Laughter Detection using Mel-Spectrogram and 2D-CNN

Sarthak Garg
School of Electrical Engineering
Indian Institute of Technology Palakkad
Pudussery West, Palakkad, India, 678623
122101037@smail.iitpkd.ac.in

Dr. Sabarimalai Manikandan
School of Electrical Engineering
Indian Institute of Technology Palakkad
Pudussery West, Palakkad, India, 678623
msm@iitpkd.ac.in

***Abstract*—This paper explores an innovative approach to generate highlights from audiovisual content by using laughter as an identifying metric using Mel-spectrogram and a 2D Convolution Neural Network(CNN).By using the unique signal characteristics of laughter using mel-spectrogram and the spatial learning capabilities of a 2D CNN. The system accurately identifies instances of laughter in a given audio stream. These detected laughter instances can then be used to extract relevant segments to create a condensed output highlight.**

I. Introduction

A. Significance of the Project

Generation of highlight from multimedia content is very important for user participation and consumption efficiency. Laughter is a universal expression of amusement and enjoyment, while serving as a key indicator of content quality and emotional resonance. However, manual identification and extraction of laughter instances for highlight creation poses significant challenges, characterised by extensive labour and time-consuming efforts. Therefore there is a pressing need and value for algorithms or methods that can accurately detect laughter instances in any given audiovisual input. The applications for such a technology span across diverse fields, including social network platforms, entertainment media among others. Furthermore by detection the duration of laughter in content will be helpful especially in ranking humorous content especially those performed live or in front an audience.

B. Survey of Existing methods and technologies

Traditionally researchers in the speech recognition community have mainly focused on laughter detection which attempt to automatically track who spoke in a multi-person conversation [1]. Published approaches to this task use traditional MFCC and pitch-based features, with the most recent work employing fully-connected or convolutional neural networks on top of the MFCC's [2], [4]. These works treat laughter as a variable length event with performance metrics computes at from level.

Though in recent year there have been certain new papers that have used newer machine learning techniques to significantly improve the performance of laughter detection. The most significant out of these is the this paper on laughter detection in noisy environments [5], this approach recognises laughter in noisy and uncontrolled real-life scenes, while at the same time detecting variable-length event boundaries with fine granularity. This setting for the detection problem is more challenging but is important for real-world use cases. They use a ResNet model for classification which performs significantly better than the traditional approaches. The main focus of this paper was also comparing the two most widely available laughter datasets the Switchboard dataset and the AudioSet. We have use the newly annotated precisely labeled dataset that they have published as part of their paper as the dataset for our research.

A more in-depth discussion on existing technologies that are being used has been done in this paper [6]. This mainly focuses on exploring the model developed by Jon Gillick [5] and retrains the data on the ICSI corpus data to improve performance. There are also other papers that have explored audio recognition using Mel-spectrogram and 2D-CNN but none have been implemented to recognise laughter.

C. Problems and Statements

The primary challenge addressed by this research lies in developing a robust and efficient system for automated laughter detection. This involves developing an algorithm capable of effectively detecting laughter in the presence of noise and other speech signals while accounting for the variability in laughter expressions across different individuals. The problem statement proposed was to design algorithm that can accurately detect laughter in any environment which can then be further used for the process of highlights generation. The main idea was to use Mel-spectrogram to extract laughter features while also utilising a machine learning model in this case a 2D-CNN to be able to accurately detect laughter segments from any given audiovisual clip. Another thing to explore was the use of different windowing functions to test the effectiveness of each while generating the Mel-spectrogram.

D. Motivation

The main motivation behind this project was to explore the technique of using Mel-spectrogram along with a CNN and validate their usefulness in detecting specific features which for us was laughter. Furthermore we try to understand whether laughter detection using such an approach deems any valuable

model. Given the lack of models that are accurately able to detect laughter especially in complex unregulated environments we try to find a feasible method that will able us to solve this particular problem. Furthermore, given the numerous applications that this method could explore motivates the exploration of such novel methods and technologies in this domain.

E. Major Objectives with Work Plan

The primary objectives of this work include dataset acquisition, data preprocessing, model generation and GUI implementation. We start with finding a good dataset. After exploring a lot of dataset we arrive at the re-annotated AudioSet data from Jon Gillick's Paper [5]. We use this dataset along with using AudioSet to get non-laughter segments. After this the next step is using this data to generate the Mel-spectrogram that will be used to train the CNN. We use librosa for generating the Mel-spectrogram and audio processing. Next we use a Keras - tensorflow Sequential model to develop our algorithm. After this we develop a GUI to get highlights about laughter like number of instances and duration along with separating out the laughter and non-laughter clips. Intermediate step with regards to model training involve refining the algorithm based on feedback and evaluation to optimise accuracy and precision.

