INNOVATIONS IN FOOTBALL FROM PITCHSIDE MICROPHONE AUDIO DATA

2.98 SPORTS TECHNOLOGY: ENGINEERING & INNOVATION – SPRING 2024

Dilan SriDaran
MIT Sloan, MBAn
dilan_s@mit.edu

Nicolas Stone Perez
MIT EECS & Sloan, BS
nstonep@mit.edu

Matthieu Perrin
MIT Sloan, MSMS
matp1802@mit.edu

Ethan Fahimi
MIT Sloan, MBAn
fahimi@mit.edu

Nicolas Evans
FIFA
nicolas.evans@fifa.org

Johsan Billingham
FIFA
johsan.billingham@fifa.org

Rob Oldfield
Salsa Sound
rob@salsasound.com

Max Walley
Salsa Sound
max.walley@salsasound.com

Henry Wang
MIT Sports Lab
hwang21@mit.edu

Christina Chase
MIT Sports Lab
cchase@mit.edu

Peko Hosoi
MIT Sports Lab
peko@mit.edu

May 27, 2024

ABSTRACT

Fédération Internationale de Football Association (FIFA) is dedicated to its mission of democratizing football data. This initiative is crucial for a wide range of stakeholders, including referees, coaches, analysts, broadcasters, players, and fans, who rely on comprehensive data to understand specific moments in the game of football. At the highest level, traditional methods of data collection primarily rely on a combination of optical and wearable tracking and manually annotated data, which are prohibitively expensive and create significant barriers across the football pyramid. FIFA is exploring innovative ways to help democratize this data, leveraging existing, accessible, and cost-effective infrastructure such as footage from broadcast cameras and sound from pitch side microphones. In collaboration with Salsa Sound and the MIT Sports Lab, this project explores two early stage concepts for sound data in football: can sound be used to (a) identify and (b) localize individual football events. This was explored through two independent projects. Firstly, the differentiation of kick and drum sounds to help understand whether sound events can be isolated against other noises in the stadium (a). Secondly, the precise determination of a referee's position to understand whether sound can localize events and serve as a viable alternative to expensive optical tracking systems (b). Our findings indicate that sound can provide valuable and near real-time insights within these contexts. However, sound data is inherently subject to limitations, notably interference from crowd activity, which can diminish its accuracy relative to optical-based technologies. Nevertheless, there is considerable potential for its use in football, either as an economical tool to corroborate data from other tracking technologies, to be developed further in conjunction with broadcast cameras, or to serve as a standalone tracking solution.

***K*eywords** sound data · sound localization · tracking data · deep learning · classification

1 Introduction

Fédération Internationale de Football Association (FIFA) has a mission to democratize football data, aiming to level the playing field by providing equitable access to advanced analytical tools and datasets. This initiative is crucial for a wide range of stakeholders, including referees, coaches, analysts, broadcasters, players, and fans, who rely on comprehensive

data to understand specific moments in the game of football. Currently, FIFA's approach to player tracking or event detection relies on sophisticated optical tracking or manual processes, neither of which are cost and time effective. These impose significant barriers to deployment in lower-level settings.

FIFA is interested in leveraging existing infrastructure to help democratise this technology such as footage from broadcast cameras, and most recently sound from pitch side microphones. Notably, sound data can be collected, stored, and analyzed in a relatively cost-efficient manner relative to state-of-the-art optical tracking data, and therefore offers the potential to be scaled and deployed in any stadium, either as a standalone solution or used in conjunction with other data modalities.

This project, commissioned by FIFA and performed in collaboration with Salsa Sound (herein "Salsa") – a third-party AI-driven audio mixer for live broadcast – explores two early stage concepts for ambient sound data from broadcast microphones in football:

1. **Kick & Drum Classifier:** Sound data from pitch-side microphones offers significant opportunities to identify on-field events, such as the referee's whistle, ball kicks, passes, and shots, as well as off-field crowd-related events essential for fan engagement and crowd management. A critical challenge in this endeavor is the ability to isolate specific events from the ambient noise in the stadium, as the unique atmosphere created by drums, cheers, chants, and whistles can interfere with the detection of these events. Previous work by Salsa has included developing algorithms to extract kick events from sound data; however, distinguishing kicks by players from drum sounds in the spectator stands remains a major challenge due to their similar acoustic profiles. To address this shortcoming, this research develops a robust deep learning classification model to distinguish between these two sounds, which can serve as an additional screening layer to refine the outputs of Salsa's existing model.
2. **Referee Localizer:** Identifying specific sounds and determining their exact locations can significantly enhance game analysis and management. If kicks can be accurately identified in audio data, could the location of these sounds serve as a cost-effective alternative to expensive optical tracking systems, providing real-time ball positioning? Additionally, could kick data help identify the exact moment the ball is played for offside decisions? Similarly, could tracking the location of the referee's whistle improve understanding of officiating decisions and game flow by pinpointing the referee's position? While research continues on distinguishing between kicks and drums, we concurrently explore the potential of extracting the referee's location using whistle sounds captured by multiple microphones. Neither FIFA nor Salsa currently use sound for positional localization, and this therefore represents a novel investigation of the potential of sound data within this context.

The data used for this project are obtained from FIFA and Salsa. The Referee Localizer uses full match audio (sample rate of 48,000 Hz) from 18 microphones positioned around the field at two 2022 FIFA Men's World Cup matches: Brazil vs. Croatia and England vs. France. This was complemented by a structured dataset of manually detected whistles and their approximate timestamps, curated by Salsa. For validation, FIFA provided known referee tracking data (sample rare of 25 Hz). The Kick & Drum Classifier uses a sample of 0.4 second audio clips (sample rate of 48,000 Hz) of both kicks and drums, sourced from various matches across the English Premier League, Scottish Premier League, and FA Cup. There is no overlap in data between the two projects.

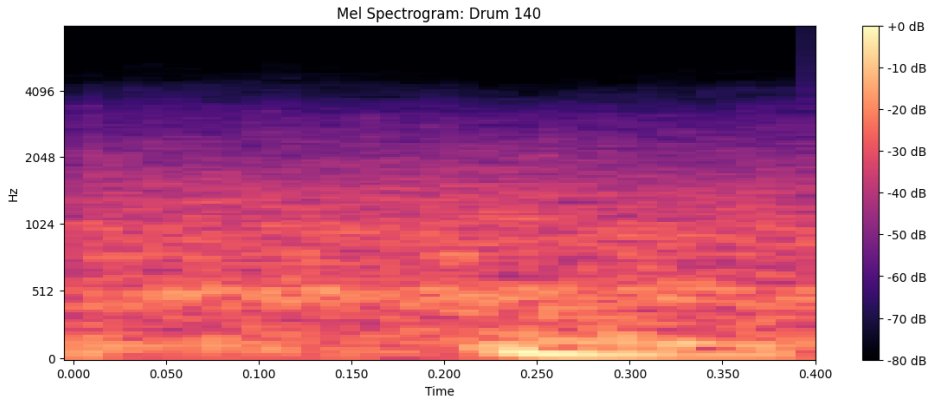
This report outlines our development of models for these two tasks, together with a broader assessment of the potential use cases and limitations of sound data. In Sections 2 and 3, we explore the Kick & Drum Classifier and Referee Localizer respectively, discussing the methodologies employed, baseline results achieved, sensitivity analyses performed, and limitations identified. Section 5 includes a broader discussion of other potential explorations using sound data, which may guide the strategic roadmap for both FIFA and Salsa.

2 Kick & Drum Classifier

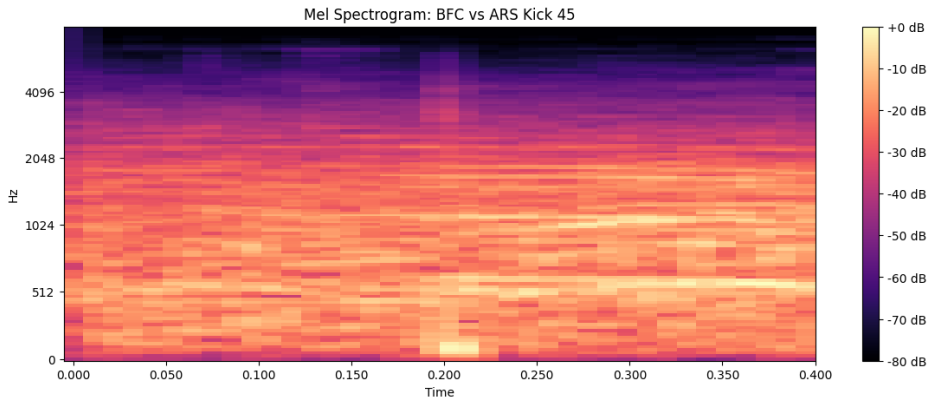
The Kick & Drum Classifier uses deep learning architectures to classify 0.4-second sound clips into either player kicks or external drum noises from the audience. It transforms each sound clip into a Mel spectrogram, a visual representation of frequency spectra over time. Convolutional neural networks analyze these images to recognize patterns, enabling the classifier to accurately identify similar sounds in new samples.

