

MASTER THESIS  
EMOTION RECOGNITION USING FUSION OF AUDIO AND VIDEO  
FEATURES

Grigor Keropyan

Yerevan State University  
May 2021

Master thesis submitted to the department of  
Mathematics and Mechanics  
Applied Statistics and Data Science program

Supervisor: Vazgen Mikayelyan

# Abstract

Emotion Recognition is ability to understand what feelings are others want to transfer. Emotion Recognition is possible to done from a different modalities (e.g. speech, video, text, etc). Understanding others transferred feelings is important for the reasons: in human-human interaction and human-computer interaction. As understanding emotion even is not easy for humans it is a great challenge for computers being able to recognize emotions with a high accuracy. However, with the rapid development of the field of Machine Learning and Deep Learning it seems reasonable to apply this techniques in Emotion Recognition. Existing methods for Emotion Recognition mainly are divided into three parts: feature level fusion, decision level fusion and hybrid fusion (which will be discussed in the following chapters in more detail). First two missing the principle to use the whole content of multi-modality, However, hybrid fusion could reasonably use the existing information in multi-modality and have a chance to outperform existing methods. This work explores multi-modal (Audio and Audio-Video fusion) Emotion Recognition using Deep Learning techniques and the method is based on the hybrid fusion.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Existing Datasets and Methods</b>	<b>3</b>
2.1 Datasets . . . . .	3
2.2 Methods . . . . .	4
2.2.1 Features . . . . .	4
2.2.2 Bi-modal Approaches . . . . .	7
<b>3 Proposed Method</b>	<b>10</b>
<b>4 Experiments and Results</b>	<b>14</b>
<b>5 Conclusion</b>	<b>16</b>

# Chapter 1

## Introduction

Emotions are indivisible part of humans life and play an important role in different types of interactions [18]. In human-human interactions, recognizing others emotion is deciding factors how deeply people understand others. Moreover, in human-computer interactions (HCI) [3, 6], emotions are becoming an vital as the interaction between humans and computer growing rapidly. Taking into account COVID-19 situation, this last year most of the universities have changed their studies from onsite to online and teaching is becoming more challenging. Having a good autonomous emotion recognizer could help improve the current situation by making people's life easier providing some feedback about their emotional state during a conversations and lectures. Also, Emotion Recognition (EM) is promoting to a harmony of HCI [22] (e.g. using a phone which understands the owner emotion and provides corresponding recommendations).

Usually expressing emotion is possible through speech, text, video, etc. It is an important point to state that emotions can vary from culture to culture and so, in particular models that learned on in English does not mean they perform well also on Spanish. As a consequence it is possible that deciding factors for emotions can be different from language to language. In the literature emotion expression activities are known as modalities. If the model is recognizing emotion based on the only one activity it is called uni-modal and if it is based on more than one it is called multi-modal. Logically, uni-modal models should not have high accuracy as they have less information (e.g. only facial expression, audio, etc.). In fact, as [22] describes in the introduction for the overall emotion semantic content contributes only by 7%, while the semantic content and facial expression contribute 38 and 55% respectively. On the one hand, facial expression and gestures are important for emotion expression as new born child can only see such things and mainly get impression from these modalities, however, they are not enough for getting a state of the art models. That is why in recent year researchers are using data fusion strategies in order to get high performance as the authors have done in [23]. In data fusion strategies the main point is to extract important features for emotion from a different modalities and using combination of them to recognize emotions.

As recently Machine Learning (ML) and Deep Neural Networks (DNN) are growing in a constant high rate and having a good results of them in a variety of fields, it is reasonable to try to ML and DNN in Emotion Recognition. During last two decades a lot of research have been done in the direction [1, 7, 8, 11, 15, 16, 19, 21, 23]. Most of mentioned researches using some kind of features from an audio and/or videos and giving these as an input for Machine Learning or Deep Learning models. With the time accuracy of the models are getting higher, however, there is no state of the art work that have been done yet in this field and here a lot of research is needed. As one expects, for the above mentioned methods proper datasets play a key role. As described in [22] already there are some datasets which is reasonable to use for the training of the new models.

Datasets that are used in this work are described in following chapter in more detailed. One of the problems is the lack of many speakers: as having a few speakers in the existing datasets will not guaranty generalization toward new speakers that the model have not seen. That is why in this work the test data is consists only from the speakers whom the models have not seen during a training time. If the performance of a model is good for this test dataset it is more logical to think that this model generalization error is less than the others which have been tasted only on data which is familiar to the model. This is one of the key difference between this work and the others mention above.

Following are the key developments that have been done during in this work.

1. Use hybrid method for modality fusion on the raw data (audio and pictures from video)
2. Train and test sets separation based on the speakers (this is important as models tend to overfit to speakers and so the generalization error will be high in this cases)
3. Use mixture of different datasets with augmentation of real world noise in order to provide robustness

## Chapter 2

# Existing Datasets and Methods

### 2.1 Datasets

This section is an overview of existing datasets for an audiovisual emotion recognition. As most of the datasets are in English and this work is based on English datasets here is only a description for English datasets. For more general class of datasets refer to [22]. Table 2.1 shows the datasets and brief information about them.