II. Materials and Methods

A. System Architecture with Description

Our dataset contains of two folders one with clipped laughter audio files and the other with a large set of non laughter clips. In our first step we load all the audio files and generate their Mel-spectrograms(mels) so as to give to the CNN later. Our first problem that arises is that given our troubles with finding a good dataset we use data augmentation to get a wider set to train on. We use overlapping blocks which allows us to expand our dataset while still keeping a SLD(strongly labelled dataset). Along with generating mels we also label the dataset simultaneously. After we get a balanced SLD we then use a 2D-CNN to classify the data into the two labels. We have defined the parameters as given. The parameters that are described were found to have the best model. We tried varying 3 parameters with respect to Mel-spectrogram — block-size(the length of audio segments that is taken),windowing function (windowing function applied to the signal before Mel-spectrogram), fft size. Along with this we also varied the model parameters to reach the best model possible given the training dataset.

B. System Specification Table

TABLE I
System Specification

Component	Specification
Sampling Rate	44.1 kHz
Processor	Apple M2
Memory	16 GB RAM
Storage	512 GB
Operating System	MacOS
Deep Learning Framework	Keras - Tensorflow

TABLE II
Model Architecture of 2D CNN

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 128, 43, 16)	160
max_pooling2d (MaxPooling2D)	(None, 64, 21, 16)	0
conv2d_1 (Conv2D)	(None, 64, 21, 32)	4,640
max_pooling2d_1 (MaxPooling2D)	(None, 32, 10, 32)	0
conv2d_2 (Conv2D)	(None, 32, 10, 64)	18,496
max_pooling2d_2 (MaxPooling2D)	(None, 16, 5, 64)	0
conv2d_3 (Conv2D)	(None, 16, 5, 128)	73,856
max_pooling2d_3 (MaxPooling2D)	(None, 8, 2, 128)	0
flatten (Flatten)	(None, 2048)	0
dense (Dense)	(None, 1024)	2,098,176
dense_1 (Dense)	(None, 512)	524,800
dense_2 (Dense)	(None, 2)	1,026

C. Description of Sensors or Other Modules Related to the Project

The main module we use for audio processing is librosa. We mainly use its feature.melspectrogram function followed by converting the power to db scale for better resolution. This function converts the raw waveform into a Mel-scaled power spectrogram, which provides a compact representation of the frequency content over time and using the db scale further enhances the relevance for human interpretation. These representation input act as the main input features to the sequential Convolutional Neural Network(CNN) model which is implemented using Tensorflow and Keras. Using annotated audiovisual data, the CNN model is trained to acquire discriminative characteristics for the detection of laughter. TensorFlow acts as the backend for effective model parameter computation and optimisation during training, while Keras offers a high-level interface for creating and training neural network models.

Further we also use Seaborn, Sklrean and matplotlib as libraries that help when evaluating the model by providing easy to use function that help us compute various performance metrics such as precision, recall and f1 score. Furthermore they also help us in plotting graphs and confusion matrix.

D. Block Diagram or Flowchart with Description

The sequential method for creating and assessing an audio-visual data-based laughter detection system is shown in this flowchart. The method starts with the input data and includes preprocessing stages including scaling power spectrograms and computing Mel-spectrograms. Next, training and testing sets of the data are created for the purpose of training and assessing the model. A Convolutional Neural Network (CNN) architecture is created, assembled, and trained using training data during the model training process. After training, the model's efficacy is evaluated using the testing data, and metrics like accuracy, precision, recall, and F1-score are calculated. Lastly, the model's performance is revealed by the output performance measures, which direct future optimization and refinement efforts.

