International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Audio Event Identification and Classification for Cricket Sports using LSTM

Pooja N.K, John Sahaya Rani Alex.

Abstract*: Audio event identification is an emerging research topic to augment the automation of audio tagging, context-based audio event retrieval, audio surveillance and much more. In this research work, audio event classification for cricket commentary is done by using long short term memory (LSTM) neural network. Mel-frequency cepstral coefficients (MFCC) features are extracted from the audio commentary and trained with LSTM neural network. The trained LSTM network is validated and attained an accuracy of 95%.

***Keywords* : Acoustic Event, LSTM, MFCC,NN**

I. INTRODUCTION

Digitalization of audio data and video data created a humongous data. Searching for a particular event in the video or audio is an enormous task that involves a lot of man-hours. Besides, creation of highlights of the video content encompasses mammoth manual work. In recent years researchers have explored to automate this work with machine learning algorithms. In the machine learning domain, Classification task requires features from the data and they are trained with any one of the machine learning algorithms. An audio event identification/classification from the audio/video data involves dominant/valid features from the data. Literature shows that the efficacy of the features for audio event classification has been analyzed for Support Vector Machine (SVM) as classifier[1]. Bag-of-features principle which takes Mel and Gamma tone frequency cepstral coefficients that imitates psychoacoustic models are used for audio event identification [2]. The local binary pattern (LBP) from the logarithm of the Gamma tone like spectrogram is used for sound event classification [3]. But, this LPB feature is sensitive to noise. Another feature representation is arrived by sparse decomposition of sounds into atomic filter functions in the time-frequency domain [4]. The acquired sparse point pattern applied for identification of the given sound using graph matching algorithm. The classifiers used for event identification are SVM[5], neural networks[6] and deep neural. networks[7]. Deep Belief Network (DBN) with several hidden layers is applied for audio event classification in soccer videos [7] and the results are compared with SVM classifier. India is a country with a passion for cricket sports, this research aims to identify events from cricket commentary-audio. The first classification is identification of advertisement that comes in between the commentary. Next, we aimed to identify boundaries, sixer, and wicket. Section 2 explains how this work is implemented to achieve the aim. Section 3 talks about the experimental setup and results. The last section lays down the conclusion and future work.

Manuscript published on November 30, 2019.
* Correspondence Author
Pooja N.K* , School of Electronics Engineering, Vellore Institute of Technology, Chennai, India. Email: nkpooja13@gmail.com
John Sahaya Rani Alex*, School of Electronics Engineering, Vellore Institute of Technology, Chennai, India. Email: jsranialex@vit.ac.in

II. PROPOSED WORK

In the domain of supervised machine learning algorithms, the dominant representation of the data which is called feature has to be extracted from the raw data. These features are further taken as input to the classification algorithms. In this research, Mel-Frequency Cepstral coefficients (MFCC) are computed from the audio commentary because of its wide usage in speech recognition [8]. In literature, various neural networks have been adopted for classification. Here, we propose Long-Short Term Memory (LSTM) Neural Network (NN) for classification. The reason for choosing LSTM is, it works better for sequential data [9] and Cricket commentary audio also sequential in nature. The LSTM NN is trained with annotated MFCC features. Further, tested with unknown data. The block diagram of the proposed architecture is shown in fig .1.

A. Feature Extraction

In general, there are various methods such as MFCC, Perceptual Linear Predictive (PLP), wavelet cepstral coefficients[11] are employed for extracting features from the audio. In this work, the popular widely used MFCC method is adopted.

B. Mel Frequency Cepstral Coefficients

MFCC takes the perceptual nature of human ears. Using short frames to frame the video. Each frame is divided into a hamming window to preserve the signal accuracy. For through this discontinuity, a segmentation overlap is used. The time-domain structure is modified to the frequency domain by applying FFT on the container. The human vision spectrum in various bandwidths is fitted with a series of triangular filters passing the Mel frequency band. The spectrum is then applied to the function of the logarithm. The Mel filter banks all converge, which means the energies of the filter bank are directly related to each other.

