Detection of Boxing Events from Acoustic Data

John Stephen.V
Computer And Communication
Panimalar Engineering College
(Anna University)
Chennai, India

Pothireddy Vidya Sagar Reddy
Computer And Communication
Panimalar Engineering College
(Anna University)
Chennai, India

Syed Tameem Basha
Computer And Communication
Panimalar Engineering College
(Anna University)
Chennai, India

Yogaram R A
Computer And Communication
Panimalar Engineering College
(Anna University)
Chennai, India

Srivarsan Kumar.J
Computer And Communication
Panimalar Engineering College
(Anna University)
Chennai, India

***Abstract*—** Boxing is a risky sport that incorporates punch and kick blows, with elite boxers capable of producing punches that travel at an average speed of about 25 mph. For contrast, 43 Ultimate Fighting Championship (UFC) events were staged abroad in 2021, with 509 fights having a maximum of 12 rounds, each lasting 3 minutes. Tracking the strikes taken in a fight can be difficult in boxing matches .The only component that is essentially constant and stable, making this detection feasible, is the sound component. In this study, we describe a system that examines sound data from a boxing match to identify events like punching and kicking. We begin by outlining the pipeline for processing audio and basic (atomic) event detection. After that, we go over the user interface, which provides a workbench for audio and model management as well as a device for data identification to quickly build the dataset accordingly. Our system's effectiveness is tested in studies using actual boxing punch and kick sound data and has an average accuracy and recall rate of around 82.1% p;.for atomic boxing matches. . Boxing players and coaches can locate and extract game highlights associated with specific actions using this method, which will allow them to study the highlights and enhance their performance.

Keywords—cloudant, architecture, acoustic, docker, label

I. INTRODUCTION

Numerous hours of boxing contests were taped from 2021 to 2022. Additional thousands of hours of boxing training are gathered in sports facilities run by associations like the world boxing association. The majority of these recordings simply provide high-level summary data like play outcome, arena, participants, and date instead of gameplay remarks. Only 60% of them possess the necessary information. Boxing players practice their forms and develop their play strategies for hundreds of hours prior to the competitions. Coaches commonly use the footage from tournaments and training games to improve their technique We propose a pipeline for boxing event recognition that comprises sound preprocessing to hunt for prospective events, feature extraction from noises, detection of basic atomic events, and the start and finish timings of a punch or kick, which is composed of a sequence of basic events. We describe the overall system architecture that successfully applies container technology to the cloud-based implementation of the boxing event detection service. The training of the models uses these labeled sound segments. In contrast to the majority of earlier studies, which assume that the training dataset is already supplied to the system in advance, we draw attention to the fact that our work in this research analyses the whole life-cycle of the system, where initially the system has no labelled data.

2.AUTOMATIC VIDEO ANNOTATIONS

Automatic interpretation of videos is becoming more and more necessary as the amount of video content increases. In the world of sports, Videos that included multisensory characteristics including performer motion, commentator speaking, and crowd noise were scored highly in Merler et al video's meta tagging and excitement ranking. To determine the beginning and conclusion of highlights, the approach involves streaming professional boxing films using a variety of deep learning algorithms. However, as indicated in Section 1, These video processing-based techniques frequently have very high levels of computing difficulty, which might not adapt well to situations with a lot of input that are video.

2.1 Acoustic Recognition

For a variety of application scenarios, the literature has paid a lot of attention to the categorization of acoustic signals into distinct classes. The state of the art for sound recognition now involves use of several deep neural network designs. For instance, Oines et al. uses Connectionist Temporal Classification (CTC) criteria in conjunction with Long Short-Term Memory (LSTM), Feed forward Sequential Memory Network (FSMN), and a combination of LSTM and FSMN for acoustic modeling. They demonstrated that their hybrid model outperforms all three other models for auditory categorization. There were 69 teams competing in an acoustic event categorization Kaggle competition, and the winner team used a weighted

ensemble that makes use of several Convolution Neural Networks (CNNs). CNNs also played a significant role in the competition's identification of uncommon sound occurrences and categorization of acoustic events.

