Feasibility of a shared model for speech and song emotion recognition

Muhammad Asim Ali
Dept. of Software Engineering
University of Victoria
maali@uvic.ca

Second author
Retain these fake authors in submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

In an era where we have ai companion programs like Siri and Alexa becoming a part of everyday life the subfield of developing AI to understand emotion is becoming more important. Emotion is an important aspect of communication expressed over many different domains, the focus of this study will be on vocal expressions of emotion through speech and songs. Research has shown that there is a clear link between emotion expressed in songs and speech [1],[2].A shared model would be useful because it would generalize across vocal communication which would help "combat data scarcity" as worded by Zheng et al.[3].

1. INTRODUCTION

Despite the clear association between songs and speech there is still work to be done on developing a generalized model that works well for classifying emotion for both speech and song.The experiments in this project will be conducted on 3 models (see figure 1), a model trained on only speech data with speech specific features, a model trained on only song data with song specific features, and a third model trained on both song and speech data with features appropriate to the training domain. The accuracy of emotion recognition in these experiments are expected to give indications on the similarity between expressions of emotion by mediums of speech and songs and also provide an insight into the feasibility of generalized emotion classification models.

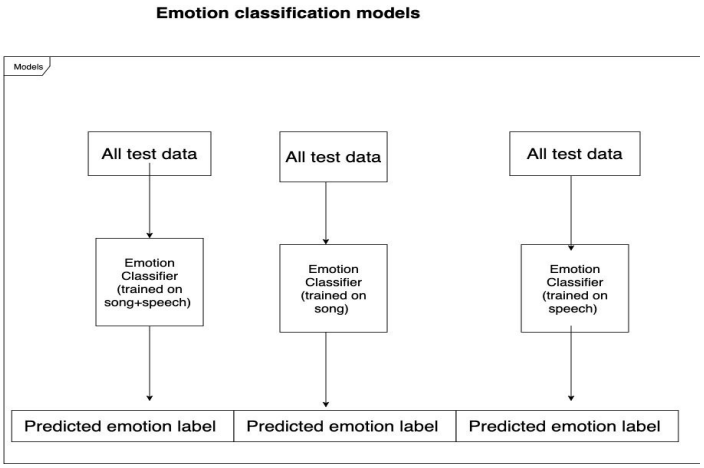

Figure 1. The 3 models as experiments for this project

2. RELATED WORKS

2.1 How expressions of emotion in speech and sound is related

There have been many studies that have shown the existence of a relationship between emotion expressed through speech and singing. Livingstone et al. [1] found that emotion was conveyed similarly in many acoustic features across speech and songs. Furthermore, Schere et al.[2] found that there was a high degree of similarity between patterns of sung expressions and spoken expressions of emotion. The work of Wenniger et. al. [4] confirmed that acoustic parameters marking certain emotions are quite similar in songs and speech. All these studies support the idea that it may be possible to build a model that can work well for recognizing emotion from both speech and songs.

2.2 Shared model for emotion classification through speech and sound

There has been a significant amount of work done on developing emotion recognizing models that are speech specific or song specific. [5]-[7]. However, there has been comparatively less research done into developing a generalized model that can recognize emotion from both speech and songs. Zhang et al [3] explored three techniques for shared emotion models for speech and song; in their research they report as a conclusion to their studies that speech and sung expressions of emotion are related and can be considered together in a shared model. In another paper, Zheng et al.[8] explain that even though spoken and sung emotion recognition are different tasks they are related and by taking advantage of their relatedness the classification accuracy of models can be improved.

3. Dataset

This project will use the Ravdees dataset [9]. This dataset contains 1440 audio files with 24 actors and each actor having 60 recordings. The files are recorded by 12 male and 12 female actors. The dataset is based on the actors repeating predefined statements in different emotions. The following emotions are available in speech: calm, happy, sad, angry, fearful, surprise, and disgust. In the ravdees dataset songs are limited to only the following emotions: calm, happy, sad, angry, and fearful. This project will be studying the emotions that are common to both speech and song (calm, happy,sad,angry,and fearful emotions).

