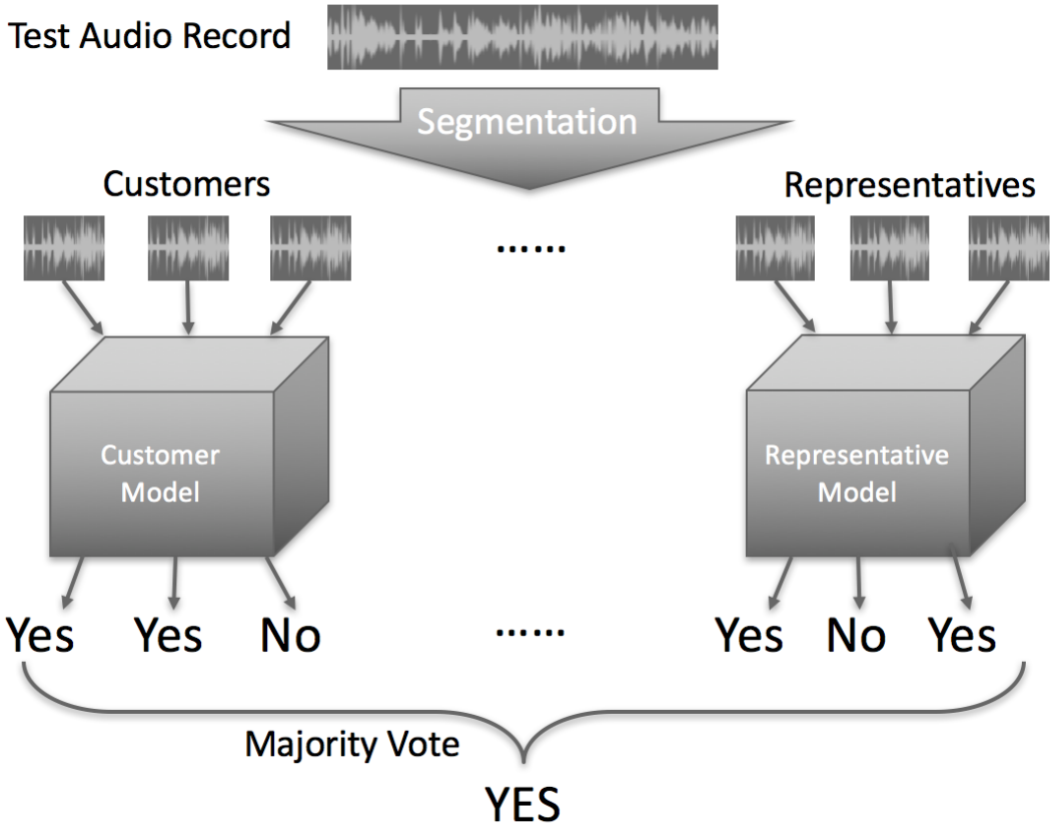

Figure 5. Majority Vote Method for Model Testing

3.2 Acoustic Feature-Matrix-Based Sentiment Recognition Using Deep Convolutional Neural Network

3.2.1 Deep Learning

It has been shown in [29] that deep convolutional neural network can be applied to character-level feature matrix for text classification problem. It has

achieved competitive results with current state-of-art methods. Evidence shows that temporal information can be captured by applying deep convolutional neural network on feature matrix of time series dataset. With the advancement of machine learning, especially deep learning methods, audio signal analysis can take advantage of it to build competitive models or even better models than other approaches.

3.2.2 Feature Matrix

Convolutional neural networks usually take two-dimensional matrix as input. We stack up audio feature vector into feature matrix to meet that. Short-term features are extracted every 50 milliseconds from each audio record and then stacked up in time order to build feature matrix for that audio record. The window size is 300, which means for every 300 rows we cut the matrix. Every matrix generated from same audio record shares the same sentiment as the audio record. If the number of rows or the remaining rows after cutting is less than 300, we use zero-valued patch to compensate. Each training matrix has the size of 300 by 34 since we have 34 short-term features in total. The figure below illustrates the process for building feature matrixes.

