Speech Sentiment Analysis

Vaijyant Tomar
College of Computing and Informatics
University of North Carolina at Charlotte
Charlotte, US
vtomar@uncc.edu

***Abstract*—Humans have developed the ability to communicate with each other. Most of the communication involves speaking. As the world moves towards an AI first world, it becomes natural that the humans would want to communicate with the machines. There has been a considerable advancement in speech recognition technology, however, speech sentiment recognition is still in its nascent state. This study demonstrates the model for Speech Sentiment Analysis. There are two ways a speech can be analyzed. First, analyzing the text transcript of the speech. This would strip off the acoustic characteristics and in turn the underlying sentiment associated with speech. Second, identifying and analyzing the acoustic characters hidden in the speech. This would preserve the acoustic characteristics. These acoustic characteristics can be then leveraged to identify the sentiment associated with human speech. In this article, the latter method is explored.**

***Index Terms*—artificial intelligence, emotion analysis, emotion classification, sentiment analysis, speech, speech processing**

I. Introduction

Humans have the ability to modulate their vocal sounds using the larynx, also known as the voice box. The larynx houses the vocal cords which allows humans to manipulate pitch and volume which is crucial in phonation. This allows humans to talk, sing, laugh and express themselves.

During a speech, the vibrations in vocal cords allow modulating the air flow out of the lungs to generate complex phonations. The characteristics of human voice such as the pitch, timbre, loudness, and tone make human voice a versatile to communicate. It can be observed that humans can also express their emotions by varying the stated characteristics. This allows for identifying human emotion by analyzing speech.

Pitch is the characteristic of sound that allows us to assign musical tones to the sound wave. It can be quantified as a frequency. *Loudness* is the characteristic of sound that allows us to assign how quite or loud a sound is. It can be quantified as the amplitude of a sound wave. *Timber* is the characteristic of the sound that allows us to distinguish between various sound sources. Timber of a sound also gives us insight into how complex the sound wave is.

With different emotions and moods, not only does the tonal quality vary, but the associated speech patterns change too. For instance, people may tend to talk in loud voices when angry and use shrill or high-pitched voices when in a scared or panicked emotional state. Some people tend to ramble when they get excited or nervous. On the contrary, when in a pensive emotional state, people tend to speak slowly and make longer pauses, thereby indicating an increase in time spacing between consecutive words of their speech [9].

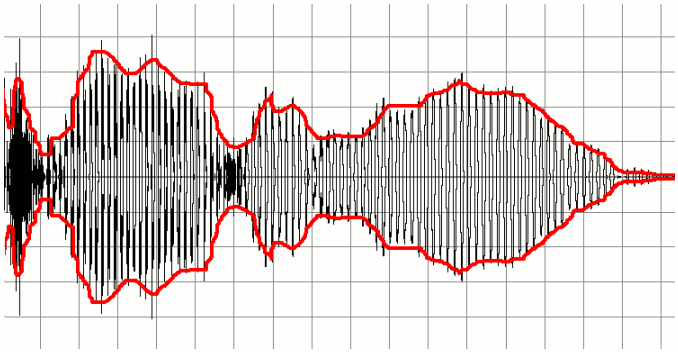

Fig. 1. A sound waves envelope marked in red [10].

As the world progresses towards an AI driven world, it becomes natural that humans would want to communicate in the natural human language with artificial agents. For an effective communication to occur, it is crucial that these agents understand human emotion. The current research and technology can easily understand the context of a conversation, however, understanding human emotion is still not stressed in the mainstream application of artificial agents or virtual agents. Having an artificial agent understand raw human emotion will contribute to enhancing current state of virtual agents.

Apart from making an artificial agent understand human emotion, speech sentiment analysis can also be employed in making humans more aware of the emotion of the person talking to them. For example, customers service centers can gather insights on their customer satisfaction by simply analyzing the speech of their customers. Also, the scores received as a part of this analysis can be used to assess the overall opinion of a company/product/services.

Sound characteristics of speech can be used in scenarios where face to face communication is not feasible or where there are a language constraint and proper model for lexicon-based speech analysis not readily available. Following are such scenarios where speech characteristics can serve as a tool for identifying human emotion:

- Playing music and changing the ambient room's lighting as per the tone of the conversation.
- Implementation in social science research

II. Related Work

There has been extensive research in speech sentiment analysis. Researchers have explored classification methods including the Neural Network (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Maximum Likelihood Bayes classifier (MLC), KernelRegression and K-nearest Neighbors (KNN), Support Vector Machine (SVM) [11].

There are two ways a speech can be analyzed. First, analyzing the text transcript of the speech. This would strip off the acoustic characteristics and in turn the underlying sentiment associated with speech. Second, identifying and analyzing the acoustic characters hidden in the speech. This would preserve the acoustic characteristics.