RAVDESS [10] is audiovisual dataset where number of actors are 24 in which 12 females and 12 males. It has only audio for audio emotion recognition. In the Audiovisual part of the dataset consists of speech and song videos. In the song part of the dataset, actors express emotions like they are singing and in the other part is just a causal speech. Each actor speaks two different sentences in eight different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised). SAVEE [17] is audiovisual dataset for four different male actors. Each actor speaks 15 different sentences with seven different emotions (anger, disgust, fear, happiness, sadness, surprise, neutral). IEMOCAP [2] is one of the important datasets that exist currently in a sense that it contains 12h video recordings with ten different emotion categories (happiness, anger, sadness, frustration, neutral, surprise, fear, disgust, excitement and other). Here, depends on the emotion the dataset can contain very little or very much recording from that part. For instance, other emotion is consists less than one percent and excitement is the most common emotion in the whole dataset. In IEMOCAP ten actors (5 female and 5 male) perform a dyadic (human-to-human) conversations with the above mentioned emotions. eNTERFACE'05 [12] 44 actor speak in six different emotions (anger, disgust, fear, happiness, sadness, surprise). Each actor speaks exactly once for each emotion. Here dataset is less than compared to IEMOCAP. SEMANIE [13] is 1166 conversations with 150 participants in 27 associated emotional categories. AFEW (Acted Facial Expressions in the Wild) [4] is a dataset extracted from movies where the intention has been that in each part faces should express some emotion. It contains 1426 sequences of 330 subjects movie parts. The emotions for this dataset are

seven (anger, disgust, fear, happiness, neutral, sadness, surprise).

Here are missing some other datasets which are either not in English or could not been used from this work. The dataset that have been in this work are from these 2.1 datasets and information about them is in the next section.

Dataset	number of speakers	number of samples	emotion info
RAVDESS [10]	12 female speakers 12 male speakers	2880 utterance	Eight emotion categories: neutral, calm, happy, sad, angry, fearful, disgust, surprised
SAVEE [17]	4 male actors	480 utterance	Seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, neutral
IEMOCAP [2]	5 females 5 males	12h	Ten emotion categories: happiness, anger, sadness, frustration, neutral, surprise, fear, disgust, excitement and other
eNTERFACE '05 [12]	8 females 34 males	1166 video sequences	Six emotion categories: anger, disgust, fear, happiness, sadness, surprise
SEMAINE [13]	150 participants	959 conversations	27 associated categories
AFEW [4]	330 subjects	1426 sequences	Seven emotion categories: anger, disgust, fear, happiness, neutral, sadness, surprise

Table 2.1: Audiovisual datasets

## 2.2 Methods

Based on the existing methods for emotion recognition methods are divided into two groups. First one is Machine Learning where the researchers extract some important (important for emotion recognition) features from the audio or video and giving it as an input for Machine Learning algorithms (e.g. Support Vector Machines, XGBoost, Random Forest, etc.) and getting some results on the existing data [7, 16, 19]. Another group is Deep Learning where the input features can be as raw data as some extracted features from the raw data [1, 8, 9, 21, 23]. Next is the discussion of features extractions and above mentioned methods in more details.

### 2.2.1 Features

Feature selection is vital for Emotion Recognition as some of them can be irrelevant for expressing emotions and some of them highly relevant for specific emotion categorization. So, having a right set of features can dramatically improve model accuracy [22]. Following is described Audio and Video

features separately.

### Audio Features

Here are described some basic features of audio signal.

- ZRC: Zero Crossing Rate of audio. It quantifies the rate in which signal changes its sign (from positive to zero to negative or from negative to zero to positive). Formally it is defined as follows:

$$zrc = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{1}(s_{t-1}s_t)$$

where  $s$  is a signal of length  $T$  and  $\mathbb{1}$  is the indicator function.

- RMS: Root Mean Square of audio. It measures the frame energy for signal. RMS is somehow a mean value for digitized signal and can be calculated as follows:

$$rms = \sqrt{\frac{1}{T} \sum_{t=0}^{T-1} s_t^2}$$

where  $s$  is a signal of length  $T$ .

- HNR: Harmonics to Noise Ratio of audio. It measures the ratio between periodic and non-periodic components of a audio. HNR can be calculated as follows:

$$hnr = \frac{R_{xx}[T_0]}{R_{xx}[0] - R_{xx}[T_0]}$$

where  $T_0$  is assumed periodicity and  $R_{xx}$  is the auto-correlation function (ACF) and estimated as,

$$R_{xx}[l] = \frac{1}{N} \sum_{t=l}^{T-1} s_t s_{t-1}$$



for the biased version, and as

$$R_{xx}[l] = \frac{1}{N-l} \sum_{t=l}^{T-1} s_t s_{t-1}$$

for the unbiased version.