However, standard algorithms like GMM and Support Vector Machine (SVM) may also work effectively in some application domains when a more condensed and limited feature set is developed. Mel-Frequency Cepstral Coefficient (MFCC) has been used in some studies [6, 7] as an input for environmental sound recognition. Comparisons of several deep neural network and shallow model versions are given by Li et al. They discovered that deep neural networks perform better in their domain. However, Dai's research revealed that traditional models like GMM and SVM excel in specific classification tasks. None of the methods listed above or in this part are intended for categorizing boxing tournaments. In this study, we provide a method for identifying and categorizing boxing events.

3.BOXING EVENT DETECTION PIPELINE

In the combat sport of boxing, two competitors hit each other for a set period of time within a boxing ring while typically using protective gloves and other safety gear like hand wraps and mouth guards. Our objective is to use machine learning techniques to automatically recognize various boxing events based on the noises captured during the boxing game.

3.1Preprocessing

To classify the acoustic events from a continuous flow of sounds captured during a boxing fight, the sound must first be segmented into tiny sound frames and then subjected to a machine learning model. Using traditional acoustics, the sound is typically separated into equal-sized windows at the start of the sound recording, and the sound of interest is then located and isolated inside a single frame. The first stage of our proposed peak detection method is to estimate the sound amplitude with time from the raw waveform. A local maximum point that occurs when the slope of the amplitude changes from a positive to a negative value serves as the primary reference point for a sound categorization task. The sound is separated into one sound window starting from the center and extending for a total of one second on both sides of the middle when a peak is located. The generated sound window is then applied to the further processing described below.

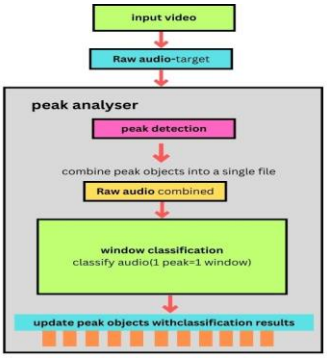

figure 1: Peak Sound Analyzer

Figure 1 depicts the complete peak detection process and how it relates to acoustic categorization. Within a search window of 10 seconds, peaks are sought; the size of this search window was established empirically. Additionally, the experimentally established window size of 1 second, which corresponds to the majority of boxing sound occurrences, was chosen. The next phase involves using the 1-second frames with likely sound occurrences which the peak detection system has provided to determine which event is present in every sound frame in the boxing match. Figure 2 depicts the whole processing pipeline for each 1-second sound frame, which consists of three primary components acoustic feature removal, acoustic classification, and fundamental evaluation.

3.2 Acoustic Feature Removal

It is necessary to extract feature representations from the original 1-second sound frames' actual waveform for machine learning algorithms to function with sounds.

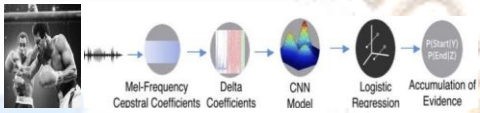

figure 2 Acoustic pipeline for boxing event detection

The acoustic signal is initially divided into time frames of 20 ms during the feature retrieval phase. The energy level inside each of the frequency areas is then calculated using Mel-Frequency Cepstral Coefficients (MFCC) characteristics with 25 bands on each 20 ms sound frame [8, 9]. We also incorporate the delta characteristics as a component of the sound attributes, which capture the variations of the acoustic signal , because a lot of the event data is embedded within the movement of the sound. The delta features are calculated as:

$$dt = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

where N is a window size to look both forward and backward , dt is the delta feature vector, and ct is the initial power spectral feature vector at time frame. To ensure that the vector encodes both static and dynamic aspects of the sound, they are added to the power spectral characteristics.

3.3 Point Boundary Detection

In a boxing competition, scoring by punching determines the winner. The judges give each boxer a score, and the boxer who demonstrates the most punches receives a bonus of 10 points. The opponent is likely to get nine points if there was no knockdown in the round, and eight if there was.so to detect the points easily this is used. Boxing coaches and players may assess tiredness, stroke efficacy, game strategy, and opponent vulnerabilities by using point detection, which is a very useful tool. Boxing Punching points are detected using characteristics extracted from the outputs of a basic event classifier and the confidence values (label probabilities) of several nearby sound windows.