Pseudo-code:

```
function feature_matricization ():
    for each audio record:
        cut it into equal length (50ms) segments
        for each segment:
            short-term = calc_algorithm(segment)
        stack up short-term features into feature matrix in time order
        cut feature matrix every 300 rows
        matrix_label = audio_record_label
        add to training dataset
```

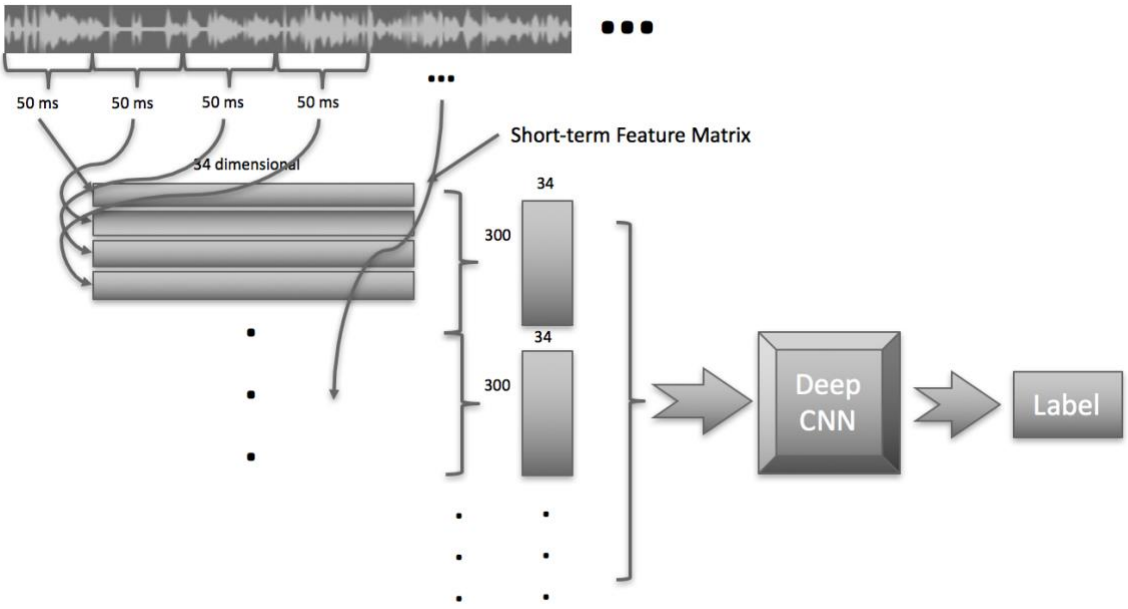

Figure 6. Feature Matrix Illustration

3.2.3 Architecture

Massive experiments have been conducted to compare performances when using different network settings. Those settings include number of convolutional layer, kernel size, max pooling kernel size, and window size when cutting matrix. The figure below is a typical network architecture of four convolutional layers. This architecture design is based on pre-trained neural networks that work extremely well on image datasets. Every convolutional layer is followed by a max pooling layer. We use 1-D convolutional kernel here since features are distinct from each other, and our experiment results show that 1-D convolutional kernel has better performance than 2-D convolutional kernel.

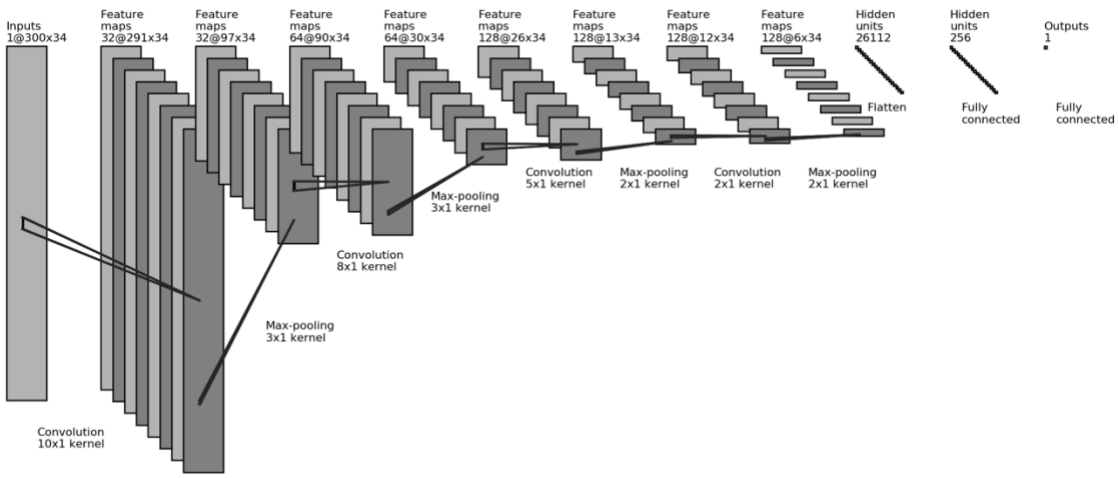

Figure 7. Deep Convolutional Neural Network Architecture

4. FEATURE SETS AND CLASSIFICAITON ALGORITHSM

4.1 Feature Sets

4.1.1 Fundamental Features

The first feature set that we consider are fundamental features for any audio data, such as decibel, mean of pitches, standard deviation of pitches, covariance of pitches, speech rate, number of pauses in a segment, length of the segment in milliseconds. These are general features that can represent an audio segment in some extent. These features have shown crucial importance on audio signal analysis from literatures a decade ago.