2.1 Methodology

2.1.1 Data Pre-processing



(a) Drum 140 from mixed collection.



(b) Kick 45 from Brentford vs Arsenal.

Figure 1: MEL spectrogram of 0.4 second audio clips.

Salsa manually identified and provided sound samples for 882 known kicks and 465 drums. Kicks were sourced from two English Premier League matches: 287 samples from Brentford FC vs. Arsenal FC and 595 samples from Southampton FC vs. Chelsea FC. Drums were provided in two batches: 99 samples from a Scottish Premier League match featuring Motherwell (with the opponent unspecified) and 366 samples from a mixed collection encompassing various English Premier League and FA Cup matches. Note, Salsa provided anecdotal advice that Motherwell drums are particularly anomalous and unusual relative to drums in most other settings. Each image is transformed into a Mel spectrogram, as illustrated in Figure 1.

For the baseline model, we randomly distribute the data into training (60%), validation (15%), and test (25%) sets to ensure that the model learns from a diverse range of contexts. Although this method aids in understanding general learning patterns, our interest moving forward lies in assessing the model's ability to generalize to novel, unseen environments. Ideally, one would train the model using data from all matches except one, which is reserved for testing the model's performance. We investigate this approach through a sensitivity analysis in our study. However, we recognize that the limited number of matches in our dataset may not fully represent broader conditions. Despite this,

FIFA and Salsa may revisit this approach as more data becomes available, potentially enhancing the robustness of the findings.

2.1.2 Model Architecture

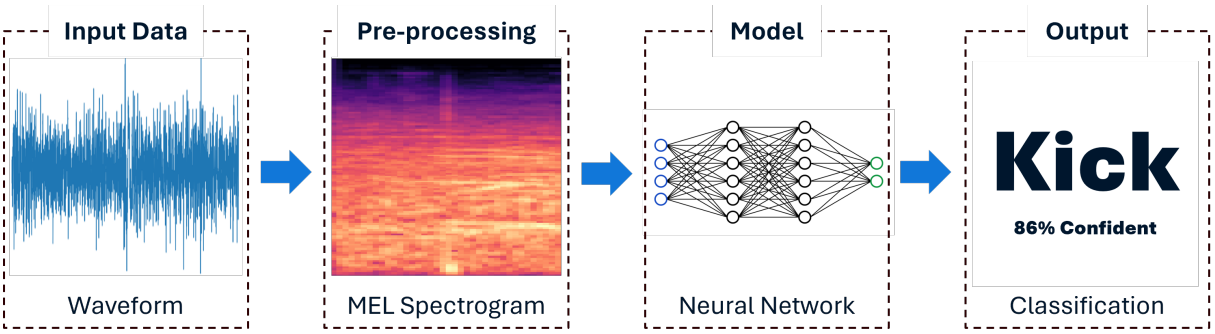


Figure 2: Model pipeline.

The baseline model takes as input the Mel spectrogram based on the full 0.4 second sound representation. The model then uses a neural network with two convolutional layers, max pooling and dropout used to minimize over-fitting, and a dense layer for binary classification. The base model is trained to minimize the categorical cross entropy loss, using 10 epochs and a batch size of 32. The output of the model is a predicted class for each sound, either “kick" or “drum," together with a quantification of the model’s confidence in its prediction.