4. Tools

4.1 Librosa Library

The librosa package [10] for python will be used in this project because it provides many features that are useful for music and audio analysis. One main use of the librosa library in this project will be the use of the MFCC feature extraction method. MFCC [11] is a popular feature extraction algorithm for audio signals that is known for creating features that resemble how humans perceive frequency.

4.2 Scikit Learn

The scikit learn package [12] for python will be used for creating a multiclass support vector machine (SVM) classification model. SVM is a supervised machine learning algorithm used for classification. [13]

Objective description	Deadline
Download data, split data into test/train (different for each of the three models) , and any other preprocessing tasks.	October 31st, 2019
Perform additional research to find which features work best for the respective domain (speech or song).	November 7,2019
Train 3 SVM classifiers as described in the 3 models of figure 1.	November 17,2019
Write report	November 30, 2019

Table 1. An overview of the milestones in this project and their tentative deadlines .

5. Progress Report

5.1 Progress to date

Since the design specification I have downloaded the RAVDEES dataset and processed the audio data into sung and speech data lists. I have also extracted the emotion labels corresponding to each of the audio files.I have extracted MFCC features from the raw audio files (see visualizations of the MFCC features in figure 2). Furthermore, I split the data into training and test sets according to the specification provided by model 1(the model trained on speech and song data) (see figure 1). And I have trained a linear SVM classifier that reported whose predicted accuracy is 48%. See figure 3 for the confusion matrix visualizing the results of the classifier.

5.2 Minimally viable project

In the worst case of the project I would create all three models without applying any data processing that would help improve the classifier accuracy. As has been

described in section 7.1 I would extract MFCC features on the raw audio files and do the appropriate test/train split as described by the models in figure 1 and use the same linear svm classifier.The results of this project would still be indicative of the feasibility of a shared model,between speech and sung expression, for emotion classification.

5.3 Expected goals for the project

An expected scenario for the project is to do all the steps explained the section 7.1 (take mfcc features, do train/test split) and further process the data to improve the classification accuracy. This can be done by normalizing the data and using a uniform length of audio across speech and sung (included in this is removing the beginnings and endings of audio that are completely silent). Another step that can be taken to improve the accuracy of the classifier is to choose a classifier that works well with the data such as SVM classifier with the RBF kernel. The results from this project would provide enough information to write a good rationalization on the feasibility of a shared model for emotion classification.

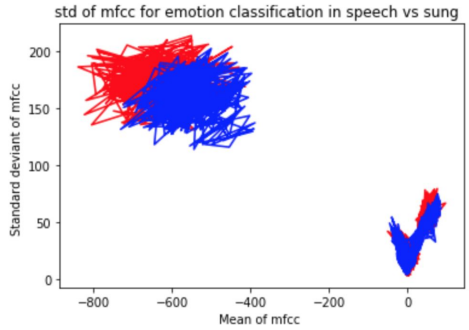

Figure 2. MFCC features visualized. (Blue is song MFCC features and Red is speech MFCC features).

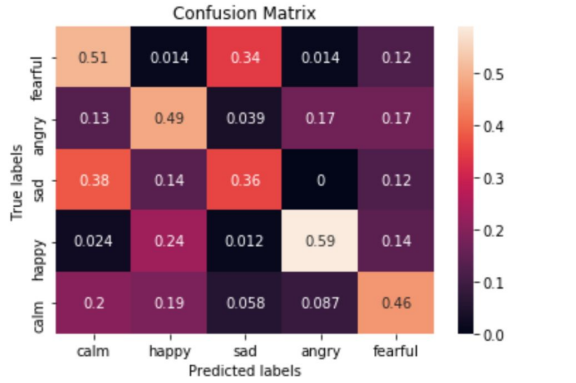

Figure 3. Confusion matrix representing results for the classifier trained on speech and song data.

