The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 4, No. 1. Online ISSN: 2831-4794

Detecting Referee Whistle Sounds in Soccer Videos Using Machine Learning Models

Tan Dat Trinh[*a], Nguyen Thi Minh Thi [a], Ngo Tuan Long [a], Ly Ai Ngoc [a], To Quoc Phong [a], Nguyen Ho Khanh Nhu [a]

[a]Department of Computer Science, Faculty of Information Technology, Saigon University, Ho Chi Minh City, Vietnam
[*]Corresponding author: trinhtandat@sgu.edu.vn

ABSTRACT

In this study, we propose a simple and efficient approach for detecting referee whistle sounds in soccer match footage. Our method combines a fixed-length sliding window with feature extraction based on the short-time fourier transform (STFT). Specifically, we use a 1-second sliding window with 50% overlap to divide the audio input into smaller segments, which are analyzed to identify whistles characterized by prominent frequency components in the range of 3300–4200 Hz. For classification, we employ well-known machine learning algorithms, including Naïve Bayes, Support Vector Machines, and Neural Networks. To improve system performance, we integrate a post-processing technique based on Power Spectral Density (PSD) and predefined thresholds to reduce misclassification. Experimental results demonstrate that our method is highly effective at detecting whistles during soccer matches.

Keywords: Whistle sound detection, short-time fourier transform, machine learning

1. INTRODUCTION

Sports videos such as baseball, soccer, basketball, etc., have attracted significant attention from audiences [1] . These videos typically consist of sequences of images and audio segments that convey the content and detailed information about the events taking place. In sports videos, audio streams often contain various sound events, including announcer commentary, fan cheering, referee whistles, crowd applause, and background noise. Among these, referee whistles, fan cheering, and applause are typical events that can be utilized to extract highlights from a sports video [1].

Referee whistles are among the most prominent audio events in sports, particularly in soccer. Referees use whistles to signal key moments in a match, such as the start or stoppage of play due to player injuries, the end of a half, the beginning of the second half, or a goal being scored. This highlights that a whistle typically corresponds to significant events in the game. Consequently, detecting whistle sounds during a match can facilitate the automatic extraction of highlights. However, whistle detection presents challenges, such as interference from commentator speech, fan cheering, music, and other background noises. Additionally, short highlight videos are often more popular, as they allow viewers to quickly catch the most exciting moments of the match. This underscores the importance of research into whistle sound extraction from soccer videos.

Automatic whistle detection falls within the domain of sound classification, with significant research efforts in recent decades focused on identifying referee whistles during sports events. These studies have employed various feature extraction techniques, including short-time energy [2], short-time fourier transform [3], power spectral density [3, 4], linear predictive coefficients [5] and mel-frequency cepstral coefficients (MFCCs) [5]. These features are typically combined with machine learning classifiers such as neural network [5, 6], support vector machine (SVM) [7], gaussian mixture model (GMM) [8, 9] and hidden markov model (HMM) [1, 10] for audio event detection.

In [1], the authors introduced a framework for extracting highlights in three sports—basketball, golf, and soccer through sound detection. The approach involves: recognizing background noise based on energy and intensity, classifying sounds using MPEG-7 audio features and Entropic Prior Hidden Markov Models. Finally, a post-processing method is applied to enhance the system's performance. In [2], the authors used a bandpass filter to enhance the sound of the whistle while reducing the influence of other audio events in sports videos. Subsequently, Short-Time Energy (STE) is utilized to extract features from audio events. Finally, the authors apply two thresholds based on amplitude and duration to identify audio segments containing whistle sounds. The method achieves results with a recall of 92.5% and a precision of 80.5%. In [4], the authors propose a system for summarizing sports videos, specifically soccer and swimming videos, based on

detecting whistle sounds, cheering/excitement from fans and commentators, and the appearance of text in the frame. For whistle detection, the authors suggested extracting power spectral density values within a predefined frequency range and comparing them to a fixed threshold to make decisions. The method achieves average recall and precision rates of 80% and 64%, respectively. Authors in [6] proposed using a convolutional neural network combined with MFCC features to recognize whistle sound events. The proposed method achieves an accuracy of 86%.