2.2 Results

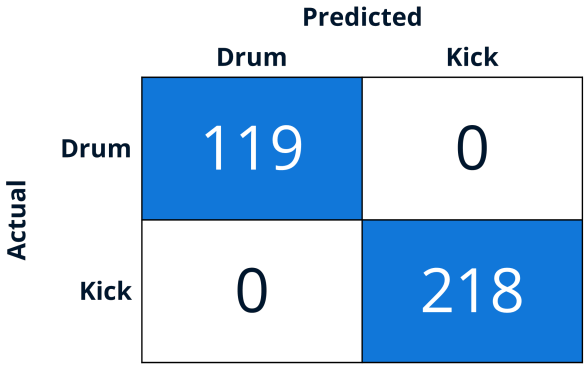


Figure 3: Out-of-sample confusion matrix.

The model achieves a perfect out-of-sample accuracy of 100% when using a randomly stratified train-test split. While promising, this result should be considered in the context of the relatively limited data available and the high sensitivity of the model, as discussed in Section 2.3.

2.3 Sensitivity Analysis

2.3.1 Sound Duration

Salsa provided audio samples, each with a duration of 0.4 seconds. In our experiments, we trimmed these audio clips to create “subsamples” by taking proportions of the full audio starting from the beginning of each clip. We then trained and evaluated our model using these subsamples. The findings were similarly strong, demonstrating that near-perfect accuracy can be achieved when the model used only a 0.16-second snippet from the original 0.4-second audio. This suggests that the model is capable of functioning with very low latency, making it suitable for near-real-time applications.

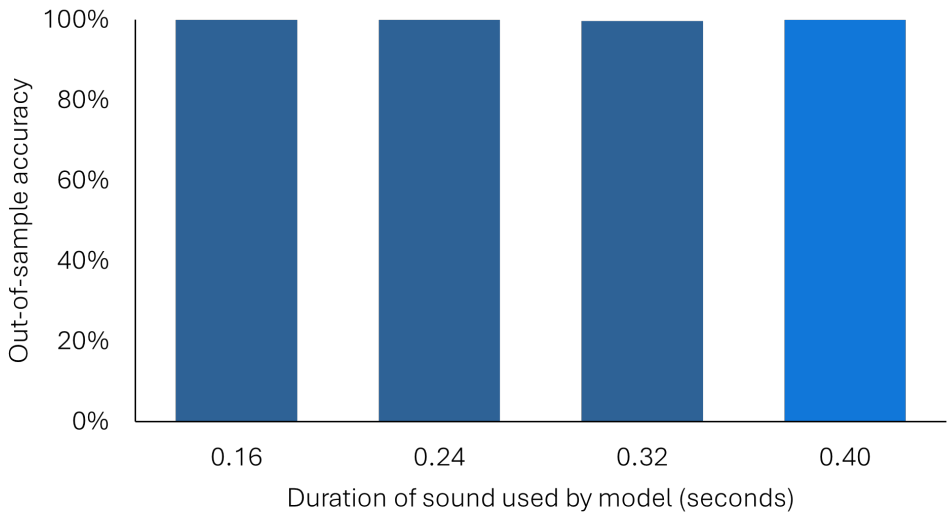


Figure 4: Model performance when varying the duration of the sound snippet.

2.3.2 Train and Test Splits

The near-perfect model results may be inflated due to the presence of sound samples from the same match in both the train and test sets, which may be leading to overfitting and a lack of generalizability. In practice, it is not realistic to have sound samples from the same match in both the training and testing datasets. Moreover, there are no matches available that include both kick and drum samples, and therefore there is a risk that the model may be learning its background environment, rather than the actual differentiation of a kick and drum.

To explore these issues, we adopted a rigorous method for assembling the train and test datasets. Rather than employing a random distribution, which is standard in many machine learning applications, we opted for a structured approach that ensures no match overlaps between the training and testing sets. This was achieved by manually selecting and segregating complete matches to one dataset or the other, thereby enhancing the robustness of the classification model against overfitting and ensuring a more realistic test of its predictive capabilities. Given two sources of kicks and two sources of drums, there are four unique ways of constructing such dataset partitions (Table 1). For each of the four combinations, we trained and evaluated the model.

Table 1: Out-of-sample accuracy by train-test distributions.