Trends in sound classifier confidence were among the extra characteristics that were discovered through experimentation with the point boundary detection models detailed below. The characteristics employed to find the the point during a punch are listed in Table 1. We employ two logistic regression techniques for identifying the start and finish boundary locations, respectively, using these features as input.. The output of the logistic regression model is defined as $p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+.....+\beta_k x_k)}}$,

where the vector x = [x1, x2, ..., xk] denotes the input of the model (i.e., the features defined above), β1, β2, ..., βk are trainable parameters that represent the predictive power of each feature. We define that a hit sound serves as the trigger for the start and finish blow. Depending on whether the model is used to forecast the punch the output can be understood as the likelihood that the sound of a hit will define the point. The borders of various punching points may then be gathered by external systems for automatic content tagging and highlight window development.

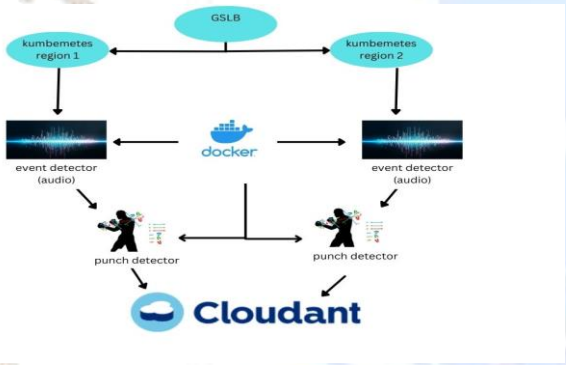

figure 3:Overall system architecture

Model	Features
Start	Sound duration, announcer average label probability, applause label count, duration to next label, duration to previous label, duration to previous two labels, duration between hit labels
End	Label probability, hit label average probability, non play label count, non play average label probability, next label probability, next two label probabilities, previous two label probabilities
Both	Announcer label count, applause average label probability, feet label count, feet average label probability, out label count, hit label count, out average label probability, previous label probability, current label probability, next label probability, previous two labels probability, next two labels probability

Table 1: The features used for point boundary detection

Our algorithm would need to classify over 1,000,000 sounds and find 100,000's of boxing points in order to evaluate all of the professional fights from the last 13 years of Boxing Grand Slams. Almost 900 hours of video and many player and data monitoring points from Hawkeye would need to be analysed in order to identify all point events connected boxing statistics. If we are to achieve the goal of a production system that is continuously accessible and extensible, each of the acoustic detecting components of our system must be built in a distributed design. This makes it simple to deploy and organise changes to our system's model, code, data, and configuration.

4.SYSTEM ARCHITECTURE

4.1 Containerization

A restful web service is provided for adding audio and video files to our system. It is spread and directed first to models for basic event detection .the application's results can be accessed through a restful interface and are stored in a flexible data storage system like cloudant; the application's docker copy can be put up as a service.

4.2 Boxing Event Detection as a Service

Kubernetes technology exposes our system as a scalable and distributed service in the cloud, as seen in Fig. 3. A Kubernetes image library receives the Docker image push. Every area on any cloud may have the picture dragged there and moved about. Our image is used to roll out any number of Kubernetes pods with the required number of workers to the cloud. Our system is open to the public for the classification of boxing events when we give our node port. Traffic to Kubernetes clusters may be distributed among many locations using a global server load balancer (GSLB). The traffic is subsequently sent to workers using Kubernetes. The outcomes of our system are stored in a shared data storage like Cloudant.

The Domain Name Service (DNS), which generates the request for our service, receives the sound files from a client. Every consumer can use the fundamental event and point boundary detection features. By upgrading the docker image, any modifications to models, settings, or code are propagated to the service.

5. User Interface

5.1 Data Labeling Tool

Before being used to analyse boxing sounds with unclear labels, the machine learning models outlined in Section 3 must first be taught on labelled boxing sounds. Watching (and listening to) lengthy boxing match records when events are only sporadically present would be too time-consuming for a person to do. We created an annotation tool, as shown in Fig. 4, to help with the human tagging process. The instrument utilises the peak detection method we developed, which is detailed in Section 3. 1. It takes in video and audio from a selected match and displays snippets that are centred on sound peaks to the viewer.