Authors in [8] developed a hierarchical framework specifically for sound event detection in tennis. The authors [9] proposed detecting highlights in rugby matches through audio analysis using a two-stage classification approach. In the first stage, incoming audio frames are classified as either "speech" or "non-speech" events. In the second stage, "speech" events are further classified as "excited speech" or "unexcited speech," while "non-speech" events are categorized as "whistles" or "others." Five GMM models are trained for classification, and a decision window is used to determine highlight onset and offset based on the percentage of "whistle" and "excited speech" events within the window exceeding a predefined threshold. SVM and HMM methods are proposed to detect audio events such as whistle sounds, fan noises, commentator speech, and the cheering/excitement sounds of both commentators and fans [10].

In this study, we aim to detect the sound of a referee's whistle in soccer match videos by analyzing audio data. We use a sliding window technique with a duration of 1 second and 50% overlap to segment the audio into smaller parts. Each segment is then processed to detect whistle sounds. We extract features from each audio segment using STFT and predefined frequency bands, as whistle sound energy is primarily concentrated in the 3300-4200 Hz range. To classify whether a segment contains a whistle sound, we employ well-known machine learning models such as Naïve Bayes, SVM, and Neural Network. To enhance system performance and reduce the impact of noise, we employ a post-processing method that involves comparing the power spectral density values with a predefined threshold

2. PROPOSED METHOD

In this section, we propose a feature extraction technique that utilizes the Short-Time Fourier Transform (STFT), focused on a specific frequency range, to detect the referee's whistle. By analyzing features derived from the input audio data, we apply machine learning algorithms such as SVM, Neural Networks, and Naïve Bayes to identify referee whistle sounds. Finally, we introduce a post-processing strategy using power spectral density (PSD) to enhance the overall performance of the system. Figure 1 illustrates the proposed model for detecting the sound of the referee's whistle.

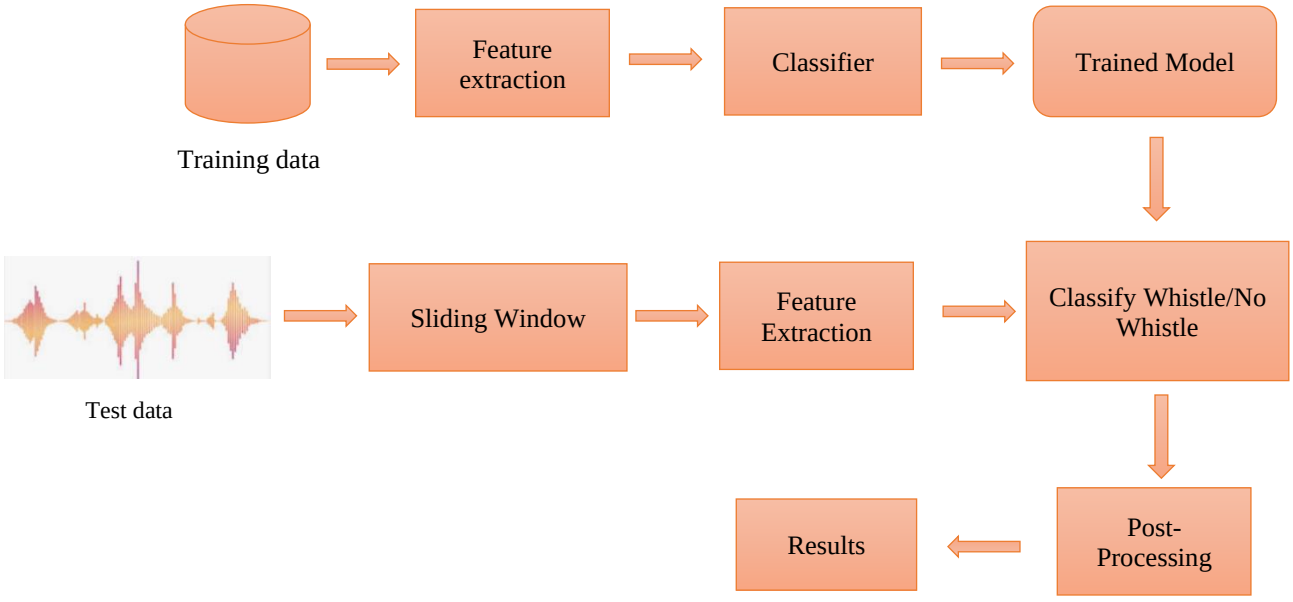


Figure 1. Our approach for detecting the sound of the referee's whistle in audio stream.