- Mel Spectrogram: this is a Mel scale of a Spectrogram which measures the spectrum of the frequency of a audio signal over time. The transformation from the Hertz scale to Mel scale is the following:

$$m = 1127 \log\left(1 + \frac{f}{100}\right),$$

where the  $\log$  is in natural base.

Spectrogram is computed as Fast Fourier Transform (FFT) of overlapping windows in a audio signal.

- MFCC: Mel-frequency cepstral coefficients. They are the coefficients of the MFC (Mel-frequency cepstrum) which is a short-term power spectrum of a audio, based on linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. MFCC can be calculated as follows:
  1. Fourier transform of a audio signal.
  2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
  3. Logs of the powers at each of the mel frequencies.
  4. Discrete cosine transform of the list of mel log powers, as if it were a signal.
  5. The MFCC are the amplitudes of the resulting spectrum.
- Spectral Centroid: This is a measure to characterize a spectrum. Spectral Centroid indicates where the center of mass of the spectrum is located.
- Chromagram: This is a the projected spectrum onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

## Video Features

The most relevant features for Emotion Recognition (ER) in a video are body movements and faces expressions. However, the latter one should be more important for ER. The most common approach


Extracted facial feature points (FFPs)	Facial regions	FAPs Num.	Euclidean distance between FFPs	Comparing FFPs displacement with neutral frame
	Eyebrows	1, 2	$D_{vertical,1}(22, 30), D_{vertical,2}(16, 35)$	$D_{v,1\_Neutral}-D_{v,1}, D_{v,2\_Neutral}-D_{v,2}$
		3, 4	$D_{vertical,3}(25, 30), D_{vertical,4}(19, 35)$	$D_{v,3\_Neutral}-D_{v,3}, D_{v,4\_Neutral}-D_{v,4}$
		5, 6	$D_{vertical,5}(22, 28), D_{vertical,6}(16, 33)$	$D_{v,5\_Neutral}-D_{v,5}, D_{v,6\_Neutral}-D_{v,6}$
		7, 8	$D_{vertical,7}(23, 28), D_{vertical,8}(17, 33)$	$D_{v,7\_Neutral}-D_{v,7}, D_{v,8\_Neutral}-D_{v,8}$
		9, 10	$D_{vertical,9}(25, 28), D_{vertical,10}(19, 33)$	$D_{v,9\_Neutral}-D_{v,9}, D_{v,10\_Neutral}-D_{v,10}$
		11, 12	$D_{vertical,11}(23, 30), D_{vertical,12}(17, 35)$	$D_{v,11\_Neutral}-D_{v,11}, D_{v,12\_Neutral}-D_{v,12}$
		13	$D_{m,13}(19, 25)$	$D_{h,13\_Neutral}-D_{h,13}$
	Eyes	14, 15	$D_{vertical,14}(29, 31), D_{vertical,15}(34, 36)$	$D_{v,14\_Neutral}-D_{v,14}, D_{v,15\_Neutral}-D_{v,15}$
		16, 17	$D_{vertical,16}(28, 49), D_{vertical,17}(33, 55)$	$D_{v,16\_Neutral}-D_{v,16}, D_{v,17\_Neutral}-D_{v,17}$
		18, 19	$D_{horizontal,18}(28, 30), D_{horizontal,19}(33, 35)$	$D_{h,18\_Neutral}-D_{h,18}, D_{h,19\_Neutral}-D_{h,19}$
	Nose	20, 21	$D_{vertical,20}(52, 68), D_{vertical,21}(58, 68)$	$D_{v,20\_Neutral}-D_{v,20}, D_{v,21\_Neutral}-D_{v,21}$
		22, 23	$D_{vertical,22}(49, 68), D_{vertical,23}(55, 68)$	$D_{v,22\_Neutral}-D_{v,22}, D_{v,23\_Neutral}-D_{v,23}$
	Mouth	24, 25	$D_{vertical,24}(52, 58), D_{horizontal,25}(49, 55)$	$D_{v,24\_Neutral}-D_{v,24}, D_{h,25\_Neutral}-D_{h,25}$
	Facial	26, 27	$D_{horizontal,26}(5, 58), D_{horizontal,27}(11, 58)$	$D_{h,26\_Neutral}-D_{h,26}, D_{h,27\_Neutral}-D_{h,27}$
	Contours	28, 29	$D_{horizontal,28}(2, 68), D_{horizontal,29}(14, 68)$	$D_{h,28\_Neutral}-D_{h,28}, D_{h,29\_Neutral}-D_{h,29}$
		30	$D_{vertical,30}(8, 68)$	$D_{v,30\_Neutral}-D_{v,30}$

Figure 2.1: The example of 68 facial feature points extracted using AAM alignment and related facial animation parameters.

is to extract faces from the video then using directly the faces or some other important features from the faces. In Figure 2.1 (taken from [22]) is shown 68 facial feature points (FFPs) from five facial regions (eyebrows, eyes, nose, mouth, and facial contours). As different people have different facial animation parameters (FAPs), researchers do further normalization. For each face FAP's expressed by FAP units which is the fraction of key distance in the face. From these points it is possible to extract various types of features which will be the video features. For instance, outer raised eyebrow FAP's is calculated as the distance between 22, 28 and 16, 33 FFP's. Another feature extraction method is to take the whole pictures of faces and force a model to learn emotion form the entire face.