4.1.2 Shimmer and Jitter

Shimmer measures the variation in frequency of vocal fold movements while jitter measures the variation in amplitude of vocal fold movements. These two features can represent characteristics of an audio segment in another perspective.

4.1.3 Short-term Features

Mel-frequency cepstrum coefficients, Chroma and timbre features are short-term features used in this work. The mel-frequency cepstrum is a representation of the short-term power spectrum of a sound. Mel-frequency cepstral coefficients(MFCCs) are coefficients that are used to make up the spectrum.

Different combination of those coefficients has different indications for some sentiment analysis, so we want to include them in our feature set. Chroma features represent spectral energy of different pitch classes. Timbre features show the character or quality of a sound that are different from pitch feature and intensity feature. These three groups of features have been frequently used in many classification problems as explained in related work in section 2.

Feature	Description
MFCCs	Mel Frequency Cepstral Coefficients from a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale
Chroma	A representation of the spectral energy where the bins represent equal-tempered pitch classes
Timbre	The character or quality of a sound or voice as distinct from its pitch and intensity

Table 1. Short-term Feature Descriptions

4.1.4 Emotion Features

We also find that emotion features can be critical indicators for some sentiments. Positive emotions, like happy and satisfaction, can lead to a good result in the end of the conversation while negative emotions, like upset and angry, can turn things around in the end. We want to train our model on these features too.

4.1.5 Spectrograms

Besides of valued inputs, we also generate image features from each segment. They are spectrograms. Spectrogram is a visual representation of the spectrum of frequencies of sound varied with time. Some sentiments may have specific patterns in the spectrograms. We want our model to be able to identify those patterns in the spectrograms if they exist.

4.2 Classification Algorithms

In our experiment, we use both legend machine learning algorithms and neural networks. For legend algorithms, we have used support vector machine with cubic kernel and k-nearest neighbor classifier. For neural networks, we have used classic shallow feed forward neural networks and deep convolutional neural network.

Different learning algorithm has different advantages and characteristics. We want to experiment on all of them and compare their results.

5. EXPERIMENT RESULTS

5.1 Dataset

Our datasets consist of 88 audio recording files from Penske cooperated offices. The duration of these files ranges from a few minutes to ten minutes. The average length is about three minutes. After we split these audio files into segments by speaker turns, we have 2859 segments in total. Among them, 1438 are for customers and 1421 are for salespersons. Features extracted from customer group are used to train customer model, so as salesperson group is for salesperson model.

	Positive	Negative	Total
Audio Records	31	55	86
Segments	1284	1304	2588

Table 2. Dataset Summary

5.2 Feature Extraction Libraries

Praat is the main software that we used to extract fundamental audio features as well as shimmer and jitter values. We use a python library named

pyAudioAnalysis on GitHub to extract MFCC features. We use OpenSmile package to extract emotion features. Another python library pyplot in matplotlib package is used to generate spectrograms for each audio segment.

5.3 Training Tools

We use machine learning package in Matlab to train models using classic machine learning algorithms. For neural networks, they are coded in Keras platform which is based on tensorflow.

5.4 Per-Segment Results

Since customer behaviors different than representative, we want to two separate models to represent each of them. Speaker diarization process helps on that. Features extracted from customer segments are grouped into one cluster while other features extracted from representative segments are grouped into another cluster. Features in customer cluster are used to train customer model and so as the representative model. Models' performances are compared by using different feature set for training. Below is the figure that summarizes the comparison based on prediction accuracy for segments:

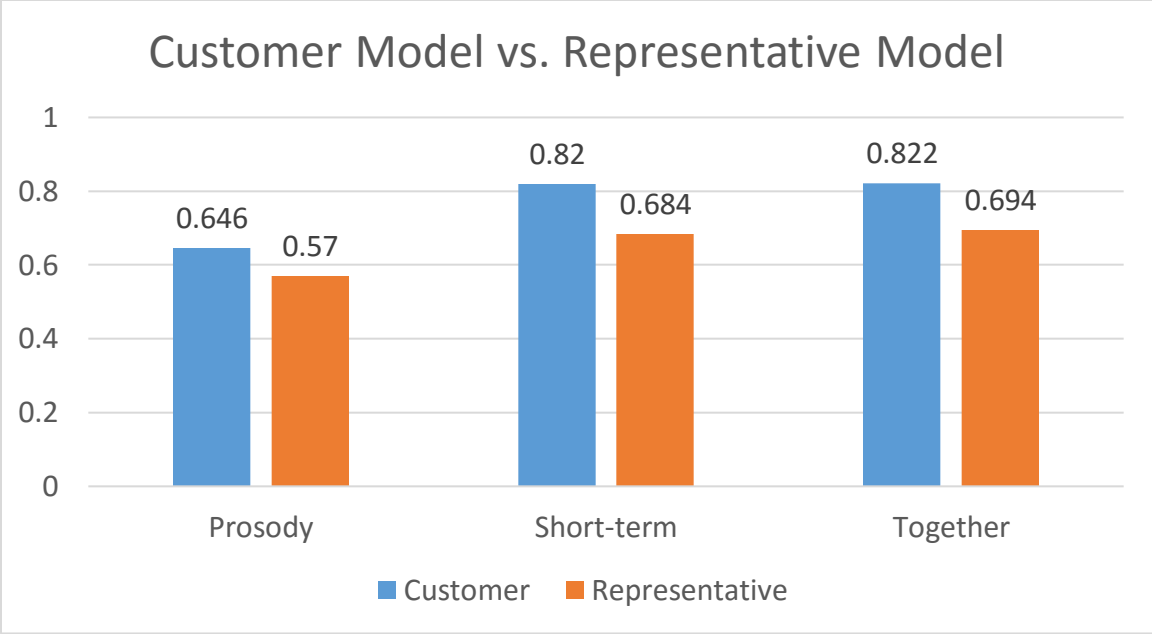

Figure 8. Comparison of Customer Model and Representative Model on Different Feature Sets

From above figure, we can easily conclude that short-term features are superior indicators than prosody features. When both of them are used for training, accuracy is slightly improved. That does not mean combination of them has better indication strength. More experiments should be conducted for a larger dataset before we can say anything about it.

5.5 Per-Record Results

Segment-wise prediction is not our final goal here. We want to predict whether representative is persuasive or not in a whole audio record. To accomplish this, we predict each segment of a test audio record using models trained in 5.2.1

and take the majority vote of all the predictions as our final prediction of that audio record. We run our experiments on based on using different model to predict. Here we have three model settings: 1) use customer model only to predict customer segments and take the majority vote as the final predication for that audio record, 2) use representative model only to predict representative segments and take the majority vote as the final prediction for that audio record, 3) use both customer and representative model to predict all segments in an audio record and take the majority vote as the final prediction for that audio record. Below is the comparison result.