2.1 Feature extraction based on STFT

The short-time fourier transform (STFT) is a widely used method in the analysis of signals and the extraction of features from audio signals. It facilitates the conversion of signals from the time domain into the frequency domain [3]. Assuming

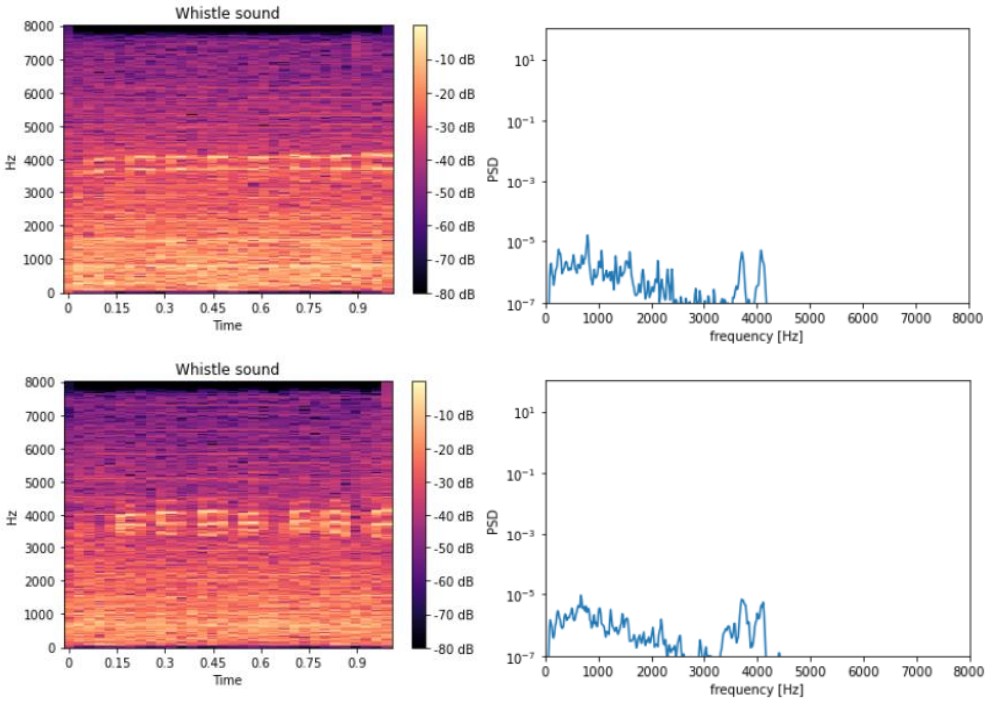
an input audio signal s(t), periodic processing of window frames can be processed through the windowing method ω(n), which is defined by the following formula:

$$s_l(n) = s(n + lM)\omega(n), 0 \leq n \leq N - 1 \quad (1)$$

where, the length of the window is denoted by N, the frame index is represented as l, the hop size is indicated by M, and the time index is referred to as n. The STFT subsequently transforms these window frames into the frequency domain through the application of Discrete Fourier Transform (DFT), with the STFT coefficients computed based on the following formula [3]:

$$S(l,k) = \sum_{n=0}^{N-1} s(n + lM)\omega(n)\, exp\left(-j\frac{2\pi}{N}nk\right), \quad (2)$$

where STFT coefficients are represented by $S(l,k)$, with DFT size indicated and k denoting the frequency bin index. In this study, we focus on deriving feature vectors utilizing STFT. We specifically implement feature extraction using a predetermined set of bin indices that reflect the properties of referee's whistle sounds. Through our experiments and observations, we determined that the predominant frequency information of whistle sounds is primarily found within the frequency range of [3300, 4200] Hz, where all sound samples were recorded with a 16-bit resolution and a sampling frequency of 16 kHz. Figure 2 illustrates the spectrogram and the PSD of audio signals containing whistle sounds. Figure 3 depicts the spectrogram and the PSD of audio signals without whistle sounds. From these figures, we observe that there are two prominent frequency bands of whistle sounds, primarily concentrated in the frequency range of 3300 to 4200 Hz