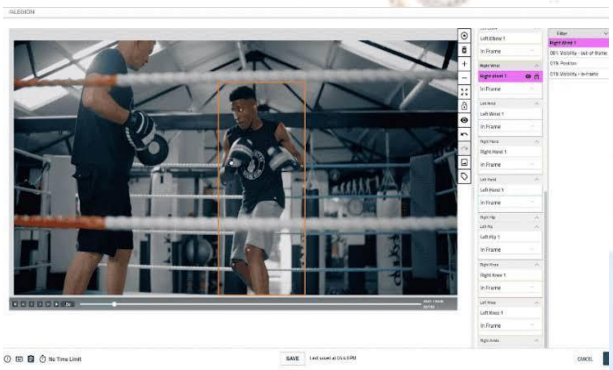

figure 4 Data labeling tool

5.2 Sound and Model Management Workbench

The system receives the annotated sound dataset created by the labeling tool mentioned in Section 5.1 and uses it to train models for fundamental event recognition and point boundary detection. New models may be trained using different purpose as the dataset expands (for instance, over time, recordings of additional boxing matches may be added). Moreover, sounds in the collection can already have incorrect labels that need to be changed. Also, noises that the user records while using the system could be fascinating event occurrences that the system was unaware of or labelled with uncertainty.

6. Experiments and Results

6.1 Setup

The study made use of videos from various boxing matches that were conducted on concrete floors in the same general acoustic location environment. This dataset offered a significant amount of sonic variation in the setting of a boxing bout on a specific surface. The ambient noise in the dataset was varied. The match footage and audio production quality likewise ranged from professional broadcast level to production using a remote-controlled robo camera. Each sound lasted one second, matching the typical duration of a boxing strike sound as described in Section 3.1. As labels, we used the set of fundamental events that are discussed in Section 3.3. The labels were tracked in a temporal sequence in a match using the match identifier and timestamp.

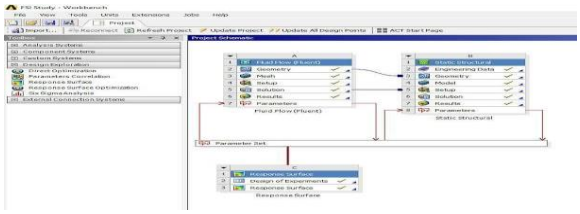

figure 5: Sound management workbench

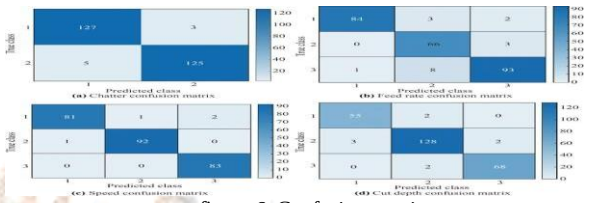

figure 6: Confusion matrix

Table 1: Distribution of acoustic classes

Class	Count	Duration(min:sec)
Announcer	3883	56:23
Applause	2842	47:22
Feet	2455	40:55
Hit	3596	59:56
NonPlay	3710	61:50
Out	3137	52:17

6.2Basic Event Classification Results

In order to evaluate fundamental event categorization, we tested our proposed method of CNN with MFCC-Delta-Acceleration features against a variety of models, such as GMM, closest neighbour, and SVM, as well as feature extraction techniques, such as log-mel energy regardless of delta/acceleration.. The MFCC log-mel energy feature is DCT transform minus the last step. Several of these shallow models that we utilise for this comparison were also employed in the related research that is detailed in Section 2. Eight Gaussians were utilised in the GMM model used for comparison. Hybrid modelling. A different GMM is educated with data matching to every label. The test phase probability of various labels is calculated using the probability values that each GMM provides, which is the standard method in audio categorization. in the vicinity. We used a Lp norm distance metric with p = 0.5 for the neighbour model. Regarding distance calculation, which isn't really a serious problem, strictly speaking because the triangle inequality does not hold with distance metrics this value of p, however our empirical data showed that it performed the best study. A linear kernel was used to define the SVM model.