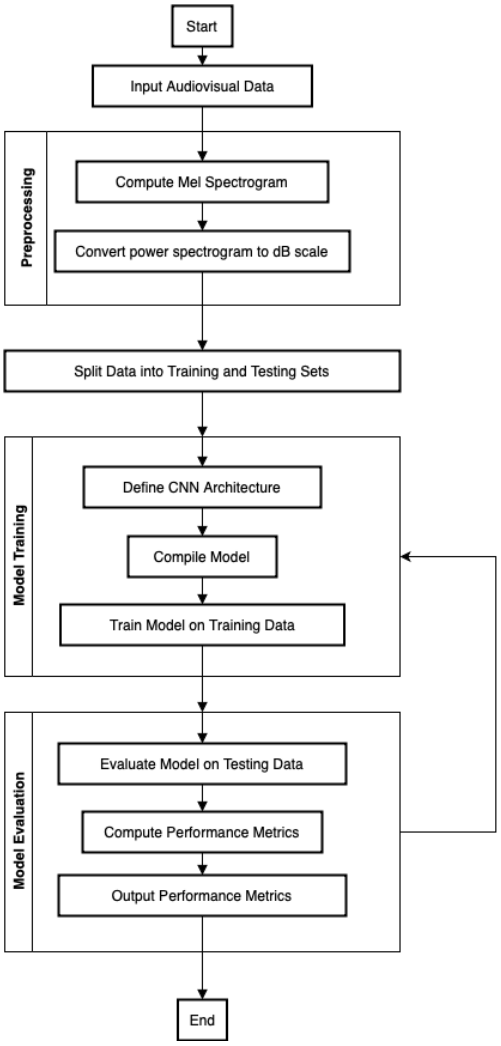

Fig. 1. Flowchart of workflow

E. Different Modules of the Proposed Methods

1) ***Data Pre-processing****:* This module converts unprocessed audio signals into Mel-spectrogram representations by using the librosa library. Mel-spectrograms give the frequency content of audio signals over time a condensed and perceptually relevant representation. This procedure entails breaking up the audio data into brief time intervals, giving each frame a Fourier transform, and then projecting the resulting spectrum onto the Mel scale.This module converts the Mel-spectrogram into a decibel (dB) scale as a next step in processing the spectrogram after it has been computed. By scaling the power values logarithmically, this conversion improves the spectrogram's perceptual relevance and makes it easier for training.

2) ***Model Training****:* The Convolutional Neural Network (CNN) architecture for laughter detection is designed and specified in this module. It specifies how many convolutional, pooling, and fully connected layers there are in the network as well as their arrangement.This module defines the CNN architecture, then specifies the loss function, optimiser, and evaluation metrics to compile the model. Model parameter optimisation and learning are influenced by the selection of optimiser (e.g., Adam) and loss function (e.g., binary cross-entropy).This module uses the preprocessed training data to train the CNN after the model has been compiled. The model learns to minimise the loss function by optimising its parameters and extracting discriminative features from the Mel-spectrogram representations during training.

3) ***Model Evaluation****:* This module evaluates the trained CNN model's performance using test data that hasn't been seen yet. In order to assess the model's accuracy in identifying instances of laughter, it entails feeding the testing data into the model and computing a number of performance metrics, including accuracy, precision, recall, and F1-score.This module computes performance metrics based on the model's predictions and ground truth labels after assessing the model on the testing data. These metrics offer valuable insights into the overall performance of the model, encompassing its ability to detect instances of laughter with accuracy, sensitivity, and specificity.The calculated performance metrics are finally shown in this module in an organised manner, like a table or report. These metrics facilitate the interpretation and evaluation of the model's performance by acting as quantitative markers of how well it detects laughter.

4) ***User interface****:* The user interface makes it easy for users to import video files and provides an option to automatically create clips when it detects instances of laughter. After uploading, users have the option to activate the system's ability to recognize and select the parts of the video that make people laugh, which expedites the process of creating brief highlights. The UI then presents users with important data, such as the total length of all the clips that were created and the total amount of time that included laughing in each of these sections. This feature gives users a thorough rundown of the length of the summarized content in addition to insights into how frequently laughter occurs throughout the highlights. This kind of functionality improves user engagement by facilitating quick decisions about what to watch and guaranteeing a more personalized and engaging experience.

F. Mathematical Expressions and Equations

1) Equations for Fast Fourier Transform (single-sided):

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-j \cdot \frac{2\pi}{N} nk} \tag{1}$$

$$MagX(k) = |X_k| \tag{2}$$

$$X_{norm}(k) = \frac{X(k)}{|X(k)|} \tag{3}$$

2) Hanning Window:

$$w(n) = 0.5 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) \tag{4}$$

3) ReLu activation function:

$$\sigma(x) = max(0, x) \tag{5}$$

G. User Interface Related to Project Tasks

The user interface is in short to make it easy to view the output of the model and also allows to create clips with the click of a button. The figure here neatly shows how the interface looks. Using tkinter in python itself it's a neat and functional design.