a) Spectogram b) Power spectral density

Figuer 2. The spectrogram and the PSD of audio signals containing whistle sounds

In our experiments, we used a Hanning window consisting of 1024 samples, performed a Fast Fourier Transform (FFT) with 1024 points, and utilized step size of 512. With a 1-second audio segment at a sampling frequency of 16,000 Hz, we were able to extract an STFT magnitude size of 513x32, where each spectral feature vector has a length of 513. The number of windows is based on the length of the audio, and for 1-second duration, we acquire 32 vectors, each with length 513. The formula for calculating the frequency bins is depended on the sampling frequency of the signal and the number of FFT points (num_fft), as follows:

$$frequency = \frac{bin_index \times sampling\ rate}{num_fft}, \quad (3)$$

The whistle sound exhibits a strong frequency concentration within the range of [3300, 4200] Hz, which encompasses 57 frequency blocks derived from [212-269] bins. To minimize noise and eliminate redundant data, we focus solely on this specific range rather than the full [0, 8000] Hz spectrum. Then, z-score normalization is applied to adjust the feature values, resulting in a distribution with a mean of 0 and a standard deviation of 1. This process effectively reduces the feature dimensions from 513x32 (totaling 16416) to 57x32 (totaling 1824). Finally, these features are transformed into a flattened vector of length 1824, using as input for the machine learning model.

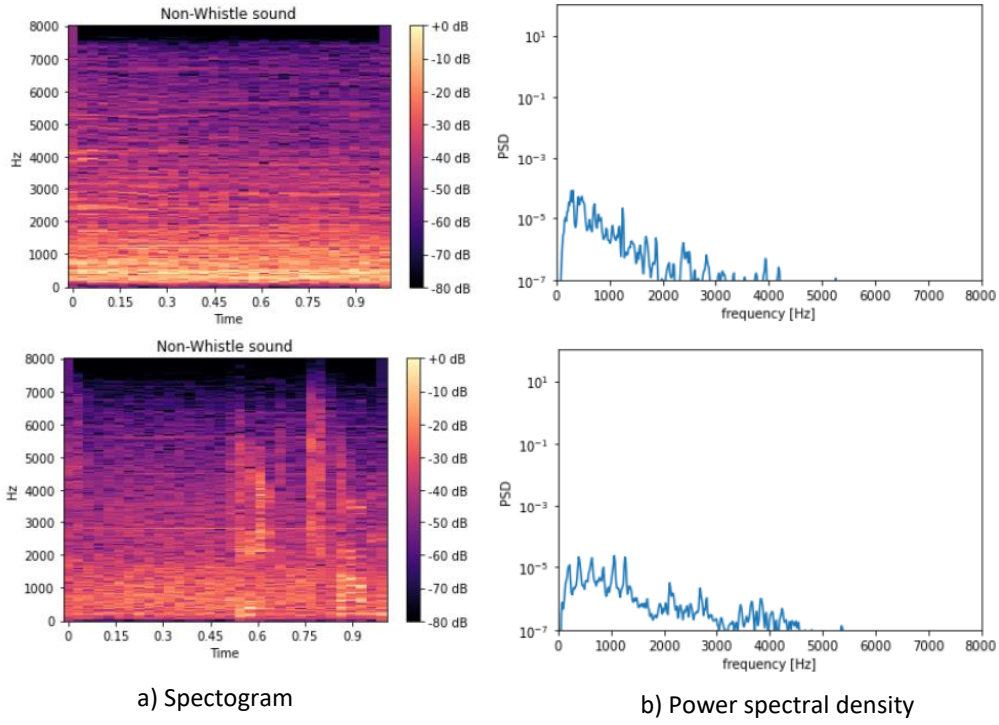


Figure 3. The spectrogram and the PSD of audio signals without whistle sounds

2.2 Whistle sound detection based on machine learning model

In this study, we focus on audio data extracted from soccer videos to identify the timing of the referee's whistle. Given that a typical match lasts at least 90 minutes, we derive the audio clips from this extensive duration. We implement a fixed-size sliding window technique to segment the audio data into shorter intervals. Then, we employ feature extraction methods using the STFT along with predefined frequency bands. Finally, we utilize a pre-trained machine learning model to classify these audio segments, determining the presence of whistles.