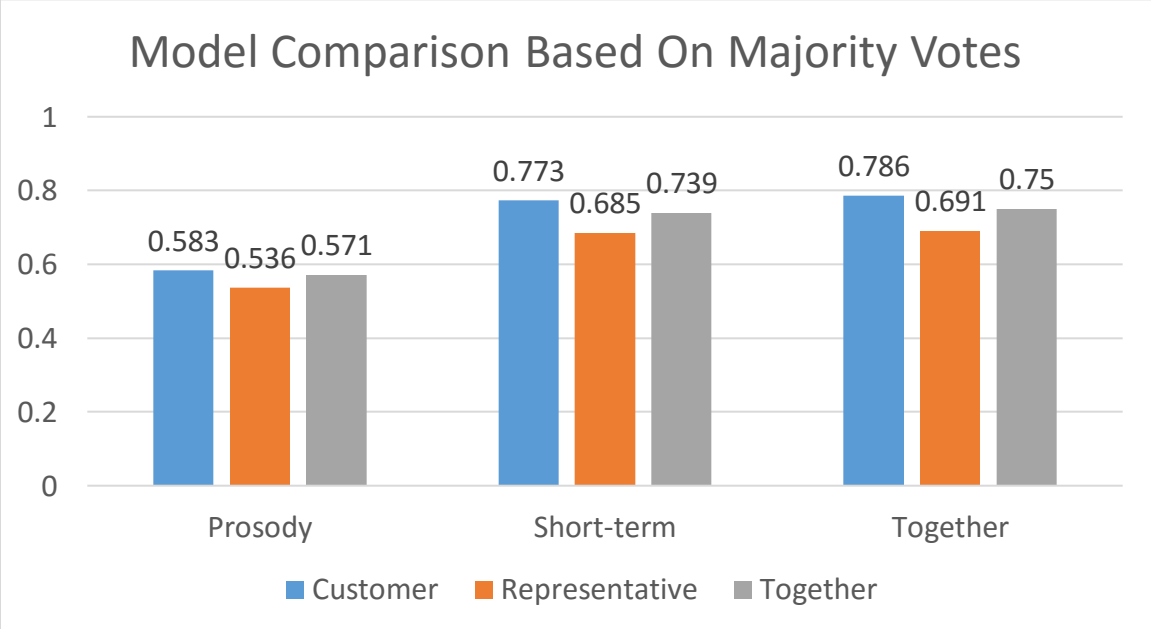

Figure 9. Model Comparison Based on Majority Votes Per Audio Record

Above figure shows that using customer model only has the best performance compared with other models. It matches intuitive instinct that representatives have similar behaviors independent of customers behaviors. Representatives are more consistent no matter what customers respond while customers have different characteristics when they are sentimentally positive or negative. Using both of them has compromised performance.

5.6 Deep Learning Results

The figure below is the result comparing different number of convolutional layers when using either feature vectors or feature matrixes as input to the neural network.

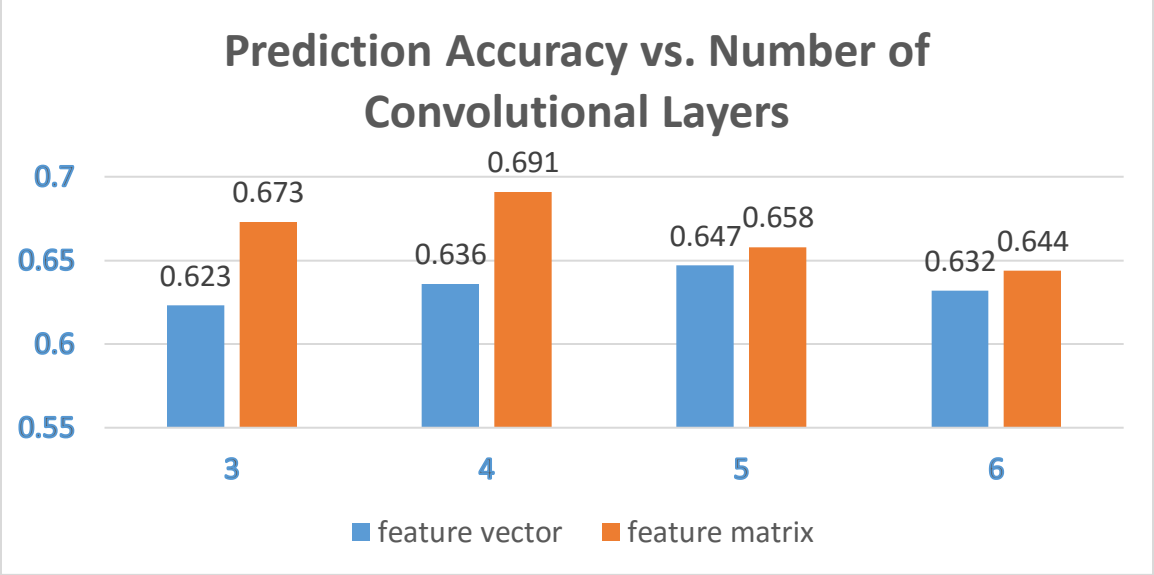

Figure 10. Prediction Accuracy vs. Number of Convolutional Layers

It can be seen that four-convolutional layer has the best performance among the others, and using feature matrixes has better performance than using feature vectors, which infers that temporal information between audio clips can be captured by using deep convolutional neural networks.