### 2.2.2 Bi-modal Approaches

For the case of uni-modal Emotion Recognition there is only one type of features (only audio, only video, only text, etc.) which will be the input of any model. However, in multi-modal cases careful fusion of data is important in order to get better accuracy. Based on the existing methods it is possible to divide these approaches into three groups: feature level fusion, decision level fusion and hybrid fusion [1, 5, 7, 8, 11, 14–16, 19–21, 23]. Each of them are described in the following:

1. Feature level fusion: Here fusion strategy is that features are merged in a early level of the model Figure 2.2 and then based on that model start to learn how to recognize emotions [5, 14, 20]. Here are some problems as features are from different modalities can be incompatible for a model and at the end model decision could be based only on one modality which will not

improve the accuracy of the model.

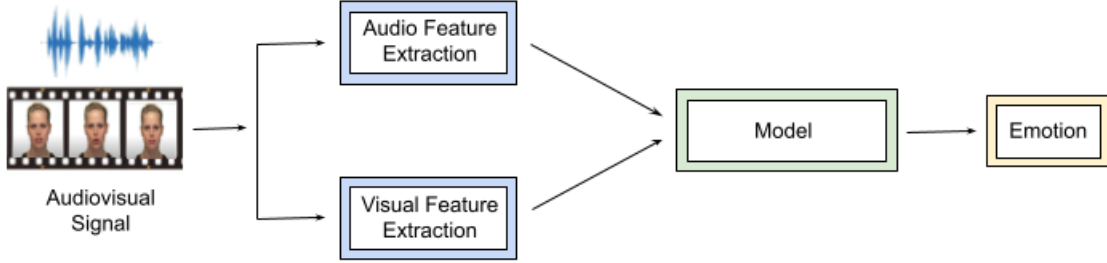


Figure 2.2: Feature level fusion.

2. Decision level fusion: Here fusion strategy is totally separates features and actually whole model is based on two sub-models Figure 2.3 and fusion is hold at the last stage [7, 16]. Having two networks for two different modalities each of them can predict emotions with some accuracy. Combining their predictions can improve the accuracy in some level. Image on of them predicts happy class more precisely and the other sad class. Final result can be taken happy if the first model predicts happy and sad if the latter predicts sad which defiantly will improve existing models. However, there are other decision-level fusions as some combination of two predictions, majority vote, etc. In these approaches actually models learn separately and the best is that when taking the best result from each of them, which means they do not use the whole existing content of the data.

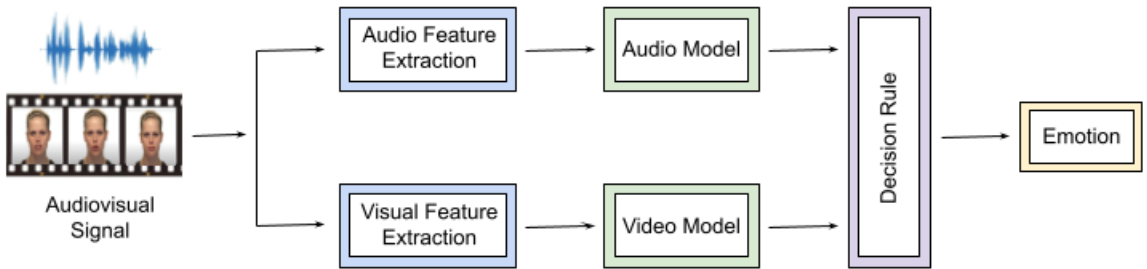


Figure 2.3: Decision level fusion.

3. Hybrid fusion: As it is clear form the previous approaches there is trade-off around the fusion of the data and the hybrid method is the combination of them [1, 8, 11, 15, 19, 21, 23]. The idea is that at the beginning two sub models extract important features from different modalities

and then combine them at some stage then merged model learn based on that Figure 2.4. This method could combine best sides of two previously described ones as in the first part models can learn features from videos and audio which will be appropriate to combine and last part of the model will learn based on these features.

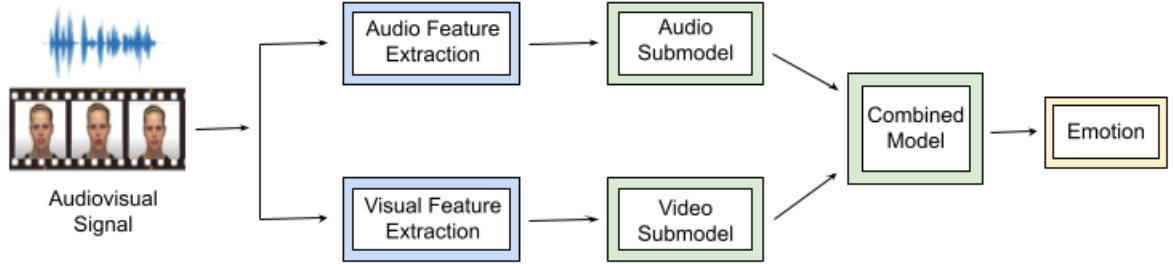


Figure 2.4: Hybrid fusion.