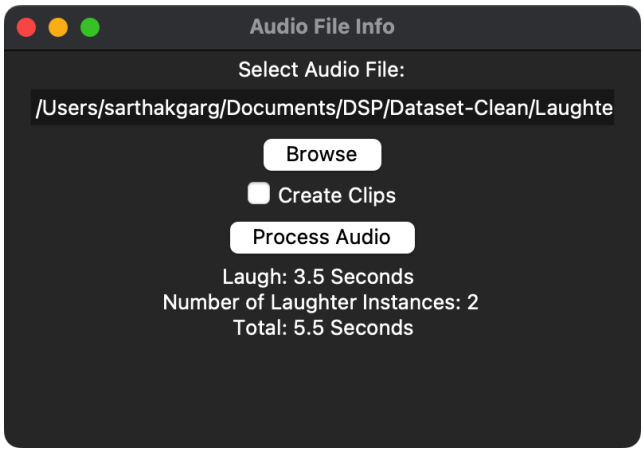

Fig. 2. GUI Interface

H. Performance Metrics

Performance metrics are essential for assessing how strong and successful the laughing detection system is. To evaluate different facets of the model's performance objectively, the system makes use of a variety of performance measures. F1-score, recall, accuracy, and precision are important performance indicators. The percentage of accurately categorized instances—both laughing and non-laughing—among all instances is known as accuracy. Precision measures the percentage of laughter instances properly predicted out of all the instances that are anticipated to be laughter, demonstrating the model's ability to reduce false positives. On the other hand, recall, also known as sensitivity, highlights the model's capacity to identify all pertinent events by calculating the percentage of correctly predicted laughter occasions among all actual laughter instances.A balanced assessment of the model's performance is given by the F1-score, which is the harmonic mean of precision and recall. This is especially useful in situations where the classes are unbalanced. Confusion matrix analysis also provides a thorough breakdown of false positives, false negatives, true positives, and true negatives, giving insights into the model's advantages and disadvantages in various classifications. Together, these performance indicators offer a thorough evaluation of the generalization, accuracy, and dependability of the laughing detection system, directing future improvements and optimizations to improve its effectiveness in practical applications.

III. Results and Discussions

A. Experimental Setup and Database Collection

First we find the dates to use for training. The main challenge faced was finding a suitable dataset to train our model over. Even though there are certain datasets available most of them are not comprehensive enough to be used for training. The two datasets freely available the Switchboard and AudioSet each face problems the former having a very controlled environment although a strongly labeled dataset while the latter is varied but is very weakly labelled. We use the dataset developed in the paper [5] which is a fine refinement of a small subset of the AudioSet data but is both varied and strongly labelled which means we can use data augmentation and use this for our training. After that we use "librosa" library to extract features and to generate the mel-spectrograms for this augmented audio.For training our model we use Keras via tensorflow for by utilising a sequential 2D CNN.

B. Model evaluation

TABLE III
SPECIFICATIONS OF THE 2D CNN MODEL

Property	2D CNN
File format	.mp3
Block length	22050
Sampling Rate	
Frames per second	44.1 kHz
Network Architecture	
Convolution Layers	4
No. of Classes	2
Dropout Layer	1
Other Layers	
Max Pooling Layer	4
Dense Layer	2
Hyperparameters	
Batch Size	5
Optimizer	Adam
Epochs	100
Learning Rate	0.0001
Model Performance	
No. of Parameters	2721154
Model Size (KB)	10380

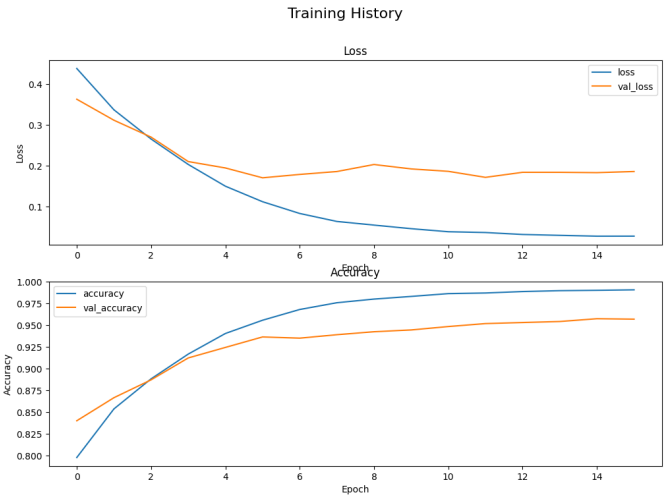

Fig. 3. Training history

TABLE IV
PRECISION, RECALL, F1-SCORE, AND SUPPORT OF THE CLASSIFICATION MODEL OF UNSEEN TESTING DATA

	Precision	Recall	F1-Score	Support
0	0.95	0.96	0.95	49801
1	0.90	0.89	0.90	22024
accuracy	-	-	0.93	71825
macro avg	0.92	0.92	0.92	71825
weighted avg	0.94	0.94	0.94	71825

