Audio-Visual Emotion and Sentiment Research

Enis Berk Çoban, Yunhua Zhao, Patrick Jean-Baptiste
The Graduate Center, CUNY

1. Introduction

Our project goal is to detect the emotion of a person from audio visual data. Emotion recognition is the process of identifying human emotion,figure1, it is a focus of attention by researchers in neuroscience, psychology, psychiatry, audiology, and computer science over the last decade. Some researches focus on audio emotion recognition while others work with video, in our project, we try both, also we try to concatenate audio and video together to detect the emotion of a person.

We use both machine learning algorithms and deep neural networks to our dataset to compare their results. For the audio-only part, we separately use CNN and RNN to train our audio models and visualize the results of different models. For the video part, we sample images from the video files, then extract the facial parts of one image, and we use the Conv2D in our network. We implement our RNN network to the audio and video concatenated features.

Our best model is the RNN model to the audio-only data set, and the best accuracy result is 75%.

figure 1. example of emotion recognition

2. Dataset

In our project, we decided to use The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).[1] The database contains 24 people (12 female, 12 male), speaking and singing in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and the song contains calm, happy, sad, angry, and fearful emotions. Recordings are

[1] https://zenodo.org/record/1188976#.XqN9yG5Fw2x

produced at two levels of emotional style (normal, strong) All records are provided in three modality formats which are Audio-only, Audio-Video, and Video-only (no sound).

Authors of the dataset Livingstone et.al. designed a validation task to see if recordings are genuinely interpreted as their intended emotions. Following heatmaps (Figure 1) shows the confusion matrices of mean proportion correct scores for actors' intended emotions as per rater chosen emotion labels for (A) Speech and (B) Song.

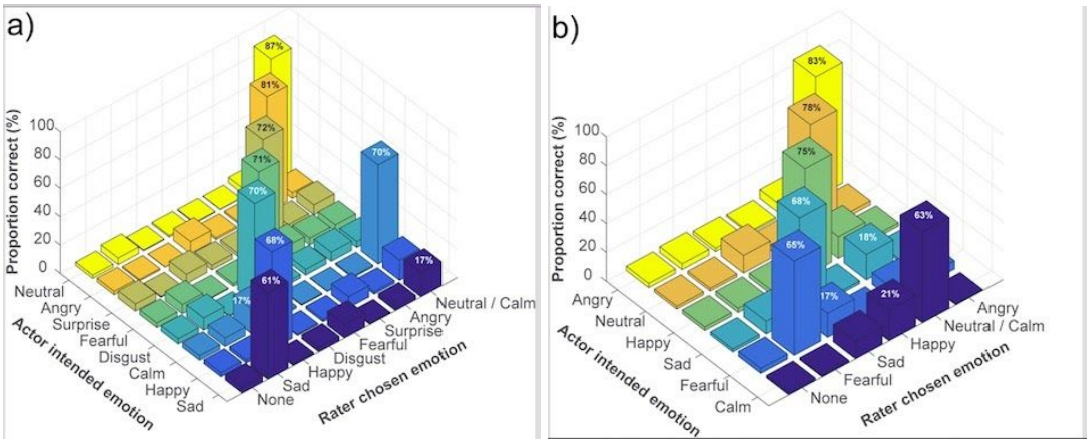


Figure-2: Heatmap of correct scores labeled by humans

Following carousel of images are sample data points of different emotions:

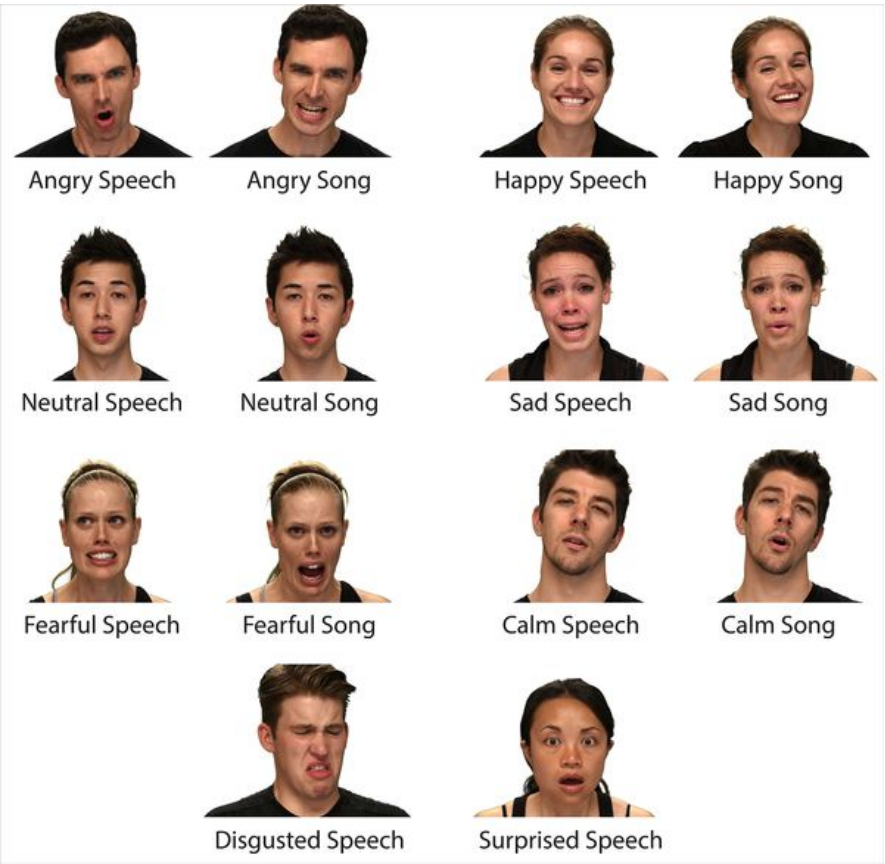


Figure-3

From images, you can see that the background is white and the actor is alone in the image. While some of the emotions are can easly be read from images like disgusted, some of them hard to distinguish such as Calm and Neutral.

3. Experiments

a. Audio Only

For the audio-song only data set, there are 1012 audio files; For the audio-speech only data set, there are totally 1440 audio files, and together there are totally 2452 audio files.

i. LSTM

Because audio is a continuous data set, and the precious audio has influence on its later audios, so we use the LSTM model to our audio data set.

For this model part, we use the LSTM model on audio-speech only data set, also we implement the same model on the audio-song only data set, then implement the same model to the merged audio speech and song data sets.

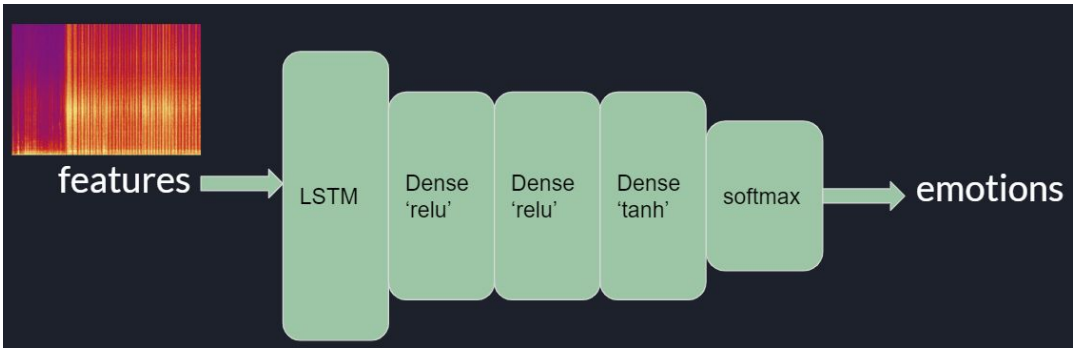
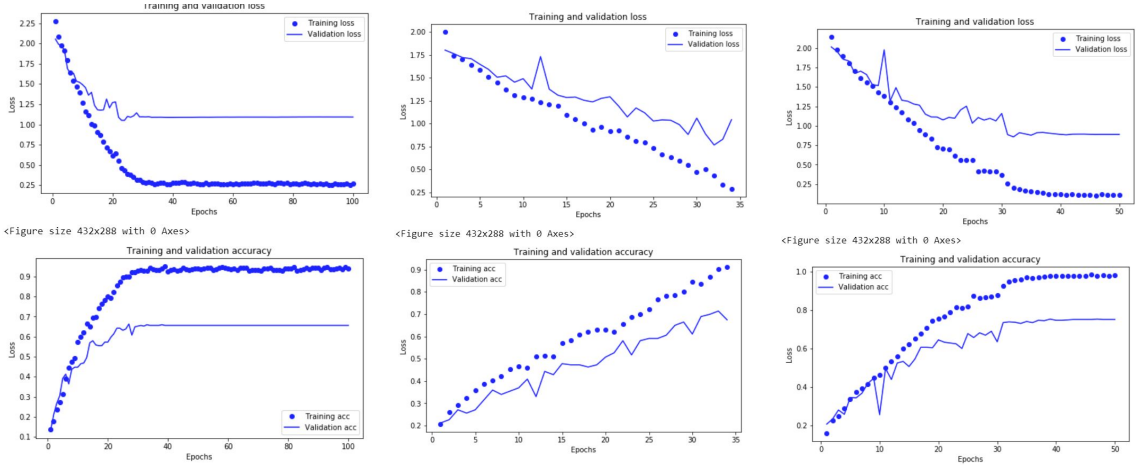