### some models

Chennoor et. al., [16] used Machine Learning methods (SVM, KNN) to predict emotions from audiovisual signals. Their methods uses decision level fusion in the following way: if predicted emotions for the audio and video models differ more than fixed threshold then video emotion is the final prediction, otherwise final prediction is audio model emotion.

In the [21] the authors used Deep Neural Networks to predict four and three kinds of emotions, that are: happy, angry, sad, neutral in the IEMOCAP [2] dataset. They used Spectrograms as feature from audio and 20 extracted faces from a video (each video is 3 seconds long or less). Model is a hybrid type consists of CNN for pictures, CNN+RNN for audio and fully connected for last stage.

Zhou et. al., [23] used AlexNet and ResNet for extracting features for audio log-mel spectrogram and video pictures coresspondingly and then used Hybrid approach. The use of AlexNet and ResNet can be dramatically improve the accuracy of a model as these networks can extract important features from the raw (pictures, Spectrograms) data, however, it has a additional layer inside model and clearly will increase prediction time.

## Chapter 3

# Proposed Method

Based on the existing methods the one which has the power to exploit the whole content of the multi-modal data is hybrid approach. In this case the method may be able to use the proper mixture of the semi-learned features and improve the emotion recognition accuracy. In order to understand whether or not the proposed methods have been able to exploit multi-modal features, the results are compared with the baseline results [21] and also best audio model accuracy is compared with the best video model accuracy. In the latter case if the video model accuracy is better than audio model accuracy it means hybrid model have been able to use features from both modalities effectively and if the accuracy is better than the baseline it means the method is also effective.

### Datasets

Datasets are chosen in way that they will be in English, contain as many speakers as possible and have been able to capture the real world examples. Being English is important as for the different languages features could vary and it could affect on the result dramatically and existing datasets are mainly in English. As explored in this work generalization of the emotion recognition algorithm is much more difficult in the case of new speakers. So, having more speaker is helpful to overcome this problem. The latter case is important as for the application in the world the algorithm should be able to recognize emotions from the real world cases. Based the above mentioned criteria the following datasets are chosen: Imeocap [2], eNTERFACE'05 [12] and RAVDESS [10]. Each video is divided into 3 seconds long sub-videos. Each sub-video is one input for the model. As the existing labels of the datasets are dramatically unbalanced only four labels (happy, angry, sad, neutral) have been chosen. They are the same labels as in [21] paper. At the end the whole dataset is about 12h long. For the sake of good generalization train and validation is divided by speakers. Speaker who are appearing in validation set does not appear in the train set at all. Around 20 percent of the whole dataset separated for the validation set. For the purpose of robustness of the models real

world noise is mixed with the dataset. After this data augmentation dataset is increased four times which is also one of the reasons that the models could have been able to overcome the problem of overfitting.

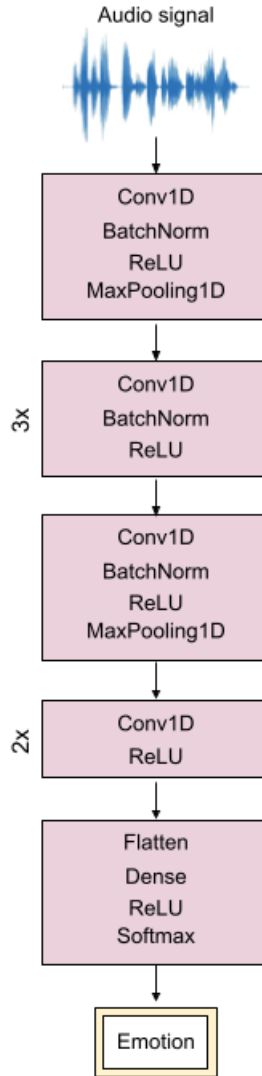


Figure 3.1: Audio model.

### Features

In the case of audio, ZRC, Chromogram, MFCC, RMS and Mel Spectrogram have been chosen. Most of the features is possible to get from Spectrogram, however, it is a raw feature and for the

models it is additional work to extract important features from Spectrogram. So, providing other features directly could potentially improve the model. At the end each audio feature is (160, 1) array.

In the case of video, from each video 20 images are taken and from each image only the faces are extracted using Dlib toolbox <sup>1</sup>. At the end for each video extracted feature is (50, 50, 20) image. The idea is to choose only faces from the whole image is that visual emotions are mostly expressing by faces and also part of the datasets missing the whole image of the person.

### Models

Throughout the work CNN, Lstm and their mixed networks have been explored.