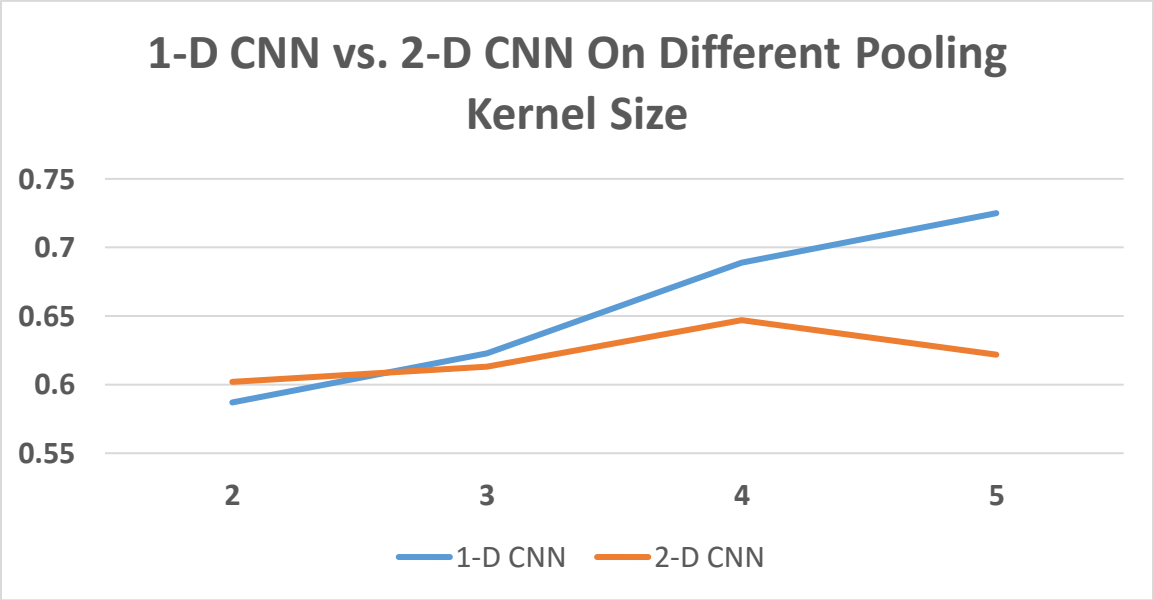

Figure 11. 1-D CNN vs. 2-D CNN on Different Pooling Kernel Size

Above figure shows the comparison result of using 1-D convolutional versus using 2-D convolutional on different pooling kernel size. From the result, we find that 1-D convolutional neural network performs better than 2-D convolutional neural network. It might be because there are no temporal meanings between each

individual short-term feature since the column in the matrix represents different short-term features. It also shows that predication accuracy can be improved by applying a larger max pooling kernel size.

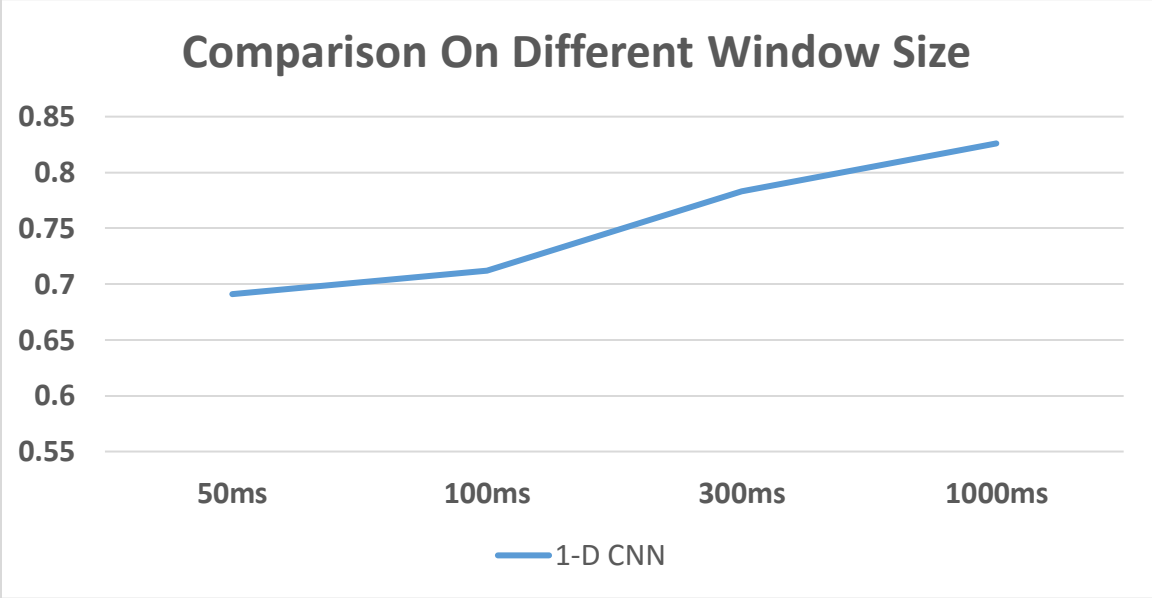

Figure 12. Comparison on Different Window Size

This figure shows the comparison result on using different window size when extracting features from audio records. It seems that a larger window size will have a better performance. However, larger window size will result smaller training dataset for neural network. Whether it is true or not needs to be proved by conducting more experiments using a larger dataset.