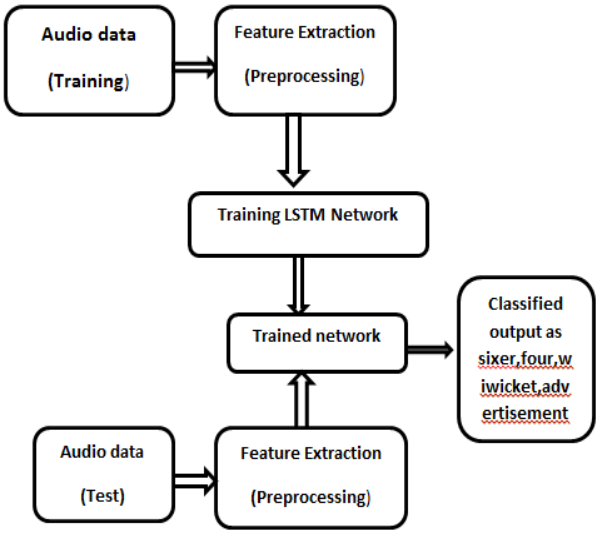

Fig. 1. Block diagram of audio event classification.

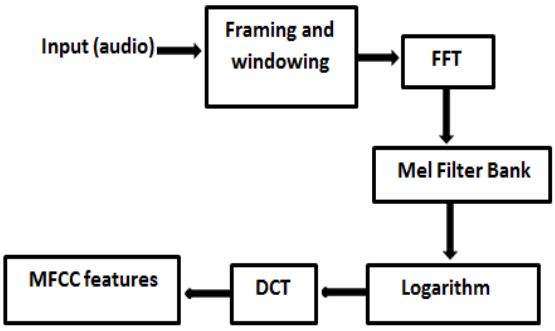

Fig.2. Feature extraction steps in MFCC

C. Audio Event Classification

We use the Long Short Term Memory (LSTM) neural network in our proposed system to train and classify the features into the tags. LSTM is a Recurrent Neural Network (RNN) with an added advantage of remembering the previous states. For this reason, LSTM is used for identification, organized data such as images, audio and video are highly efficient[9]. The training of LSTM is shown in fig 3.

1. LSTM network

LSTM is a bidirectional RNN. Deep RNN is used for general classification models. LSTM architecture consists of LSTM cells. An LSTM cell is shown in fig 4. The LSTM cell consists of regulators called gates, and they are input gate, forget gate and an output gate. The forget gate is implemented with a *sigmoid* activation function. A value of '1' to the *sigmoid* function remembers the state and a value of '0' will make the cell to forget the state. The *tanh* function makes the value between 1 to 1.

The simple function of LSTM is explained with the following equations 1 and 2.

$$c_t = \sigma(W_f [h_{t-1} * x_t] + b_f) * c_{t-1} + \sigma (W_i [h_{t-1} * x_t] + b_i) + \tanh(W_c [h_{t-1} * x_t] + b_c) \quad (1)$$

$$h_t = \sigma (W_o [h_{t-1} * x_t] + b_o) * \tanh(c_t) \quad (2)$$

where c_t is the cell vector at time t, h_t hidden vector of the cell at time t, x_t is the input vector at time t and σ is the sigmoid activation function, W_f, W_c, W_o are the weights of the forget cell, input cell, output cell respectively and b_f, b_c, b_o are the biases of the forget cell, input cell, output cell respectively. The proposed network consists of LSTM cells as much of input nodes and the output of each LSTM is given to the next LSTM unit with the previous state information.

The LSTM cells are connected to the fully connected network layer.