Our experiments indicate that the duration of whistling during a football match typically ranges from 0.3 to 1 second. We use a 1-second sliding window with a 50% overlap for analyzing the audio data. Features are extracted from this 1-second window and used as input for a binary classifier that differentiates between the presence and absence of whistles. We utilize well-known machine learning algorithms, including SVM, Neural Network, and Naïve Bayes, for classification. The outcome of these three models will be evaluated, and each audio sample will be categorized as either containing a whistle or not, based on the highest probability output.

2.3 Post-Processing method

After the classification step, we obtained results in temporal whistle detection. However, several limitations affecting the system's performance have been identified [3]:

- False alarms: The classification model incorrectly identifies the presence of a whistle sound when there is none. This misclassification primarily occurs due to the input audio being influenced by similar sound events, such as audience cheers, commentary, or whistles from fans, especially in matches with large crowds.

- Interleaving: This issue arises due to a 50% overlap in audio segments, which leads to the mixing of different audio events. The whistle sound is often masked by background noise and other sounds, causing certain portions of the whistle to go undetected.
- Overlapped boundary: The use of a sliding window with a 50% step can cause multiple audio segments containing whistles to overlap continuously, making accurate segmentation challenging

To solve these challenges, we apply fine-tuning rules and post-processing techniques based on power spectral density (PSD), along with predefined thresholds [3].

a) Fine-tuning rules

To address the initial two problems, we implement the majority voting rules utilizing the sub-segments s1, s2, s3, and s4.

Rule 1: Handling the "false alarm"

If s1 ≠ whistle and s2 = whistle and s3 ≠ whistle and s4 ≠ whistle, then s2 ≠ whistle

Rule 2: Handling the "interleaved"

If s1 = whistle and s2 ≠ whistle and s3 = whistle, then s2 = whistle

These simple rules aim to identify an audio segment as erroneous if its classification diverges from the labels of adjacent segments, given the use of a sliding window with 50% overlap. To address the issue of continuously overlapping whistles, we will combine the sounds into a longer output. While this rule-based adjustment can minimize misclassification errors, it cannot eliminate them. In fact, a sound may encompass multiple audio events simultaneously, such as overlapping whistles, commentary, and fan cheers. Therefore, we will apply additional post-processing techniques based on the PSD and established thresholds to enhance accuracy.

b) Post-processing based on the predefined PSD

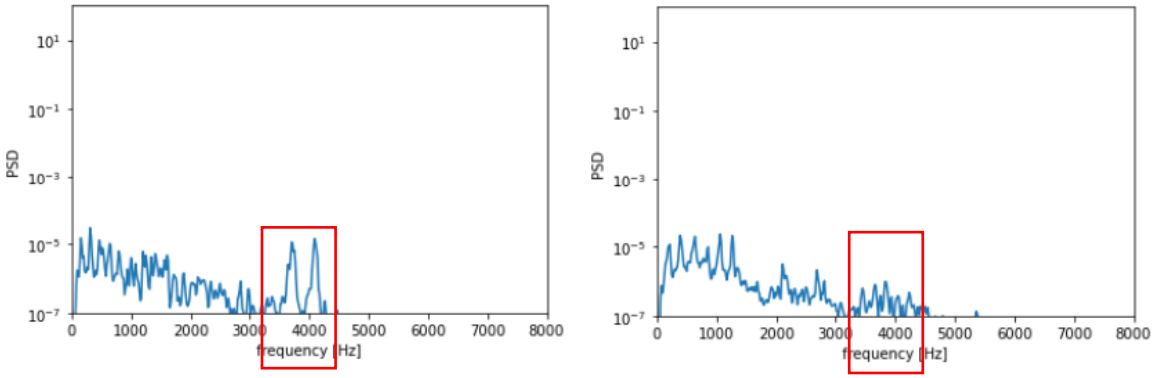
The sound of referee's whistle during a match is characterized by high frequency and intensity, specifically within the frequency range of [3300, 4200] Hz. Depending on this, we propose a post-processing technique aimed at accurately identifying the whistle sound and reducing instances of misclassification (False alarm). After the sound segments have been classified by the machine learning algorithm, we compute the total values of the PSD within the frequency range of [3300, 4200] Hz using the following formula:

$$SPSD_W = \sum_{FL}^{FU} |S(\omega)|^2 \quad (4)$$

where FU = 4200Hz, FL = 3300Hz and $S(\omega)$ representing the frequency spectrum of the audio signal in ω Hz. Then, this value is normalized in the range of [0, 1] using min-max normalization.