6.3 Point Boundary Detection Results

Based on the output of our CNN +MFCC-Delta-Acceleration model, our suggested method outlined in Section 3.4 had an accuracy of 84% and a recall of 82% for predicting the start time of a point through punching. A somewhat lower accuracy of 77% and recall of 77% were recorded for the end time prediction. Several of the beginning or ending sounds of a point overlapped. For instance, a point can begin when a punch is delivered and

the audience applauds. Applause or background noise may occasionally mask the sound of a punch or a kick, though. The performance of detecting the point border was reduced by the constructive noises during the point event.

Model Type	Precision%	Recall%	F1%
CNN + MFCC	90.1	89.89	89.88
CNN + MFCC-Delta-Acc	92.47	92.40	92.39
GMM + MFCC-Delta	78.41	77.53	77.50
GMM + MFCC	76.65	75.76	75.87
GMM + Log-Mel	72.67	71.83	71.61
Nearest Neighbour + MFCC	72.59	68.53	68.57
Nearest Neighbour + Log-Mel	73.64	70.21	70.26
SVM + MFCC	37.19	20.12	23.94

Table 2: Basic event results of different models

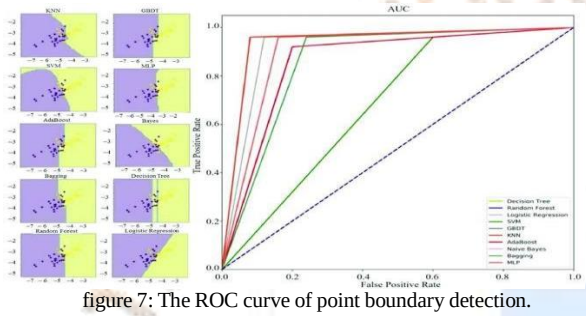

figure 7: The ROC curve of point boundary detection.

Intuitively, there is little variation in the time between a punch being given and being returned. When identifying the punches, the label confidences (probabilities) using the sound classifier were more accurate. The sound was more likely a marker if the confidences of the sound classifier's upcoming two labels and antecedent two labels were both over 50%. Fig. 7's Receiver Operator Characteristic (ROC) curve illustrates how well our punching point identification method performed.

7.CONCLUSION

In this research, we propose a method for extracting boxing events from audio data. It has a full pipeline for boxing event detection as well as facilities to facilitate data labelling and sound/model manipulation by users. The technology is offered as a cloud service that may cover a huge geographic region. The efficiency of our suggested solution has been demonstrated by experiment results. Our technology may be used to automatically identify boxing highlights with event markers. Moreover, the outcomes returned by our technology may provide extra meta-data and indexes into boxing footage to provide boxers more information and insights as they train and get ready for fights.

REFERENCES

[1] 2018. TUT Acoustic Scene Classification. (2018). https://www.kaggle.com/c/ acoustic-scene-2018

[2] 2019. Cloudant. (2019). https://www.ibm.com/cloud/cloudant

[3] 2019. Hawk-Eye Line-Calling System. (2019). https://www.topendsports.com/ sport/tennis/hawkeye.htm

[4] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. 2009. Environmental sound recognition with time–frequency audio features. IEEE Transactions on Audio, Speech, and Language Processing 17, 6 (2009), 1142–1158.

[5] Anthony Bagnall, Jason Lines, William Vickers, and Eamonn Keogh. 2018. The UEA & UCR time series classification repository. URL http://www. timeseriesclassification. com (2018).

[6] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. 2009. Environmental sound recognition with time–frequency audio features. IEEE Transactions on Audio, Speech, and Language Processing 17, 6 (2009), 1142–1158.

[7] Md Afzal Hossan, Sheeraz Memon, and Mark A Gregory. 2010. A novel approach for MFCC feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*. IEEE, 1–5.

[8] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy.2001. *Spoken language processing: A guide to theory, algorithm, and system development*. Vol. 1. Prentice hall PTR Upper Saddle River.

[9] James Lyons. 2015. Mel frequency cepstral coefficient (MFCC) tutorial. *Practical Cryptography* .

[10] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Deep learning for time series classification: a review. arXiv preprint arXiv:1809.04356 (2018).