There has been extensive research in extracting sentiment from transcribed speech however, there has been little work on speech sentiment analysis merely from acoustic attributes of the sound. It is observed that the characteristic of the sound changes with a change in emotional state of the speaker. The speaker may say a very positive sentence but may feel otherwise of what he is speaking. The algorithm was developed to account for three emotions namely, normal, angry, panicked.

There has been an attempt to formulate an algorithm to discern the emotion associated with human speech. The analysis was carried out in MATLAB and Wavepad. The proposed algorithm investigates four vocal parameters viz., pitch, sound pressure level (SPL), timbre (ascend and descend time), and time gaps between consecutive words of speech. The authors were able to discern the emotion by retrieving the quantitative value associated with each vocal parameter [9].

There has also been formation of certain library module in python which can process audio streams. One such library in *pyAudioAnalysis*. pyAudioAnalysis provides packages for feature extraction and classification (including implementation of Support Vector Machines and kNN classifier). It also has provision for regression, segmentation and visualization. What makes pyAudioAnalysis different from other audio libraries is that it has provision to extract general feature which is linked to machine learning components. It also has provision for baseline techniques which are implemented to for audio analysis task [8].

Sentiment analysis involving speech can possess many challenges [1]. These are as follows:

- Presence of background noise,
- foreign accents,
- generation of speech in real time,
- presence of a diverse range of topics.

III. Approach

The proposed approach for speech sentiment analysis is as follows:

1) The Ryerson Audio-Visual Database of Emotional Speech and Song [2] (RAVDESS) data set.
2) Extract the acoustic features over the full sample. To extract the features, pyAudioAnalysis is used. [8].

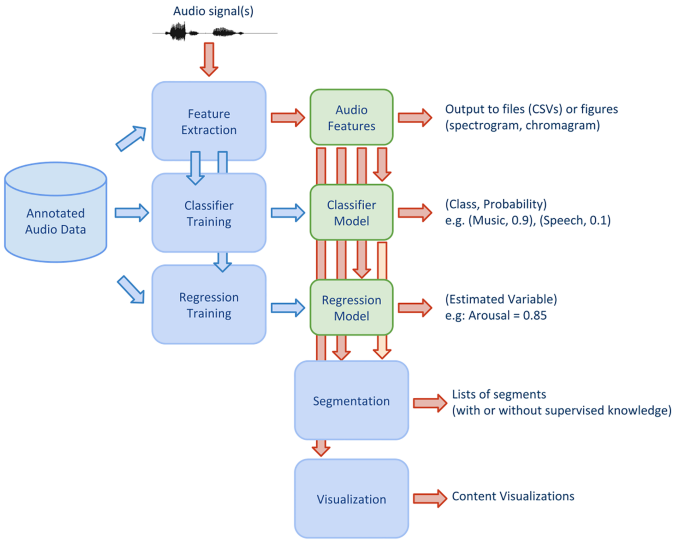

Fig. 2. pyAudioAnalysis Library General Diagram [8].

3) After feature extraction this data will be feed into training models Support Vector Machines (SVM), Random Forest and K-Nearest Neighbor.
4) To further analyze the dataset, the dataset will be spillited into male speech only samples and female speech only samples.
5) A subset of the dataset is also taken which consists of only four emotions viz., angry, happy, sad and neutral.

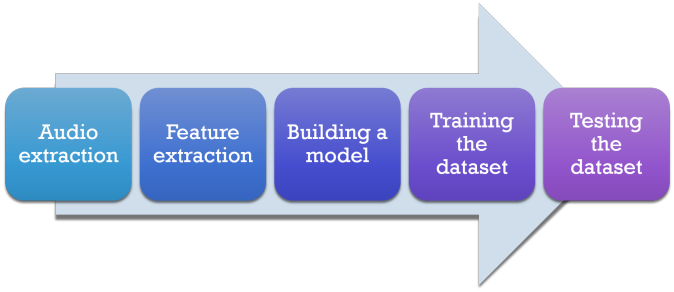

Fig. 3. Steps for analysis.

IV. Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is used as the database. It is a database comprising of voice sample of 24 actors (12 male, 12 female). The speech of these are in various emotions and all the actors speak in North American English accent. The dataset also contains 7,356 high quality video recordings which corresponds to the audio dataset. The data set voice samples for eight emotional expressions viz., neutral, calm, happy, sad, angry, fearful, surprise, and disgust. The data set for song samples for six emotional expressions viz., neutral, calm, happy, sad, angry, and fearful. All emotion expressions, except neutral, are expressed at two levels of emotional intensity normal and

strong. The database has been validated in by 297 participants [2].