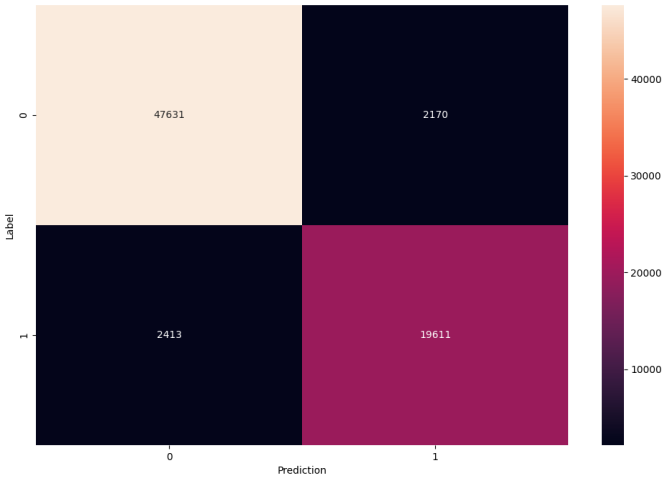

Fig. 4. Confusion Matrix for Unseen Testing Data

TABLE V
PERFORMANCE METRICS OF THE CLASSIFICATION MODEL ON ENTIRE DATA

	Precision	Recall	F1-Score	Support
0	0.84	1.00	0.91	166194
1	1.00	0.56	0.72	73222
accuracy	-	-	0.87	239416
macro avg	0.92	0.78	0.81	239416
weighted avg	0.89	0.87	0.85	239416

IV. CONCLUSION AND FUTURE WORKS

In this work, we introduced a thorough method for detecting laughter with 2D CNN models and Mel-spectrogram analysis. We proved, by means of thorough testing and analysis, that our suggested methodology is capable of precisely locating instances of laughter in audiovisual material. We were able to get strong performance in laughing detection across a range of datasets and circumstances by utilizing Mel-spectrogram representations and training deep learning models. Our findings open up new possibilities for applications in a variety of fields, including multimedia content analysis, social media monitoring, and human-computer interaction. They also highlight the potential of deep learning approaches to automate the process of extracting valuable insights from audiovisual data. All things considered, our work advances the state of the art in laughing detection and establishes the foundation for future investigation and advancement in this emerging subject.

This work opens up a number of new directions for future investigation and advancement. First off, investigating more complex deep learning architectures like attention mechanisms and recurrent neural networks (RNNs) may improve the model's capacity to extract contextual information and temporal relationships included in audiovisual data. Further research into multi-modal strategies that combine complimentary modalities, like text and image data, may enhance the resilience and generalizability of systems for detecting laughing. Deploying laughing detection systems in real-world scenarios would also need tackling issues with data imbalance, domain adaption, and real-time processing. Furthermore, there is potential for improving user experiences in multimedia applications by expanding the usage of laughing detection to tasks like sentiment analysis, emotion recognition, and content suggestion.

REFERENCES

[1] M. T. Knox, N. Morgan, and N. Mirghafori, "Getting the last laugh: Automatic laughter segmentation in meetings," in Ninth Annual Conference of the International Speech Communication Association, 2008.

[2] L.Kaushik,A.Sangwan,andJ.H.Hansen,"Laughterandfiller detection in naturalistic audio," in Proc. Interspeech. International Speech and Communication Association, 2015.

[3] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776–780.

[4] M.T.KnoxandN.Mirghafori,"Automaticlaughterdetectionusing neural networks," in Eighth Annual Conference of the International Speech Communication Association, 2007.

[5] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman, "Robust laughter detection in noisy environments". Proc. Interspeech 2021, pages 2481–2485, 2021.

[6] Lasse Wolter ,"A Machine Learning Model for Laughter Detection", University of Edinburgh, School of Informatics, Edinburgh, U.K., 2022. [Online]. Available: https://project-archive.inf.ed.ac.uk/ug4/20222999/ug_proj.pdf

[7] Dataset, https://drive.google.com/drive/folders/12Jd64OCnS4E9IblzHiUxBBT_AfGMbT1n?usp=sharing