6. Data processing

6.1 Trimming audio for Feature extraction

The original audio files had various amounts of silence in the beginning and endings of the audio, this silence was detected and removed from all files before feature extraction. MFCC features were extracted from the time series form of the audio files using default settings for the librosa library mfcc implementation**.** Each audio file was summarized by 20 MFCCS, thereafter the mean of these MFCCS was taken to summarize each audio recording with 20 mean MFCC values.

6.2 Data standardization and PCA

The PCA algorithm (Principal component analysis) is a dimensionality reduction algorithm that transforms the original data into a lesser dimensional space while still preserving the information represented by our original data. The first PCA dimension has the highest variance of the data and the variance decreases as the dimensions decrease. The standardization of data is important before PCA because the PCA gives more importance to variables with high variance.Both the standardization and PCA computations were done using the scikit learn library.

6.3 Test train split

The train_test module from scikit learn was used to create the train and test split of the data. Each of the classifier was trained with 70% of the data and tested with 30% of the data.

6.3 Pre processing for every experiment

Each audio file was summarized by 20 mean mfcc features thereafter the MFCC data was standardized and the PCA was computed which decreased the number of dimensions of the data from 20 to 10 values for each audio file.

7. Results

7.1 Speech only

The speech only experiment was performed as follows: train classifiers with only speech audio files and test the classifiers with only speech audio files. The classification accuracy of the speech only experiment was 58.6% with the svm + rbf kernel classifier and it was 65.6% with scikit neural network (using adam solver).

Figure 3 shows a visualization of a scatter plot of the MFCC features as a scatter plot. One can see from figure 3 that the emotions that appear most distinct with respect to each other (and as a result classifiable) are calm/sad, fear/happy, and angry. There is a lot of overlap between calm and sad, and fear and happy emotions which means we expect the emotions to be misclassified as each other. This intuition was verified by the results in the confusion matrix shown in figure 4. The emotion sad was misclassified as calm 35% of the time and the emotion fearful was misclassified as happy 24% of the time.

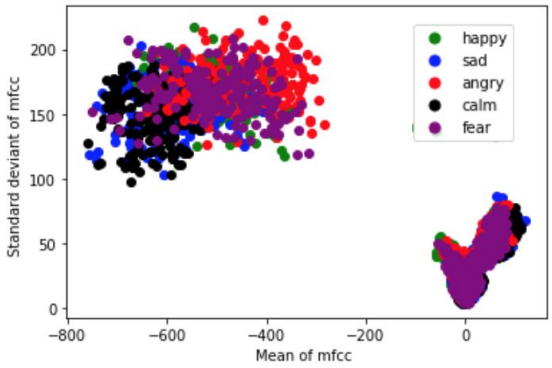

Figure 3. MFCC features visualized (All speech emotions)

Figure 4. Confusion matrix for Speech only results

7.2 Sung only

The sung only experiment was performed as follows: create a model trained on only sung audio files and test it on only sung audio files.The classification accuracy of the sung only experiment was 72.1% with the svm + rbf kernel classifier and it was 75.4% with scikit neural network (using lbfgs solver).

Figure 5 shows a visualization of a scatter plot of the MFCC features. One can see from figure 5 that there is a lot of overlap between calm and sad, fear and happy emotions, and fear and calm emotions. This overlap implied that there would be misclassification between those specific classes. This intuition was verified by the results in the confusion matrix shown in figure 6. The emotion calm was misclassified as sad 26% of the time, the emotion fearful was misclassified as sad 21% of the

time, and the emotion happy was misclassified as fearful 20% of the time.

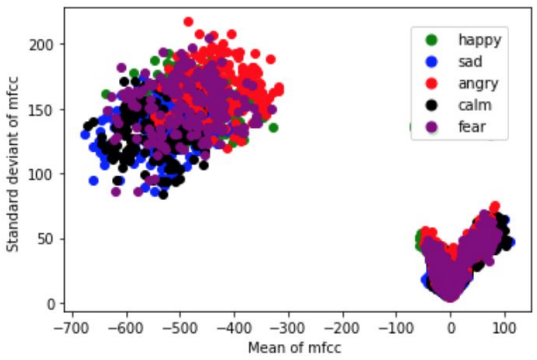

Figure 5. MFCC features visualized (All sung emotions)

Figure 6. Confusion matrix for Sung only results

7.3 Shared model trained on song only

This experiment was carried out as follows: train models on song data and test with both speech and song data. There was an equal proportion of test data between speech and song.The classification accuracy of this model was 19% with the svm + rbf kernel classifier and it was 20% with scikit neural network (using adam solver).