	A	**B**	**C**	**D**
Kicks: Arsenal vs. Brentford	Train	Test	Test	Train
Kicks: Chelsea vs. Southampton	Test	Train	Train	Test
Drums: “Normal"	Train	Test	Train	Test
Drums: Motherwell	Test	Train	Test	Train
Out-of-sample accuracy	83.6%	44.6%	99.5%	74.9%

The model exhibited notable sensitivity to how the data were partitioned for training and testing, with the model performing inconsistently across different configurations. Notably, the accuracy suffered when the training data included the anomalous Motherwell drum samples, which are characterized by unusually loud or distinct drum noises and a tightly-packed crowd near the field, which are not representative of typical match environments. These anomalous conditions likely lead the model to learn specific noise patterns of the Motherwell stadium, rather than generalizable features for accurately classifying kicks and drums. This is exacerbated by the relatively limited sample of audio files for the Motherwell match, with just 99 samples, compared to 366 “normal" samples. This means that there is less data from which to learn and extract signals when training on the Motherwell match only.

By contrast, the model appears less affected by the choice of kick samples for training, likely reflecting that match-related events are far more consistent than fan-generated or external noises.

2.3.3 Model Architecture

We explored refinements to the model's architecture to improve on the limitations identified. This included adjusting key training parameters such as the number of epochs (2 to 10) and batch sizes (16 to 64), and addressing potential class imbalances through downsampling to balance the number of kick and drum samples in the training dataset. None of the modified models outperformed the baseline established by the original settings.

2.4 Discussion

The model demonstrates potential to act as an accurate, low-latency tool for distinguishing between kicks and drums. However, the findings also emphasize the vital role of diverse and representative training data in creating a robust and generalizable solution that can confidently be applied across various tournaments, venues, and stadium environments. Extensive testing of different model architectures indicates that the primary limitation to achieving robustness lies in the volume and variety of data, not in the model structure itself. This is an area that FIFA could focus on to enhance the model's effectiveness.

Future enhancements to the model could also involve incorporating additional types of sounds, such as whistles, fan music, stadium announcements, and the distinct sounds of kicks, touches, passes, and players running. These improvements could significantly benefit multiple aspects of football, across management, refereeing, and broadcasting. Additionally, better detection of whistles could further refine and augment the capabilities of the referee localizer, discussed in the next section.

3 Referee Localizer

The Referee Localizer estimates the referee's position by analyzing the time delays in whistle detections across multiple microphones. These microphones have known positional coordinates. Therefore using the speed of sound in air (343 m/s), these time delays provide insights into the relative distances from the referee to each microphone, thereby facilitating a localization of their position. The key challenge, however, is to accurately estimate the time delays between microphones, with difficulties arising from:

1. **Inconsistent Definitions:** There is no consistent and objective definition for the "start of a whistle" with which to compare signals across different microphones.
2. **Sound Interference:** Microphones positioned at different locations may not detect consistent sounds due to different background fan noises across the stadium and obstructions by moving players.

3.1 Methodology

3.1.1 Data Pre-processing

Salsa provided 18 microphones of full match audio from two matches at the 2022 FIFA Men's World Cup: Brazil vs. Croatia (2 hr 55 min 31 sec) and England vs. France (2 hr 4 min 26 sec).[1] The exact timing of the start of each whistle within each audio file is unknown. However, Salsa provided manually approximated timestamps of whistles throughout each match. These timestamps are not sufficiently accurate to perform the referee localization directly, but do provide a starting guide.

Salsa provided 1,117 and 790 whistle identifications across the two matches respectively, noting that the same whistle may be detected by multiple microphones. We restrict the analysis to only whistles identified during normal and extra time, excluding sounds during pre-match, half-time, post-match, and penalties.[2] We assume that any whistles identified by different microphones within 0.39 seconds of each other represent the same unique whistle event, with this threshold determined based on the maximum possible distance, and by extension time delay, between microphones. This yields $W = 149$ and $W = 106$ unique whistle events per match respectively.

For each whistle event, we extract the same 4 second audio segment from each microphone, centered around the first time of detection of that whistle event by any microphone using Salsa's manual timestamps. This is sufficiently long to ensure that the full audio signal of the whistle is present (if detected) within all microphone segments.

3.1.2 Model Framework

Let o_{ijw} denote the observed arrival time lag of Whistle $w \in [W]$ at Microphones $i \in [1, 18]$ and $j \in [1, 18]$.