figure 4. LSTM model

Figure 4 shows the structure of our LSTM model, the first layer is LSTM layer with 128 neurons, then there are two dense layers which activation is Relu with 32 neurons, and following a third dense layer which activation isTanh with 16 neurons, and last layer is a softmax layer.



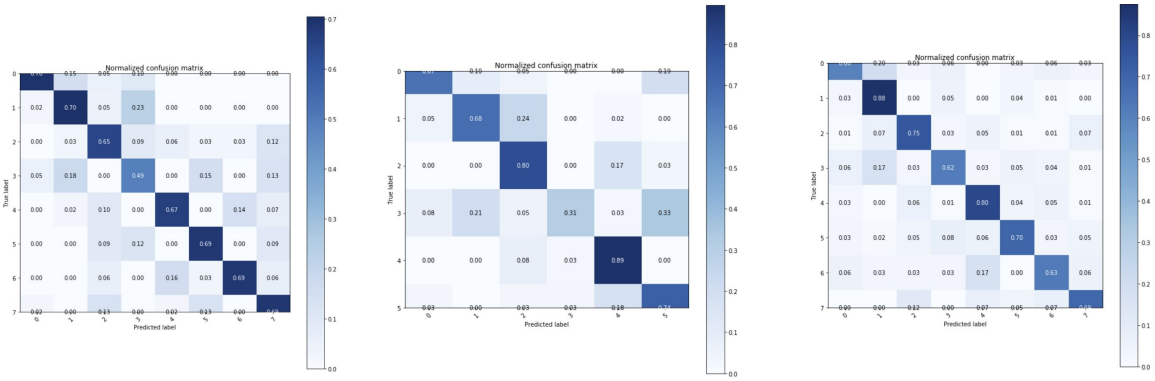


Figure 5 shows the final result of LSTM(The first column is the audio-speech only result, the second column is the audio-song only result, the third column is the audio speech and song together result)

Figure 5 shows the result, the audio-speech only accuracy is 66%, the audio-song only accuracy is 67%, the total audios mixed speech and song audios got 75%, which is the best one.

We also tried to use the transfer model to extract features first, then feed into our LSTM model, the result is just 49%, not as good as extract features directly from the original audio files.

ii. Transfer Learning

Transfer learning involves the use of a model pre-trained on a large mismatched dataset to generate input features for another machine learning system trained on the target dataset. One popular pre-trained model for audio is VGGish[2]. This part of the project investigates whether the VGGish model and the related Audio Set dataset, both based on soundtracks of YouTube videos, can be effectively applied to the analysis of emotion recordings.

VGGish's architecture is similar to the vision system VGGNet; VGGNet is a deep convolutional network originally developed for object recognition. VGGish's only difference is that it has 3087 sigmoid units in its output layer. While the original VGGNet has 144M weights and 20B multiplies, the audio variant uses 62M weights and 2.4B multiplies. VGGish is trained on the Youtube-100M dataset. We use the model to predict 128-dimensional embedding vectors from non-overlapping 960~ms segments of audio. Each segment is processed with a short-time Fourier transformation with 25~ms windows calculated every 10~ms. The generated spectrograms are converted to 64 mel-spaced frequency bins and the magnitude of each bin is log transformed.

[2] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017.

We use a pre-trained VGGish model to generate audio embedding vectors with the size of 128. We also generated a normalized set of embeddings by post processing with PCA and whitening.

We fed these embeddings as features to classical machine learning algorithms such as Nearest Neighbors, Linear SVM, and Decision Tree. We also tried different MLP architectures. We achieved 62% accuracy with MLP and the best performing classical ML algorithm achieved 55% which was Linear SVM.

b. Video Only

For the visual part of the project, the goal was to perform emotion recognition using videos of actors expressing emotions. The initial step was to detect the actor's face in the video followed by extracting the image displaying the expressed emotion. The final step involved creating visual models using TensorFlow to recognize emotions from the images. The videos used were the video only and audio-visual files of the RAVDESS dataset for both speech and song. There were a total of 4092 videos that were processed.

i. Detection and Extraction

The detection and extraction step consisted of extracting one image from each video, which displayed the detected face of an actor expressing an emotion. First, video preprocessing needed to be performed. This entailed reading a video file, extracting a video frame, and converting the video frame to grayscale. Only the middle frame of the video was used since it was the most likely frame to show an expressed emotion.