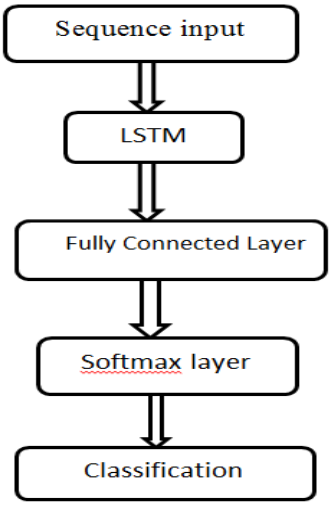

Fig. 3. LSTM network

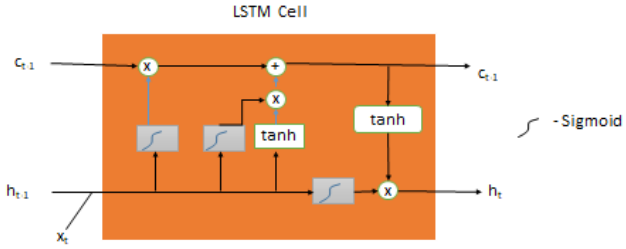

Fig.4 LSTM cell

2. Fully connected layer

The fully connected layer is a simple feed-forward neural network layer where all inputs from one layer are connected to every activation unit of next layer. It multiplies the input by a weight matrix and then adds a bias vector. It converts 2D array into 1D array i.e, it provides output as linear array.

3. Softmax layer

A Softmax layer allows the neural network to run a multiclass function and it will be able to classify as well as a probability in LSTM.

III. EXPERIMENTAL SETUP AND RESULTS

Datasets for this experiment are taken from a cricket commentary by recording. A long cricket commentary recording is fragmented into 2000 recordings of 9 seconds of duration and the time domain representation is shown in fig 5. These audio samples are manually annotated with tags for training purposes.

The audio data is split into training data and test data. Then the annotated audio data is given to MFCC to extract feature which is shown in figure 6. For a 9 second audio data 998 frames of 14 MFCC features per frame are generated and the number of samples taken for test and train is equal which is shown in Table I. The MFCC features are given as input to the LSTM network for training which is shown in fig. 8.

Initially, the network is trained with cricket commentary and advertisement audio. The accuracy achieved with the test data is 95%. Next, the network is trained for commentary audio and audio for four, the received accuracy is 70% and Finally, the network is trained for the multi-class problem. The classes are an advertisement, commentary, four, wicket and sixer, the obtained accuracy with the test data is 47.6%. and they are shown in Table II.

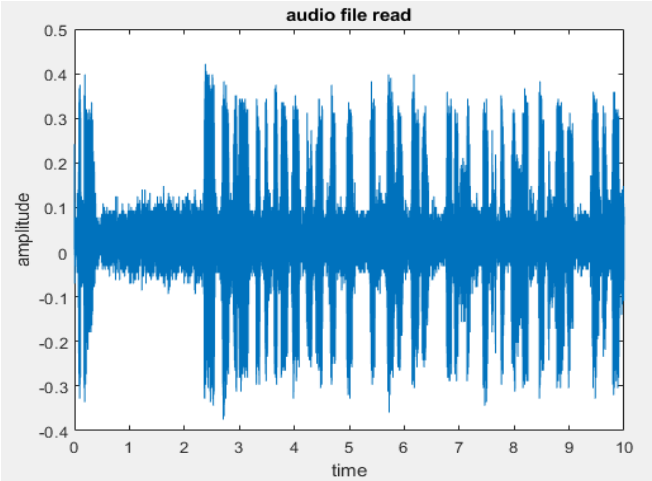

Fig. 5 Audio sample

Table- I: Number of samples used for train and test

Tags	Number of Audio files
Four	20
Sixer	14
Wicket	20
Advertisement	20
Commentary	20