For the audio case Figure 3.1 one of the CNN models that have been tried. 2x and 3x on the left of the blocks mean that there is 2 and 3 consecutive repetition from the same block respectively. Using a BatchNormalization used to overcome the overfitting problem partially. As in the case of new speaker learned model could not predict emotion in a high accuracy. MaxPooling layers help to reduce the number of weights and models become simpler. At the end of 8th block output have flattened and one dense layer added and then Softmax activation. Lstm network also tried for the only audio case, where the first layer is Lstm then few layers dense and ReLU blocks with Batchnormalization. Also networks with more weights have been tried which tend to overfit and could not generalize well.

For the video case CNN-CNN and Lstm-CNN networks have been tried. Until their concatenation audio and video parts learn separately and in some point they concatenated and then some block after it. Figure 3.2 shows a CNN-CNN model where the audio part is the same as only previously described audio network without Softmax replaced by ReLU. The video part of the network uses Conv2D layers with ReLU and Batchnormalization and some of them MaxPooling. At the end of 5th block Dropout is used in order to prevent from overfitting. After concatenation two layers was used for the hybrid level fusion and at the end Softmax activation.

---

<sup>1</sup><http://dlib.net/>

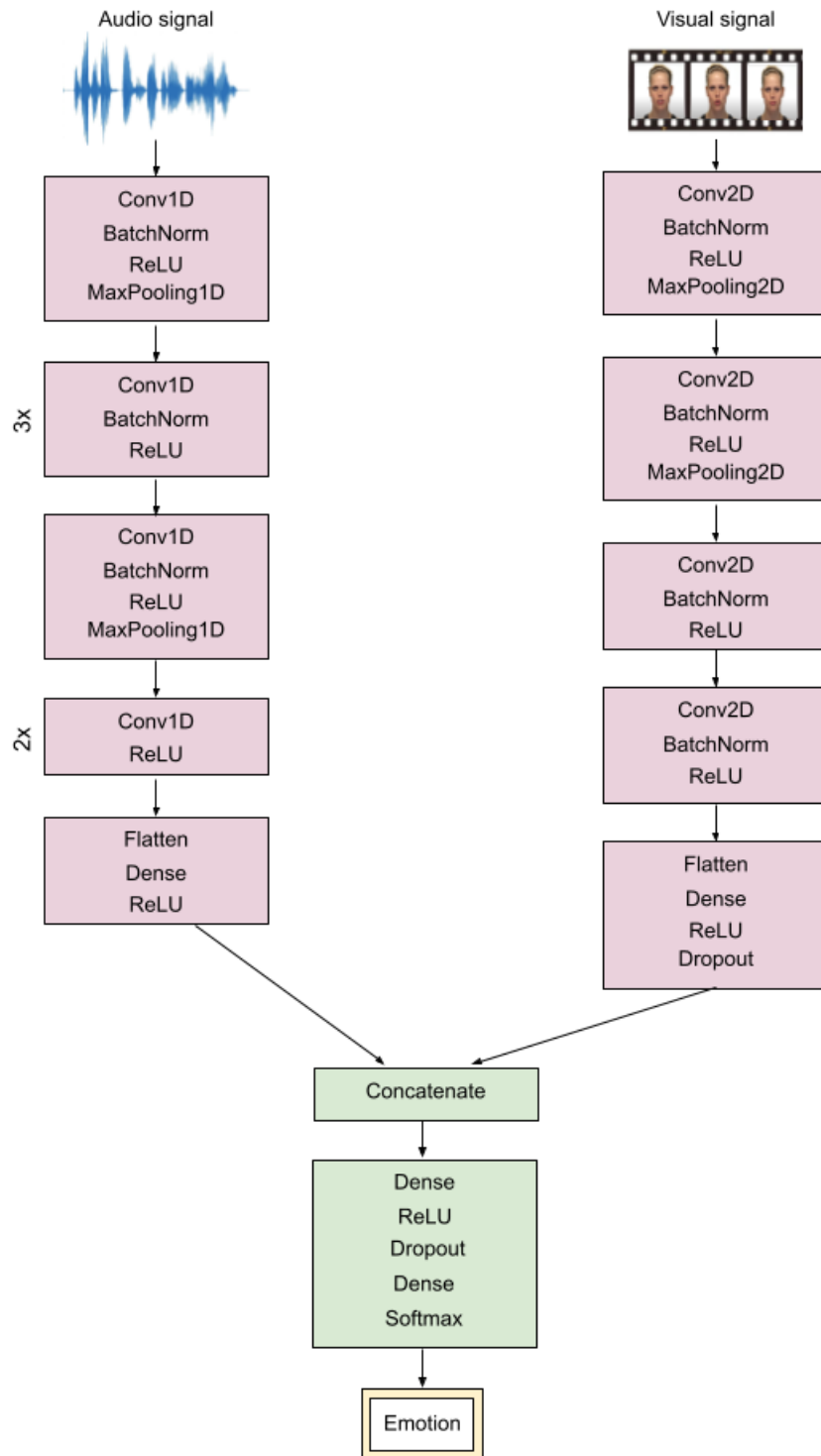


Figure 3.2: Audio model.