For our analysis, only the audio (.wav) files. The .wav file format is used in this study because:

- Simple Format, as a result these files are easy to process.
- Lossless format, hence allows for preserving the quality of the file.
- High recoding rates, this allows for having huge dynamic ranges.

V. Feature Extraction

pyAudioAnalysis is the tool which is used to extract features from the speech. pyAudioAnalysis this my extracting these features in two stages. These are as follows:

- Short-term feature extraction: pyAudioAnalysis implements this in stFetureExtraction() function of the audioFeatureExtraction. py file. It splits the audio file into short term windows (frames) and then computes features for every frame. This result in formation of short term feature vector for the entire audio signal.
- Mid-term feature extraction: pyAudioAnalysis implements this in mtFeatureExtraction() function of the audioFeatureExtraction. py file. It computes the mean and standard deviation over each short-term feature vector of the audio signal [8].

Table I shows all the featured derived using pyAudioAnalysis python package.

VI. Patterns

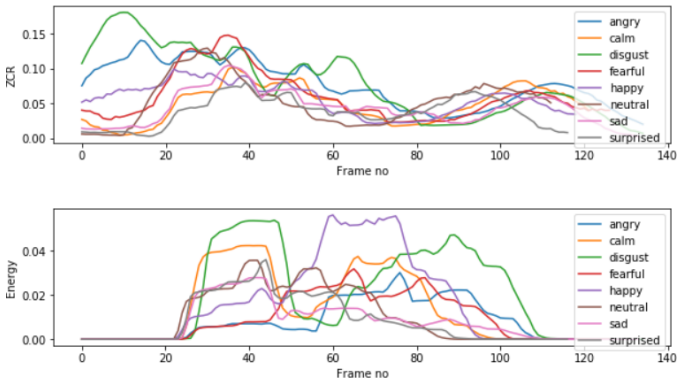

Fig. 4. ZCR and Energy against Frame No for different emoions.

Fig. 7 is the representation of features over time (frame number). These graphs are obtained for actor 1 (male) for the sentence kids are talking by the door. The graph is an overlay graph color-coded by emotions. Clear distinction between various emotions. There is almost a similar trend in ZCR for this speech sample, but it a clear difference for happy emotion in the energy graph can be seen. The graph for happy emotion rises and maintains a plateau phase for some time.

Similar patterns were observed for other features. For example, Spectral Flux for angry emotion had initial spikes which

TABLE I
List of the implemented features [8]

Feature ID	Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

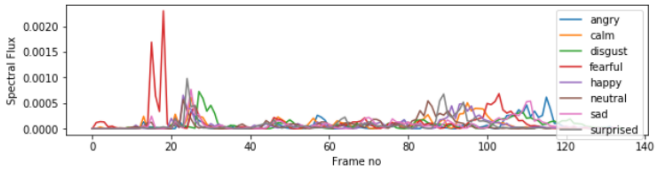

Fig. 5. Spectral Flux against Frame No for different emoions.

made it stand out from other emotions. Spikes can again be observed at the end of the sentences.

The graph of Chroma Deviation for disgust emotion also had a plateau phase making it easily traceable. For the same frame number, calm emotion has a plateau phase however it is not as prominent as the disgust emotion.

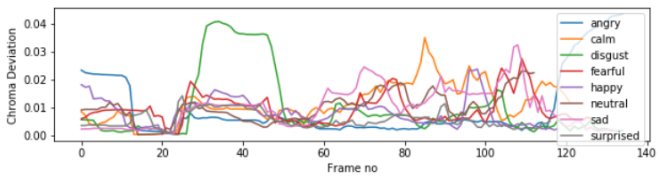

Fig. 6. Chroma Deviation against Frame No for different emoions.

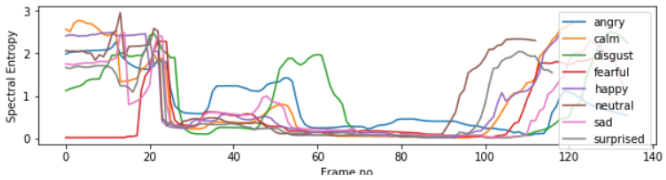

Fig. 7. Spectral Entropy against Frame No for different emoions.

VII. Result

The model initially employed in this study is Support Vector Machines (SVM). The model was trained with all the speech samples. This accounted for 1,361 speech samples. The accuracy obtained with this approach was 52.9% when regularization parameter *C* was set to 0.500. KNN approach yielded an accuracy of 53.7%. Later a random forest approach was also employed which yielded an accuracy of 54.5%.

To further investigate the data, the model was trained with only speech samples of males. This accounted for 725 speech samples produced by male actors. The accuracy obtained with this approach with SVM was 49.9% when regularization parameter *C* was set to 0.500, with KNN an accuracy of 52.3% was obtained and for Random Forest the accuracy obtained was 56.2%.