Once the video frame was obtained from the video, the next phase was to detect the face of the actor in the video frame. The face detector employed was a pre-trained classifier for faces. It detected the face of the actor in every video without error. The last part of this phase was to draw a bounding box within the video frame indicating the detected face. Figure 6 displays an example of a video frame with the detected face indicated by the black boundary box. Using the boundary box, the detected face was extracted or cropped from the image. A sample of an extracted image is shown in Figure 7. The extracted images depicting the expressed emotions of the actors were later collected and stored in a folder for easy access.

Figure 6. Example of a detected face.

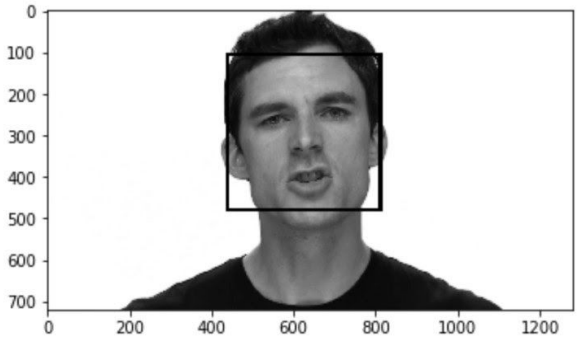


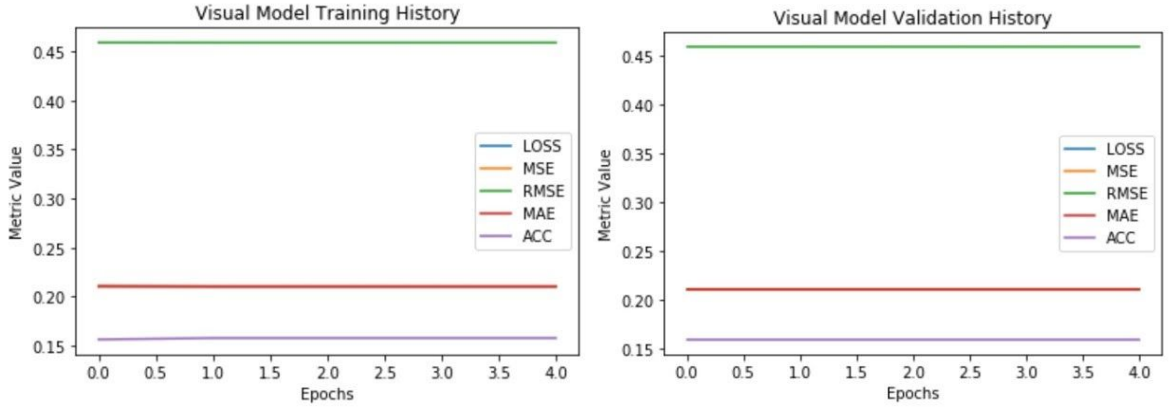
Figure 7. Example of an extracted face.

ii. Visual Emotion Recognition

With the images of emotion extracted, it was now time to proceed to the second step of building models for the task of visual emotion recognition. Then, create a model to recognize the emotions expressed in these images. Several models were built with different convolutional and dense layers as well as differing numbers of nodes. Different training parameters including the loss function, optimizer, learning rate, epochs, and batch size were utilized as well.