## Chapter 4

# Experiments and Results

Models have been trained using Adam, RMSProp and SGD optimization algorithms with different learning rates (1e-3, 1e-4, 1e-5). The result for the RMSProp and SGD is significantly worse compared to Adam. The best Audio model outperformed [21] best video model slightly having more than 54% accuracy Table 4.1. Important thing here is that the result is on the unseen speakers. In the case of random split of data models accuracy is more than 87% Table 4.2 and shows that described approach is much more powerful than existing ones.

Video models show better performance than audio models, however, depends on the model they have tendency to overfitting. Best Video model is CNN-CNN model more than 57% accuracy Table 4.2. The accuracy plot of the model is showed in the Figure 4.1.

models	Accuracy
Baseline	54
Audio model	54
Video model	<b>57.7</b>

Table 4.1: Results on speaker split.

models	Accuracy random.s.	Accuracy speaker s.
Baseline	54	-
Lightgbm Audio	<b>87.7</b>	<b>57.6</b>

Table 4.2: Results on random split.

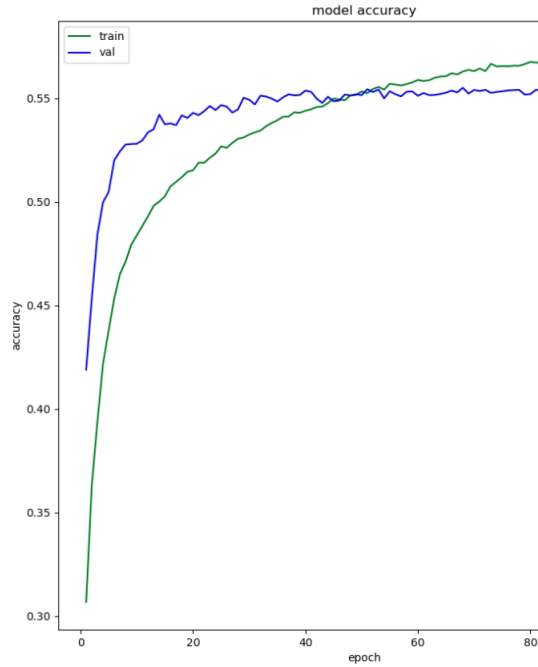


Figure 4.1: Video model accuracy.

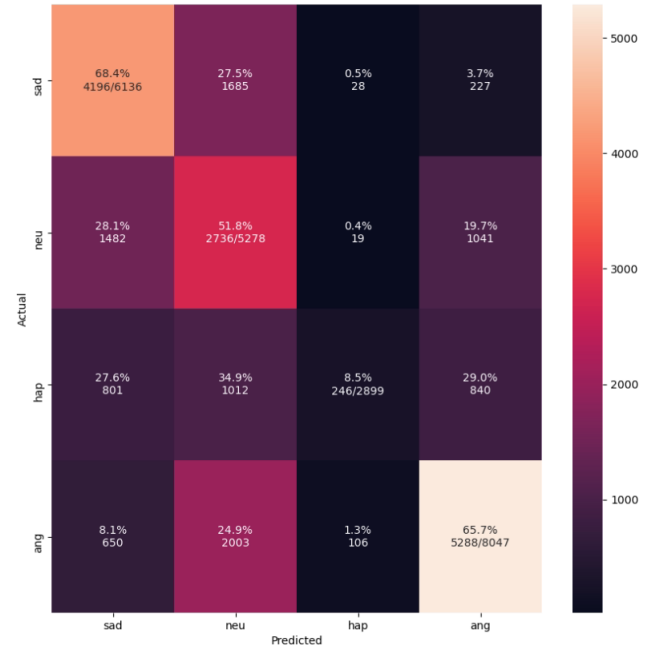


Figure 4.2: Confusion matrix.

All the models have trained maximum 100 epochs. In the cases when the train and validation accuracies differ more than 20 percent model training have been stopped for the sake of not overfitting. The codes are available in GitHub. <sup>1</sup>

<sup>1</sup><https://github.com/grigor97/multimodal.emotion.recognition>

## Chapter 5

# Conclusion

In this work new approach have been explored for emotion recognition. Existing methods mainly use the random split of the datasets and they are open to overfitting for the speakers and cannot generalize well. In the proposed method validation dataset is split by speakers and also compared with the random split of the dataset and experimentally showed that the latter one is much more easier to get high accuracy. Based on the experiments it turned out that using a hybrid level fusion help to use the existing content of the multiple modalities effectively by improving model accuracy. Moreover, using fusion of few datasets and similar to real world scenarios help to improve model performance and give an opportunity to use these kind of model in a real world applications. This work has been carried out in a short period of time for the master thesis and having more time would enable to explore existing methods and their performance in a better way. Moreover, being able to effectively use another modality (e.g., text) is promising research area in order to get state-of-the-art models.