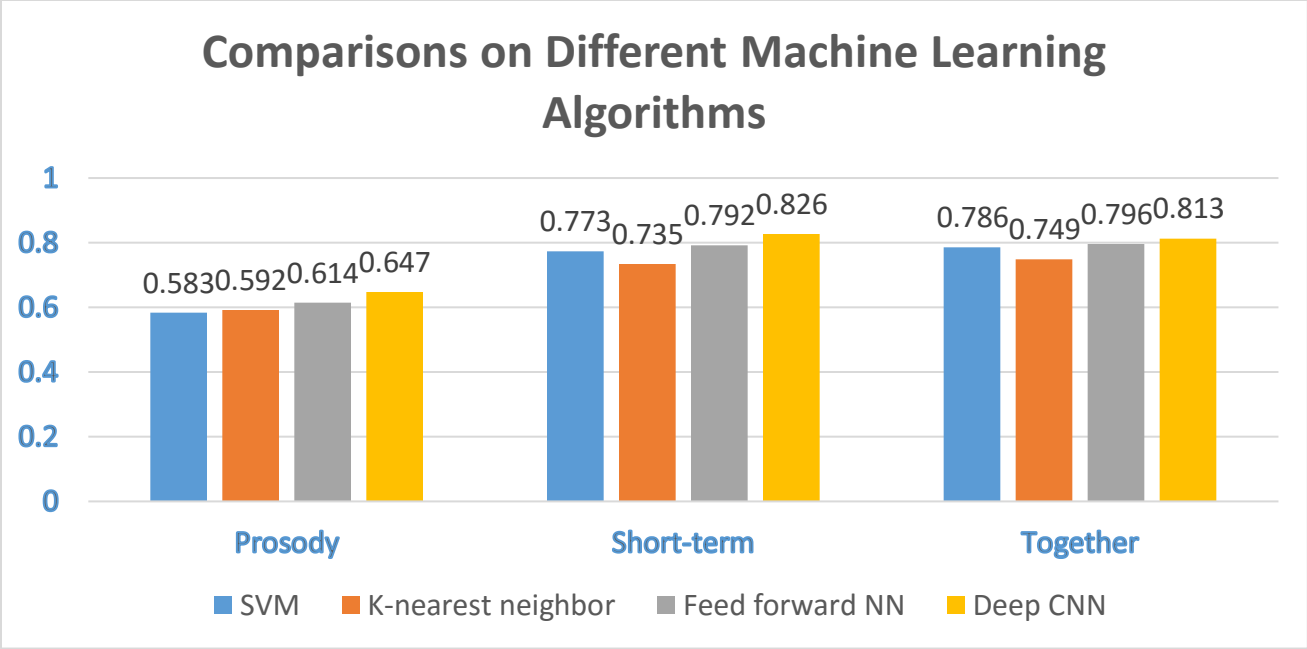

Figure 13. Comparison on Different Machine Learning Algorithms

This figure compares model performance using different machine learning algorithms. It can be seen that deep convolutional neural network has a slightly better performance over other methods listed here. Short-term features have a stronger indication effect of the sentiment than prosody features.

6. Conclusion

Audio sentiment analysis is a method of inferring opinions of customers in conversations with customer support representatives. Having a mature system to analyzed sentiments would be very beneficial for business world. This thesis work presents two methods to address this problem. One is based on segment-wise feature vector classification. Every audio recording is split into segments by speaker turns before analyzing. Different sets of features are extracted from each segment and then a classification model is trained based on these features. Another is to feed feature matrix into deep convolutional neural network and let it learn the temporal information between consecutive audio clips. Each feature matrix is constructed by stacking up feature vectors which are extracted from equal length audio clips of audio records. We compare the performance of models using different machine learning algorithms. We show that there are strong connections between the sentiment we analyze and the features we extract. Short-term features, such as MFCCs, Chrome and timbre features have better indications than prosody features. Deep convolutional neural network can not only be applied to image dataset, but also be applied to non-image dataset, like feature matrixes. Temporal information between consecutive audio clips within a feature matrix can be captured by using deep convolutional neural network from our experiment results. We also present the problems that we have encountered

during our research process. Further research work can be conducted to address these problems to improve the overall system performance.

7. FUTURE WORK

Automatic speaker diarization process can be embedded into this audio sentiment analysis process since we want to split each audio data into segments by speaker turns. It is even important when transcripts are costly and time-consuming to get. Accurate text transcription could be another direction since much work has been done in regard of text-based audio sentiment analysis. If we can get a good text transcription of an audio file automatically, then we can combine those techniques in text-based analysis with our acoustic feature based techniques to improve the performance of analysis system. Feature matrix is not the end. There could exist a better representation of features to feed into deep neural networks. Moreover, we can find other acoustic features and better indicators to characterize audio data. More experiments should be conducted for a larger dataset to build more reliable models.

8. REFERENCES

[1]. Lapidot, Itshak & Aminov, L & Furmanov, T & Moyal, Ami. (2014). Speaker Diarization in Commercial Calls.