$$norm_SPSD_W = \frac{SPSD_W - min(SPSD_W)}{max(SPSD_W) - min(SPSD_W)} \quad (5)$$

Finally, within each output segment, an audio segment is identified as potentially including the referee's whistle if its $norm_SPSD_W$ value exceeds a specified threshold. This threshold is determined through experimental analysis. Figure 4 illustrates the power spectral density (PSD) values of two audio segments, one containing a whistle and the other without a whistle



a) Whistle sound b) Non-whistle sound

Figure 4. The power spectral density (PSD) values of two audio segments, one containing and the other not containing a whistle sound.

3. EXPERIMENTAL RESULTS

3.1 Dataset and analysis of experiments

Soccer match footage has been sourced from YouTube, featuring events such as the FIFA World Cup in 2014 and 2018, the Emirates FA Cup, and the UEFA Champions League. These videos are used for research purposes, with a specific focus on audio analysis to identify referee whistles. All footage is converted to WAV format at a sampling rate of 16 kHz, using a mono channel configuration. In total, there are 18 videos that have been gathered, which are categorized into training and testing sets. The information on soccer matches in our collected dataset is described in Table 1.

The training set comprises 15 videos, where the audios have been manually segmented into fixed-length segments of 1 second, organized into two layers: segments containing whistles and those without. The testing set consists of 3 videos, which include time-stamped annotations for whistle occurrences. Training data includes 15 videos, yielding 1,208 audio samples that contain whistles and 5,208 samples that do not. The testing data consists of 3 videos, with whistle events documented in an accompanying annotation file.

Table 1. The data collection of soccer matches sourced from YouTube.

No	Name of the video	Link source	Time (hh:mm:ss)
Training data			
1	Tunisia v England | 2018 FIFA World Cup |	https://youtu.be/u3wfrhjoIJg	00:02:10
2	Brazil v Costa Rica | 2018 FIFA World Cup	https://youtu.be/u2v_mb5Xx00	00:02:10
3	Germany v Mexico 2018 FIFA World Cup	https://www.youtube.com/live/xPfs2JL_4ws?feature=share	01:39:12
4	France v Croatia 2018 FIFA World Cup	https://www.youtube.com/live/SvV6aUki6LU?feature=share	01:41:55
5	Belgium v Japan (2018 FIFA World Cup)	https://www.youtube.com/live/GrkiZjoyugA?feature=share	01:42:23
6	Portugal v Spain | 2018 FIFA World Cup	https://www.youtube.com/live/OFbyNU6UQQs?feature=share	01:39:48
7	Everton vs Manchester City|Emirates FA Cup	https://youtu.be/dhbFM980HQk	01:37:43
8	Arsenal v Manchester City | Emirates FA Cup	https://youtu.be/1WTimKr6gNc	01:40:39
9	Southampton v Manchester City | Emirates FA Cup Quarter-Finals 21-22	https://youtu.be/JlboOicYHZ4	01:37:19
10	Chelsea v Liverpool | Emirates FA Cup	https://youtu.be/x8xdEOnv0Dk	01:39:32
11	The Blues Too Strong For Manchester United | Emirates FA Cup	https://youtu.be/fCG2ENyzbEA	01:53:12
12	Spain vs. Netherlands 2014 FIFA World Cup	https://www.youtube.com/live/WlNAln9mcg8?feature=share	01:39:12
13	Chelsea v Manchester City | Emirates FA Cup	https://youtu.be/CGFgHjeEkbY	01:39:36
14	Chelsea vs Leicester City | Emirates FA Cup	https://youtu.be/fS3skwfog9E	01:40:41
15	2018 FIFA World Cup: France v Argentina	https://www.youtube.com/live/Cbij3MKhdOY?feature=share	01:41:05
Testing data			
1	Brazil v Belgium | 2018 FIFA World Cup	https://youtu.be/5OJfbYQtKtk	01:50:50
2	Netherlands v Mexico | 2014 FIFA World Cup	https://youtu.be/a4Lt_hJdvNY	01:52:21
3	Liverpool -Real Madrid | UEFA Champions League	https://www.youtube.com/live/wvMNuU7HBUs?feature=share	02:21:30