# Bibliography

- [1] Avots, E., Sapiński, T., Bachmann, M., & Kamińska, D. (2018, jul). Audiovisual emotion recognition in wild. *Machine Vision and Applications*, 975–985. doi: <https://doi.org/10.1007/s00138-018-0960-9>
- [2] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., . . . Narayanan, S. (2008, 12). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42, 335–359. doi: 10.1007/s10579-008-9076-6
- [3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001, January). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18 (1)(1), 32–80. doi: 10.1109/79.911197
- [4] Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3), 34–41. doi: 10.1109/MMUL.2012.26
- [5] Eyben, F., Petridis, S., Schuller, B., & Pantic, M. (2012, March 25). Audiovisual vocal outburst classification in noisy conditions. In *Proceedings of ieee international conference on acoustics, speech and signal processing, icassp 2012* (pp. 5097–5100). United States: IEEE Computer Society. (10.1109/ICASSP.2012.6289067 ; null ; Conference date: 25-03-2012 Through 30-03-2012) doi: 10.1109/ICASSP.2012.6289067
- [6] Fragopanagos, N., & Taylor, J. (2005). Emotion recognition in human-computer interaction. *Neural Networks*, 18(4), 389–405. doi: <https://doi.org/10.1016/j.neunet.2005.03.006>
- [7] Gajsek, R., Štruc, V., Dobrisek, S., Žibert, J., Mihelic, F., & Pavesic, N. (2009, 09). Combining audio and video for detection of spontaneous emotions. *Biometric ID Management and Multimodal*, 114–121. doi: 10.1007/978-3-642-04391-8\_15
- [8] Gera, A., & Bhattacharya, A. (2014, 03). Emotion recognition from audio and visual data using f-score based fusion. *Proceedings of the 1st IKDD Conference on Data Sciences*, 1–10. doi: 10.1145/2567688.2567690

- [9] Jaiswal, M., Bara, C.-P., Luo, Y., Burzo, M., Mihalcea, R., & Provost, E. M. (2020, May). MuSE: a multimodal dataset of stressed emotion. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1499–1510). Marseille, France: European Language Resources Association. Retrieved from <https://www.aclweb.org/anthology/2020.lrec-1.187>
- [10] Livingstone, S. R., & Russo, F. A. (2018, 05). 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), 1-35. Retrieved from <https://doi.org/10.1371/journal.pone.0196391> doi: 10.1371/journal.pone.0196391
- [11] Ma, F., Zhang, W., Li, Y., Huang, S.-L., & Zhang, L. (2020). Learning better representations for audio-visual emotion recognition with common information. *Applied Sciences*, 10(20). Retrieved from <https://www.mdpi.com/2076-3417/10/20/7239> doi: 10.3390/app10207239
- [12] Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006, 02). The enterface’05 audio-visual emotion database. *Data Engineering Workshops, 2006. Proceedings*, 8 - 8. doi: 10.1109/ICDEW.2006.145
- [13] Mckeown, G., Valstar, M., Cowie, R., & Pantic, M. (2010, 07). The semaine corpus of emotionally coloured character interactions. *2010 IEEE International Conference on Multimedia and Expo, ICME 2010*, 1079-1084. doi: 10.1109/ICME.2010.5583006
- [14] Metallinou, A., Katsamanis, A., Wöllmer, M., Eyben, F., Schuller, B., & Narayanan, S. (2015). Context-sensitive learning for enhanced audiovisual emotion classification (extended abstract). In *2015 international conference on affective computing and intelligent interaction (acii)* (p. 463-469). doi: 10.1109/ACII.2015.7344611
- [15] Neumann, M., & Thang Vu, N. (2021, March). Investigations on Audiovisual Emotion Recognition in Noisy Conditions. *arXiv e-prints*, arXiv:2103.01894.
- [16] Nikhil Chennoor, S., Madhur, B. R. K., Ali, M., & Kishore Kumar, T. (2020, June). Human Emotion Detection from Audio and Video Signals. *arXiv e-prints*, arXiv:2006.11871.
- [17] Philip, J., & Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.
- [18] Picard, R. (1997). *Affective computing*. MIT Press.
- [19] Sahu, G. (2019, April). Multimodal Speech Emotion Recognition and Ambiguity Resolution. *arXiv e-prints*, arXiv:1904.06022.
- [20] Schuller, B., Valstar, M., Cowie, R., & Pantic, M. (2012). Avec 2012: The continuous audio/visual emotion challenge - an introduction. In *Proceedings of the 14th acm international conference on multimodal interaction* (p. 361–362). New York, NY, USA: Association

- for Computing Machinery. Retrieved from <https://doi.org/10.1145/2388676.2388758> doi: 10.1145/2388676.2388758
- [21] Singh, M., & Fang, Y. (2020, June). Emotion Recognition in Audio and Video Using Deep Neural Networks. *arXiv e-prints*, arXiv:2006.08129.
- [22] Wu, C.-H., Lin, J.-C., & Wei, W.-L. (2014). Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3, e12. doi: 10.1017/ATSIP.2014.11
- [23] Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K., & Qiao, Y. (2020, December). Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. *arXiv e-prints*, arXiv:2012.13912.