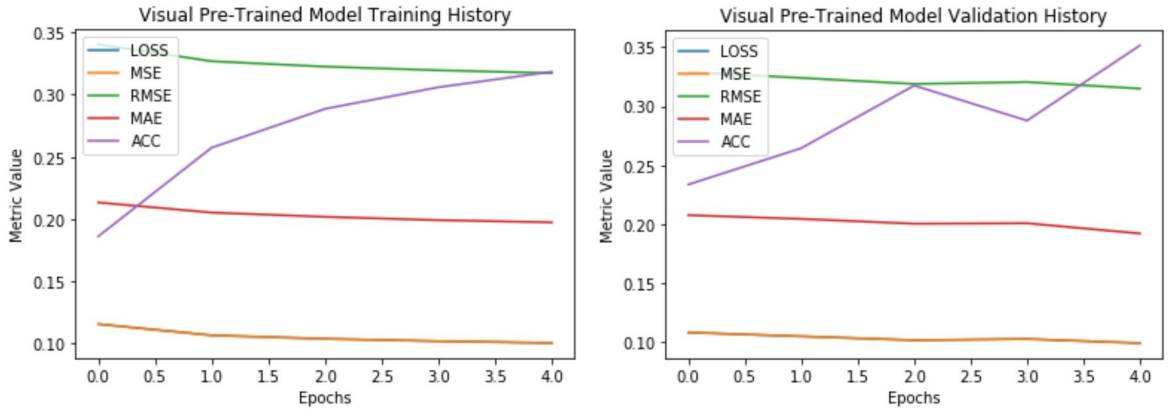
In general, the models were not trained with more than five epochs given that either the models would take more than ninety minutes to train or the program would crash. This led to underfit models, or models with accuracy and loss that would not improve. Figure 9 displays the training and validation history of a model. As seen, it is clear that the model accuracy does not improve after three epochs. These models evidently generated misclassifications on a set of test images due to the models' poor training. The visual emotion recognition accuracy varied between 15% and 19% for the song, speech, and speech + song sets of images. Some of the predicted emotions were understandable, in a way, such as the image expressing a fearful emotion, which looks like a sad emotion. Altering the model architecture and training parameters still produced poor emotion recognition performance.

In an attempt to better the visual emotion recognition results, a VGG16 model pre-trained on ImageNet was employed. Despite using the same number of epochs, the visual emotion recognition accuracy did slightly increase using the pre-trained model as shown in Figure 10. Accuracy was still increasing when training ended suggesting that this pre-trained model was underfitting as well. Final recognition accuracy utilizing this model ranged between 30% and 34% with an average accuracy of 32%. Overall, training with more epochs in addition to incorporating more images for each emotion class could significantly improve visual emotion recognition performance.



(a) (b)

Figure 9. a) The plot on the left shows the training progress of a model. b). The plot on the right displays the validation history.



(a) (b)

Figure 10. a) The plot on the left shows the training progress of the pre-trained model. b). The plot on the right displays the pre-trained model's validation history.

c. Merged

The highest accuracy of audio-only and video-only is 75%, so we want to know if the accuracy could improve when we merge audio and video files together. In our project, we merge two model together and also merge the original features.

1) Training two models together

Since merging inputs to train a single model did not perform so well, we decided to train a single model which processes two inputs separately. In this experiment the first module is based on our best performing audio processing module and the second part is a three layer CNN, at the end two modules are merged by a fully connected layer. As shown in the following figure 11. This strategy performed poorly as well and we could not improve our results.

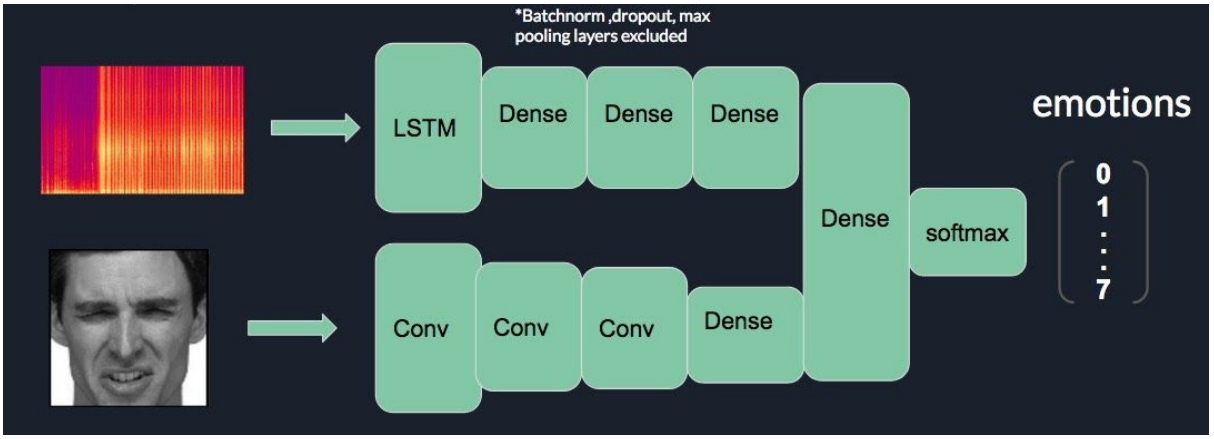


Figure 11: NN architecture that has different modules for processing audio and visual inputs

2) Combining inputs for single model

For merging features,shown in figure 12, first we use mfcc to extract audio features from the original audio files, and we use the facial images which the

previous video part gets to extract features, then we concatenate audio and video features as one and input to a dense neural network, the result is not good, shown as figure 13.