[2]. A. X. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech, and Language*, vol.20, no.2, pp.356-370, Feb. 2012

[3]. Cyrta, Pawel, Tomasz Trzcinski and Wojciech Stokowiec. "Speaker Diarization using Deep Recurrent Convolutional Neural Networks for Speaker Embeddings." *ISAT* (2017).

[4]. W. Medhat, A. Hassana, and H. Korashy. Sentiment analysis algorithms, applications: A survey. *Ain Shams Engineering journal*, pages 1093–1113, 2014.

[5]. B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proc. International Symposium on Music Information Retrieval*, 2000.

[6]. G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[7]. H. Davis and S. M. Mohammad. Generating music from literature. In *Proc. 3rd Workshop on Computational Linguistics for Literature*, pages 1–10, 2014.

[8]. D. P. W. Ellis. Classifying music audio with timbral and chroma features. In *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, pages 339–340, 2007.

[9]. D. Ververidis, C. Kotropoulos, and I. Pitas. Automatic emotional speech classification. In *Proc. ICASSP*, volume 1, pages 593–596, 2004.

[10]. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for

acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[11]. I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno. Automatic language identification using deep neural networks. In *Proc. ICASSP*, pages 5337–5341, 2014.

[12]. F. Richardson, D. Reynolds, and N. Dehak. A unified deep neural network for speaker, language recognition. In *Proc. INTERSPEECH*, pages 1146–1150, 2015.

[13]. T. Mikolov, M. Karafiat, L. Burget, J. H. Cernocky, and S. Khudanpur. Recurrent neural network based language model. In *Proc. INTERSPEECH*, pages 1045–1048, 2010.

[14]. S. J. Fulse, R. Sugandhi, and A. Mahajan. A survey on multimodal sentiment analysis. *International Journal of Engineering Research,Technology (IJERT)* ISSN: 2278-0181, 3(4):1233–1238, Nov 2014.

[15]. M. Sikandar. A survey for multimodal sentiment analysis methods. *International Journal of Computer Technology, Applications (IJCTA)* ISSN:2229-6093, 5(4):1470–1476, July 2014.

[16]. X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proc. Joint Conference on Digital Libraries, (JCDL),* pages 159–168, 2010.

[17]. J. Zhong, Y. Cheng, S. Yang, and L. Wen. Music sentiment classification integrating audio with lyrics. *Information and Computational Science*, 9(1):35–54, 2012.

[18]. A. Jamdar, J. Abraham, K. Khanna, and R. Dubey. Emotion analysis of songs based on lyrical and audio features. International Journal of Artificial *Intelligence and Applications(IJAIA),* 6(3):35–50, 2015.

[19]. T. Wang, D. Kim, K. Hong, and J. Youn. Music information retrieval system using lyrics and melody information. In *Proc. Asia-Pacific Conference on Information Processing*, pages 601–604, 2009.

[20]. Y. Xia, L. Wang, K.-F. Wong, and M. Xu. Sentiment vector space model for lyric-based song sentiment classification. In *Proc. ACL-08:HLT, Short Papers*, pages 133–136, 2008

[21]. X. Hu, J. S. Downie, and A. F. Ehmann. Lyric text mining in music mood classification. In *Proc. 10th International Conference on Music Information Retrieval (ISMIR)*, pages 411–416, 2009.

[22]. B. G. Patra, D. Das, and S. Bandyopadhyay. Unsupervised approach to Hindi music mood classification. In *Proc. Mining Intelligence and Knowledge Exploration (MIKE)*, pages 62–69, 2013.

[23]. A. Kumar and T. M. Sebastian. Sentiment analysis on twitter. *International Journal of Computer Science (IJCSI)*, 9(4):372–378, July 2012.

[24]. P. Gamallo and M. Garcia. Citius: A Naive-Bayes strategy for sentiment analysis on English tweets. In *Proc. 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 171–175, August 2014.

[25]. R. Mihalcea. Multimodal sentiment analysis. In *Proc. 3rd Workshop in Computation Approaches to Subjectivity and Sentiment Analysis*, pages 1- 1 July 2012.

[26]. V. Rosas, R. Mihalcea, L.-P. Morency, "Multimodal sentiment analysis of spanish online videos", *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 0038-45, 2013.

[27]. Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, Amir Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, In *Neurocomputing*, Volume 174, Part A, 2016, Pages 50-59, ISSN 0925-2312, January 2016.

[28]. M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L. P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context", *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46-53, 2013.

[29]. X. Zhang, J. Zhao, Y. LeCun. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.