In this research, we use a 1-second sliding window with 50% overlap to examine audio data. For the extraction of Short-Time Fourier Transform (STFT) features, we utilize Hanning window consisting of 1,024 samples, with an FFT size of 1,024 and a hop size of 512. We focus on 57 frequency bins ranging from 212 to 269, which correspond to a frequency range of 3,300 to 4,200 Hz. The extracted features are standardized using z-score normalization and flattened into a vector of length 1,824 to serve as inputs for the binary classification model.

We evaluate three models of machine learning: Support Vector Machine (SVM) with kernel RBF, Neural network (MLP), and Naïve Bayes, using the Scikit-Learn library. A post-processing technique based on Power Spectral Density is applied with a threshold of 0.01 to enhance performance of system. The effectiveness of the system is measured by comparing the classification results against groundtruth labels in the testing set, using metrics such as Precision, Recall, and F1-Score.

3.2 Results and discussion

We evaluate the performance of the audio detection system using three test datasets, both with and without post-processing. Tables 2 and 3 compare the performance of the whistle sound detection system without and with post-processing, respectively.

Table 2. Comparison of the whistle sound detection system without post-processing.

Number	Model	Number of whistles (groundtruth)	True positive	False positive (false alarm)	False negative (miss detection)	Precision (%)	Recall (%)	F1-value (%)
Test_1	**MLP**	71	56	19	15	0.746	0.788	0.767
	SVM+RBF		**53**	**7**	**19**	**0.883**	**0.736**	**0.803**
	NB		59	57	12	0.508	0.830	0.631
Test_2	**MLP**	61	44	32	17	0.578	0.721	0.642
	SVM+RBF		**46**	**12**	**16**	**0.793**	**0.741**	**0.767**
	NB		50	92	10	0.352	0.833	0.495
Test_3	**MLP**	52	42	17	10	0.711	0.807	0.756
	SVM+RBF		**37**	**4**	**15**	**0.902**	**0.711**	**0.795**
	NB		44	196	8	0.183	0.846	0.301

Table 3. Comparison of the whistle sound detection system with post-processing

Number	Model	Number of whistles (groundtruth)	True positive	False positive (false alarm)	False negative (miss detection)	Precision (%)	Recall (%)	F1-value (%)
Test_1	**MLP**	71	56	17	15	0.767	0.788	0.777
	SVM+RBF		**53**	**6**	**19**	**0.898**	**0.736**	**0.809**
	NB		59	53	12	0.526	0.830	0.644
Test_2	**MLP**	61	44	21	17	0.676	0.721	0.698
	SVM+RBF		**45**	**7**	**17**	**0.865**	**0.725**	**0.789**
	NB		50	79	10	0.387	0.833	0.529
Test_3	**MLP**	52	42	15	10	0.736	0.807	0.770
	SVM+RBF		**35**	**3**	**17**	**0.921**	**0.673**	**0.777**
	NB		44	153	8	0.223	0.846	0.353

In Table 2, SVM+RBF consistently demonstrates the highest precision and F1-value across all tests, particularly in Test_1, where it achieves a precision of 0.883 and an F1-value of 0.803. While the Naive Bayes (NB) model records the highest recall rates in most cases, its performance is hindered by significantly higher false positive rates, resulting in lower precision and F1-values. The MLP model shows moderate performance, with balanced precision and recall values; however, its F1-values remain lower than those of SVM+RBF across all tests. Overall, the results highlight that SVM+RBF is the most robust model without post-processing. In Table 3, SVM+RBF continues to outperform the other models, achieving the highest precision (0.921) and F1-value (0.809) in Test_3. Post-processing significantly reduces the false positive rates for most models, particularly for NB, which sees an improvement in precision despite its recall rates remaining high. The MLP model also benefits from post-processing, with a slight increase in precision and F1-value,

although it still lags behind SVM+RBF. These findings demonstrate the effectiveness of post-processing in enhancing the overall reliability of the detection system, particularly in reducing false alarms.