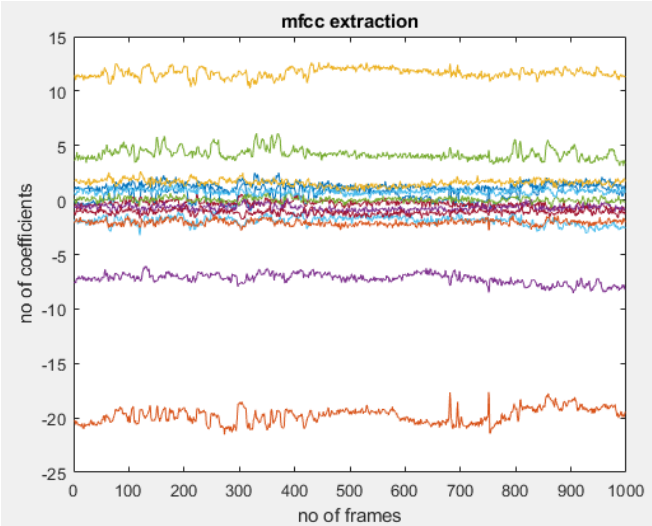

Fig .6. MFCC Features of the audio

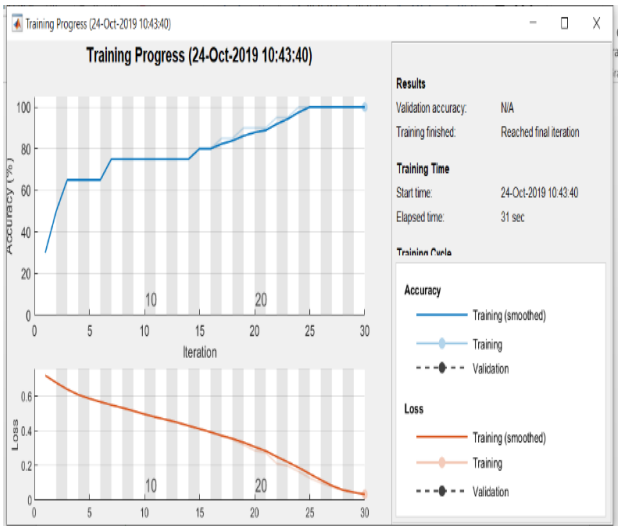

Fig. 8. Training result of the proposed LSTM network

Table-II: Proposed network accuracy with different classes

Classes	Accuracy
Commentary *vs* Advertisement	95%
Four vs Commentary	70%
All classes	47.6%

IV. CONCLUSION

This research work uses the benefits of the sequential nature of the LSTM architecture for audio event identification and classification in cricket sports. It is observed from the results that, we need more annotated data for multi-class identification. With the binary class problem as classifying the audio event as cricket commentary and advertisement obtained high accuracy of 95%. In future, the network will be trained with more data. Also a robust model will be built for automated annotation of audio data.

REFERENCES

1. Martín-Morató, I., Cobos, M., & Ferri, F. J. (2016, September). A case study on feature sensitivity for audio event classification using support vector machines. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.
2. Ren´e Grzeszick, Student Member, IEEE, Axel Plinge, Member, IEEE, and Gernot A. Fink, Senior Member, IEEE "Bag-of-Features Methods for Acoustic Event and Classification Detection" IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 25, NO. 6, JUNE 2017
3. Ren, J., Jiang, X., Yuan, J., & Magnenat-Thalmann, N. (2016). Sound-event classification using robust texture features for robot hearing. IEEE Transactions on Multimedia, 19(3), 447-458.
4. Adiloglu, K., Annies, A., Wahlen, E., Purwins, H., & Obermayer, K. (2012). A graphical representation and dissimilarity measure for basic everyday sound events. IEEE Transactions on Audio, Speech, and Language Processing, 20(5), 1542-1552.
5. D.A.Sadlier and N.E. O'Connor, "Event detection in field sports video using audio-visual features and support vector machine," IEEE Trans.on Circuits and systems for video technology,vol. 15,no,2005

6. K. Wickramaratna, M. Chen, S.-C.chen, and M.-L. Shyu, "Neural network based framework for goal event detection in soccer videos," in Proc. Of ISM,2005
7. Lamberto Ballen, Allesio Bazzica,Marco Bertini,Alberto Del Bimbo and Giuseppe Serra,Media integration and communication center,University of Florence,Italy. "Deep neural networks for audio event classification in soccer vedios".IEEE 2009.
8. Mukhedkar, A. S., & Alex, J. S. R. (2014, August). Robust feature extraction methods for speech recognition in noisy environments. In 2014 First International Conference on Networks & Soft Computing (ICNSC2014) (pp. 295-299). IEEE.
9. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.
10. John Sahaya Rani Alex, Nithya Venkatesan Robust optimal sub-band wavelet cepstral coefficient method for speech recognition. International Journal of Computer Aided Engineering and Technology,11(2), 163-173

AUTHORS PROFILE

Pooja N.K is currently pursuing MTech in Embedded Systems in VIT Chennai. She obtained a B.E from UBDTCE college of engineering in davanagere,Karnataka

Dr. John Sahaya Rani Alex is working as an associate professor in School of Electronics Engineering, VIT Chennai. Her research interests are Speech Processing, Signal Processing, IoT, implementation of DSP algorithms on an embedded board and ANN.

.