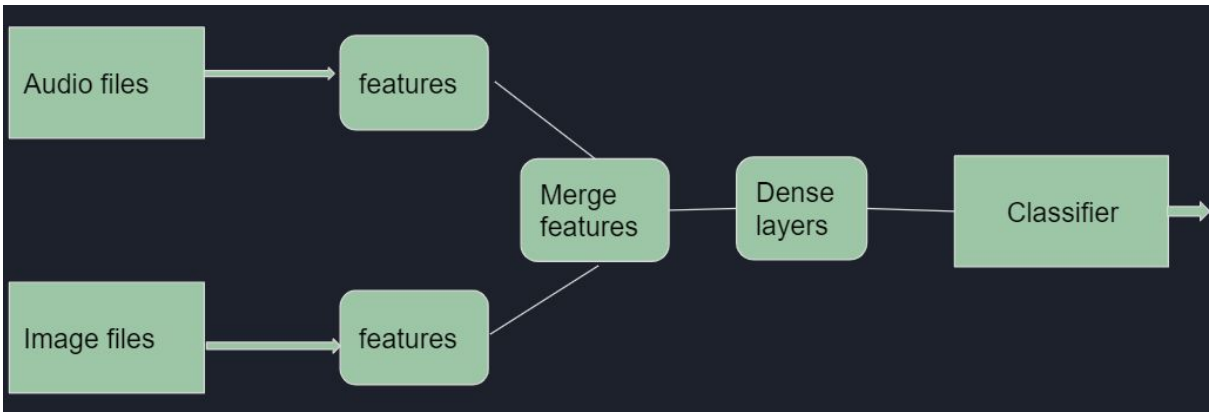


Figure 12: NN architecture that has different modules for processing audio and visual inputs

```
Classification report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        22
           1       0.00      0.00      0.00        39
           2       0.00      0.00      0.00        33
           3       0.00      0.00      0.00        42
           4       0.11      1.00      0.21        34
           5       0.00      0.00      0.00        42
           6       0.00      0.00      0.00        32
           7       0.00      0.00      0.00        52

    accuracy                           0.11       296
   macro avg       0.01      0.12      0.03       296
weighted avg       0.01      0.11      0.02       296

Confusion matrix:
[[ 0  0  0  0 22  0  0  0]
 [ 0  0  0  0 39  0  0  0]
 [ 0  0  0  0 33  0  0  0]
 [ 0  0  0  0 42  0  0  0]
 [ 0  0  0  0 34  0  0  0]
 [ 0  0  0  0 42  0  0  0]
 [ 0  0  0  0 32  0  0  0]
 [ 0  0  0  0 52  0  0  0]]
```

figure 13. audio and video merge result

4. Conclusion

We conducted several experiments for audio, visual, and audio-visual emotion recognition. The results of the emotion recognition accuracy using distinct models are shown in Table 1. We use machine learning algorithms to compare with deep neural networks, results show that the machine learning algorithms could work as

well as deep neural networks when the data set is not very large.

Models	Song	Speech	Song + Speech
VGGish-Deep	49%	48%	62%
VGGish-ML	47%	51%	54%
LSTM	**67%**	**66%**	**75%**
Visual	17%	18%	16%
Visual Pre-trained VGG	32%		
Audio-Visual	11%		

Table 1. Emotion recognition results for audio only, video only, and audio-visual employing different models.

We try our models on audio-speech only, audio-song only and total speech-song together audio files, the total files work better, the reason may be because the data set is bigger. We implement both the LSTM model and VGGish model, the results show the LSTM model works better, it is probability because the features of the audio, and LSTM works better than CNN on the continuous data set. For the visual part, our models perform poorly on all three sets of video files. The pre-trained VGG gives the best performance for visual emotion recognition, but the recognition accuracy is, nevertheless, very low. Increasing the number of images in each emotion class and training with more epochs could significantly improve visual emotion recognition results.

We use our best model "LSTM model" from the previous experiments for performing the audio-visual emotion recognition task. Original features are merged together directly. The end results for audio-visual emotion recognition are worse than the audio only and video only performances. This is primarily due to the fact that there were not enough images extracted from the videos for training.

5. References

[1] Livingstone, S. R., & Russo, F. A. (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).