Furthermore, the comparison between Tables 2 and 3 highlights the important role of post-processing in refining the performance of whistle sound detection systems. Post-processing enhances precision across all models, which is crucial for minimizing false positives in real-world applications. Among the models, SVM+RBF consistently demonstrates the best trade-off between precision and recall, making it the most suitable choice for the whistle detection task. On the other hand, NB, despite achieving high recall rates, struggles with false positives, making it less practical even with post-processing. These results suggest that while post-processing is essential for improving precision, the choice of the underlying model remains a critical factor in determining overall system performance.

4. CONCLUSION

In this study, we proposed a simple and efficient method for extracting whistle sound features in combination with machine learning models to identify the referee's whistle during soccer matches. The model is composed of three main elements: (1) Extracting features using Short-Time Fourier Transform (STFT) and specified frequency bands, (2) Implementing machine learning algorithms such as Support Vector Machines (SVM) with Radial Basis Function (RBF) Kernel, Multi-Layer Perceptron (MLP), and Naïve Bayes for audio classification, (3) Utilizing a post-processing technique based on power spectral density to minimize errors and enhance performance. We evaluate the performance of our system by using audio recordings from soccer matches available on YouTube. The experimental results indicated that the model performed well, particularly in challenging acoustic environments. The future development direction will focus on exploring alternative feature extraction methods, particularly deep learning approaches, in comparison to other machine learning models. Additionally, integrating audio data with visual information from video footage will be considered to enhance the analytical system and provide a comprehensive overview of key moments in soccer matches.

REFERENCES

[1] X. Ziyou, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, 6-9 July 2003 2003, vol. 3, pp. III-401, doi: 10.1109/ICME.2003.1221333.

[2] K. P, M. Manikandan, and S. Kp, "Automated Referee Whistle Sound Detection for Extraction of Highlights from Sports Video," *International Journal of Computer Applications*, vol. 12, 12/01 2011, doi: 10.5120/1729-2340.

[3] T. Tan Dat, M. Xinjie, L. Hak-Jae, K. Jin Young, and C. Seung-Ho, "Cheering Event Detection in Basketball Audio Stream Using Adaptive GMM Model and Low Rank Matrix Recovery," (in English), *The Journal of Korean Institute of Information Technology*, vol. 14, no. 10, pp. 87-96, 2016, doi: 10.14801/jkiit.2016.14.10.87.

[4] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham, "Sports Video Summarization using Highlights and Play-Breaks," 01/01 2003, doi: 10.1145/973264.973296.

[5] D. Zhang and D. Ellis, "Detecting sound events in basketball video archive," *Dept. Electronic Eng., Columbia Univ., New York*, 2001.

[6] K. Hayashi and M. Mori, "Whistle recognition using convolutional neural network," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 12-15 Oct. 2021 2021, pp. 423-424, doi: 10.1109/GCCE53005.2021.9622085.

[7] L. Lu, F. Ge, Q. Zhao, and Y. Yan, "A SVM-Based Audio Event Detection System," in *2010 International Conference on Electrical and Control Engineering*, 25-27 June 2010 2010, pp. 292-295, doi: 10.1109/iCECE.2010.78.

[8] Q. Huang and S. Cox, "Hierarchical language modeling for audio events detection in a sports game," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 14-19 March 2010 2010, pp. 2286-2289, doi: 10.1109/ICASSP.2010.5495935.

[9] A. Baijal, C. Jaeyoun, L. Woojung, and K. Byeong-Seob, "Sports highlights generation bas ed on acoustic events detection: A rugby case study," in *2015 IEEE International Conference on Consumer Electronics (ICCE)*, 9-12 Jan. 2015 2015, pp. 20-23, doi: 10.1109/ICCE.2015.7066303.

[10] S. Liu, M. Xu, H. Yi, L.-T. Chia, and D. Rajan, "Multimodal Semantic Analysis and Annotation for Basketball Video," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, p. 032135, 2006/12/01 2006, doi: 10.1155/ASP/2006/32135.