Figure 7 shows the confusion plot for this model. The results show that almost all the data was classified as calm even though there was a good distribution in the training and testing data over all the different emotion labels.

Figure 7. Confusion matrix for shared model trained on song only

7.4 Shared model trained on speech only

This experiment was carried out as follows: train models on speech data and test with both speech and song data. There was an equal proportion of test data between speech and song.The classification accuracy of this model was 20% with the svm + rbf kernel classifier and it was 22% with scikit neural network (using adam solver).

Figure 8 shows the confusion plot for this model. The results show that almost all the data was classified as fearful even though there was a good distribution in the training and testing data over all the different emotion labels.

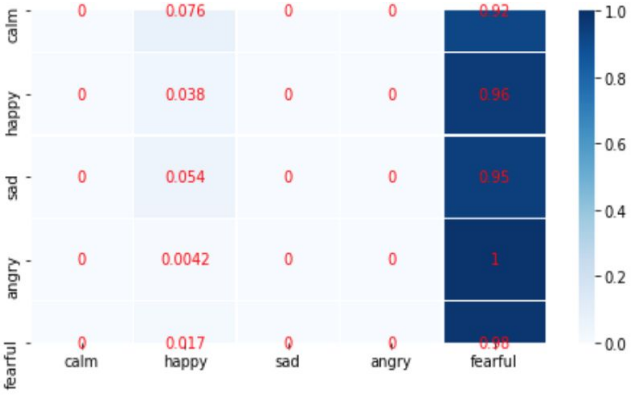

Figure 8. Confusion matrix for the shared model trained on speech

7.5 Shared model trained on speech and song

This experiment was performed as follows: train the classifier on both the sung and speech audio files and then test the classifier with both the sung and audio test files. There was an equal portion of song and speech data in both the train and test data.The classification accuracy of this shared model experiment was 42.0% with the svm + rbf kernel classifier and it was 41.0% with scikit neural network (using lbfgs solver).

Figure 9 Confusion matrix for shared model trained on sung and audio

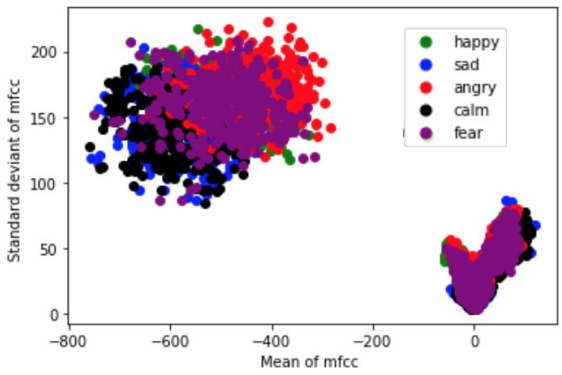

Figure 10. MFCC features visualized (All sung and speech emotions)

8. Conclusion

Figures 11 and 12 are visualizations of the principle component 1 vs principle component 2 of the sung and speech mean MFCC data respectively. The plots show that in both the sung audio and the speech audio data can be clustered into the following categories: angry, happy/fearful, and calm/sad. The results of the sung only, speech only, and the shared model classifications confirm the existence of such clusterings in the data. I have already explained these clustering in the speech only and song only models and they can also be seen in the shared model trained on speech and song in figure 9 which shows that calm is misclassified as sad 26% of the time, sad is misclassified as calm 35% of the time, happy is misclassified as fearful 33% of the time. These results imply that emotion is not discrete but is more continuous in nature. For example, the emotion fearful has dots all across the PCA plots this means that there are different intensities of fear and that fear is more continuous in nature rather. Furthermore, the results also show that the labelling of emotions may depend from person to person as the data shows quite a significant amount of overlap between calm and sad and fearful and happy emotion. After listening to some of the recordings, I too was sometimes unsure whether the emotion enacted was happy or fearful. To conclude, all these results indicate that with the model of emotion chosen to study emotion classification in this paper does not lend itself for the feasibility of a shared model for the classification of emotion from both speech and sung data. However it does indicate that the shared model may be more successful if a more continuous emotion model is chosen.