Next, the model was trained with only speech samples of females. This accounted for 636 speech samples produced by female actors. The accuracy obtained with this approach was 61.9% when regularization parameter *C* was set to 0.500, with KNN an accuracy of 5.0% was obtained and with Random Forest the accuracy obtained was 60.1%.

When a subset of the database, which included only angry, happy, neutral and sad emotions was considered an accuracy of 65.6% was obtained with SVM classifier, 68.2% for KNN classifier and of 66.4% was obtained with Random Forest classifier.

For male only subset, SVM classifier obtained an accuracy of 67.1%, KNN classifier obtained an accuracy of 69.9%, and Random Forest obtained an accuracy of 72.8%.

For female only subset, SVM classifier obtained an accuracy of 72.6%, KNN classifier obtained an accuracy of 69.9%, and Random Forest obtained an accuracy of 70.2%.

As evident from the above analysis, it is seen that Random Forest model had the best accuracy for the complete dataset. It also had a better classification for Random forest model trained only on male speech dataset. However, SVM model was only slightly better than random forest for female only dataset.

The subset dataset performed better than the dataset containing all the emotions. It is only natural because the complexity of emotions required to identified was reduced. SVM model performed better than other models when all the samples were taken account. However, Random Forest model outperformed when samples of only male speech and only female speech were taken separately.

VIII. Conclusion and Future Work

The accuracy obtained in this study cannot be the compared to other studies as there is a difference in the datasets. Also, the dataset used here was originally intended to incorporate visual cues such as facial features which were recorded in video samples and contained songs sung by various actors. The accuracies obtained for data from female speakers is more than with the male speakers. This is in accordance with other studies which found that accuracies on a data consisting of only female speakers are higher than the data consisting of only male speakers.

Majority of the datasets available for Speech Sentiment Analysis have prompted emotions. That is it involves speech samples which are produced by identical utterances of a speech in a given emotion. As these speeches are a deliberate effort it may not always be like an unprompted speech which are more natural in nature. However, the major drawback in obtaining unprompted speech samples is that it would require more human effort and time. It would also involve recording speech samples all the time which may lead to privacy concerns.

To improve the accuracy of this model the database can be expanded with other speech samples, as there is only limited number of speech corpus in this dataset. Also, incorporating a mechanism which would in real time update the models could also be looked into. However, this approach would require the active interaction of a participant to annotate the data in real time.

The possible use case of such an approach could be identifying the emotion of a customer on the other side of the phone, making intelligent home agents identify emotional state of humans from mere simple sentences, automating call analysis to assess customer satisfaction, assessing if a person has to post-traumatic stress disorder by extrapolating speech cues, lie detection.

Acknowledgment

I would like to express our appreciation to Dr. Samira Shaikh, who helped us during the process of this paper formation.

References

[1] L. Kaushik, A. Sangwan, and J. H. Hansen, "Sentiment Extraction from Natural Audio Streams" https://www.utdallas.edu/~john.hansen/Publications/CP-ICASSP13-KaushikSangwanHansen-Sentiment-0008485.pdf

[2] S. R. Livingstone, K. Peck, and F. A. Russo (2012). RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song. Paper presented at the 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS), Kingston, ON.

[3] W. Medhat, A. Hassan, H. Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, Sciencedirect, 27 May 2014

[4] Y. Chavhan, M. L. Dhore and P. Yesaware, "Speech Emotion Recognition Using Support Vector Machine." International Journal of Computer Applications 1(20):69, February 2010. Published By Foundation of Computer Science.

[5] Mp3tag. Florian Heidenreich. March 20, 2018. Retrieved March 20, 2018.

[6] Audacity® software is copyright ©1999-2018 Audacity Team. The name Audacity® is a registered trademark of Dominic Mazzoni.

[7] C. Sobin and M. Alpert, “Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy.” Journal of Psycholinguistic Research, Vol. 28, No. 4, 1999

[8] T. Giannakopoulos (2015) pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLoS ONE 10(12): e0144610. https://doi.org/10.1371/journal.pone.0144610

[9] P. B. Dasgupta, “Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing.” International Journal of Computer Trends and Technology (IJCTT) Volume 52 Number 1 October 2017

[10] R. Lyon and S. Shamma, “Auditory Representation of Timbre and Pitch”. Harold L. Hawkins & Teresa A. McMullen. Auditory Computation. Springer. pp. 22123. ISBN 978-0-387-97843-7, 1996

[11] S. S. Jarande, S. Waghmar, “Speech based Human Emotion Recognition Using Hybrid Classifier Technique” International Journal of Advanced Research in Computer Science and Software Engineering