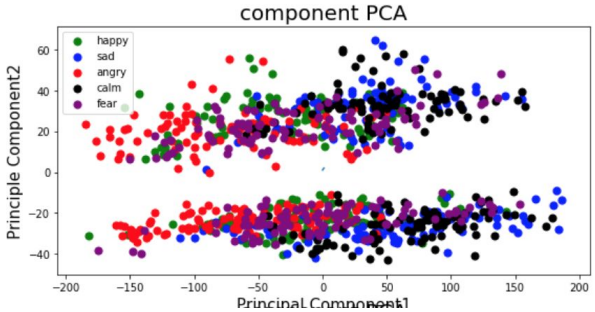

Figure 11. PCA component scatterplot for song data (PC1 vs PC2)

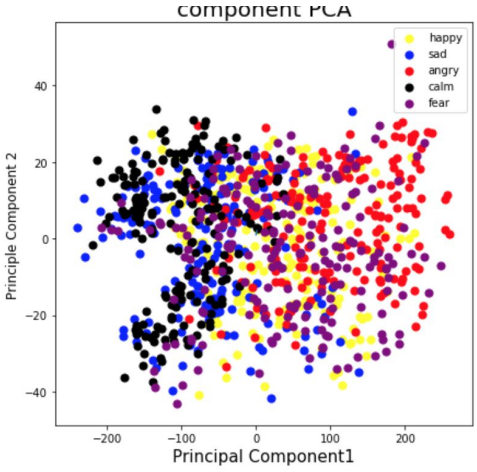

Figure 12. PCA component scatterplot for speech data (PC1 vs PC2)

9. REFERENCES

[1] S. R. Livingstone, K. Peck, F. A. Russo, "Acoustic differences in the speaking and singing voice", *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013.

[2] Klaus R. Scherer, Johan Sundberg, Lucas Tamarit, Gláucia L. Salomão,"Comparing the acoustic expression of emotion in the speaking and the singing voice" Computer Speech & Language, Volume 29, Issue 1, 2015, Pages 218-235,

[3] B. Zhang, G. Essl and E. M. Provost, "Recognizing emotion from singing and speaking using shared models," *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, 2015, pp. 139-145.

[4] Weninger, Felix, et al. "On the acoustics of emotion in audio: what speech, music, and sound have in common." *Frontiers in psychology* 4 (2013): 292.

[5] Kim, Youngmoo E., et al. "Music emotion recognition: A state of the art review." *Proc. ISMIR*. Vol. 86. 2010.

[6] Yang, Yi-Hsuan, and H. Homer. "Chen,"Machine Recognition of Music Emotion: A Review", TIST, May."

[7] Mower, Emily, Maja J. Mataric, and Shrikanth Narayanan. "A framework for automatic human emotion classification using emotion profiles." *IEEE Transactions on Audio, Speech, and Language Processing* 19.5 (2010): 1057-1070.

[8] Zhang, Biqiao, Emily Mower Provost, and Georg Essl. "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[9] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[10] McFee, Brian, et al. "librosa: Audio and music signal analysis in python." *Proceedings of the 14th python in science conference*. Vol. 8. 2015.

[11] MFCC : N. Lulla and N. Purohit, "An improved algorithm for efficient computation of MFCC," *2014 Annual IEEE India Conference (INDICON)*, Pune, 2014, pp. 1-4.

[12] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12.Oct (2011): 2825-2830.

[13] Scholkopf, Bernhard, and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.