To compute o_{ijw}:

1. **Bandpass Filtering:** Filtering is applied to the 4 second sound segments from Microphones i and j, isolating frequencies between 3,750 Hz and 4,250 Hz that represent the typical frequency range of a standard whistle. By utilizing a Butterworth filter of order 5, we ensure a maximally flat response in the passband, thereby minimizing distortion within this critical frequency range.
2. **Cross-correlation:** Cross-correlation is used to measure the time offset of the whistle between two microphones. Cross-correlation measures the similarity between two signals and identifies the time offset that minimizes the inner product of their sound representations (i.e., the time offset required to minimize the subsequent difference between the two audio profiles). We impose a microphone pair-specific upper limit on the time offset permissible based on the Euclidean distance, d_{ij} between Microphones i and j, and the speed of sound in air (343 m/s): $o_{ijw} \leq d_{ij}/343$.

This is repeated for all microphone pairs to determine $\mathbf{O}_w$, an 18×18 matrix of "observed" time lags for each whistle.

The football pitch is then divided into non-overlapping and field-exhaustive square tiles, each with a known positional coordinate. Assuming the whistle originated in Tile t, the time of flight to each microphone can be computed using their Euclidean distance, d_{it}. A theoretical time lag can therefore be computed between Microphones i and j as $(d_{it} - d_{jt})/343$. This is repeated for all microphone pairs to determine $\mathbf{E}_{tw}$, an 18×18 matrix of "expected" time lags for each whistle assuming the sound originated in tile t.

[1] The Brazil vs. Croatia audio files are longer in duration as the match went to extra time and penalties.

[2] Timestamps for kick-off and end-of-half whistles were not provided. These times are approximated using manual review of the audio data, and have been validated against publicly available match summaries.

The sound source for Whistle w can then be predicted as:

$$\text{argmin}_t \|\mathbf{E}_{tw} - \mathbf{O}_w\|^2 \tag{1}$$

3.1.3 Algorithm Implementation

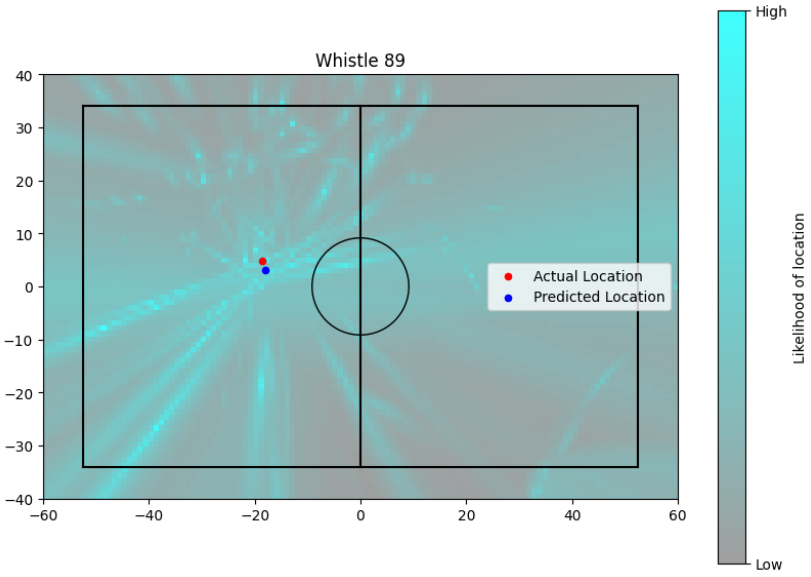


Figure 5: Implementation of localization approach for Whistle 89 of Brazil vs. Croatia, using a single grid size of 1x1m.

The methodology presented herein is conceptually robust and effectively addresses the primary challenge of aligning sound waves. Nonetheless, the practical application of this method encounters a secondary issue, as some calculated cross-correlations prove to be erroneous or spurious due to distortions by crowd noise and sound obstructions. Such inaccuracies can significantly skew the predicted location.

To mitigate the influence of unreliable microphones, we have adopted an iterative approach. Rather than relying on the complete 18×18 matrix for localizing the referee, we employ a subset strategy wherein only three microphones are selected at a time. The localization is then conducted using the corresponding 3×3 matrix derived from these microphones. This procedure is systematically repeated across all 816 unique triadic combinations of the microphones, resulting in 816 distinct predictions. Predictions located on or beyond the boundary of the pitch are excluded from consideration, and the median of the remaining coordinates is determined to establish the final predicted location, shown in blue (Figure 5).

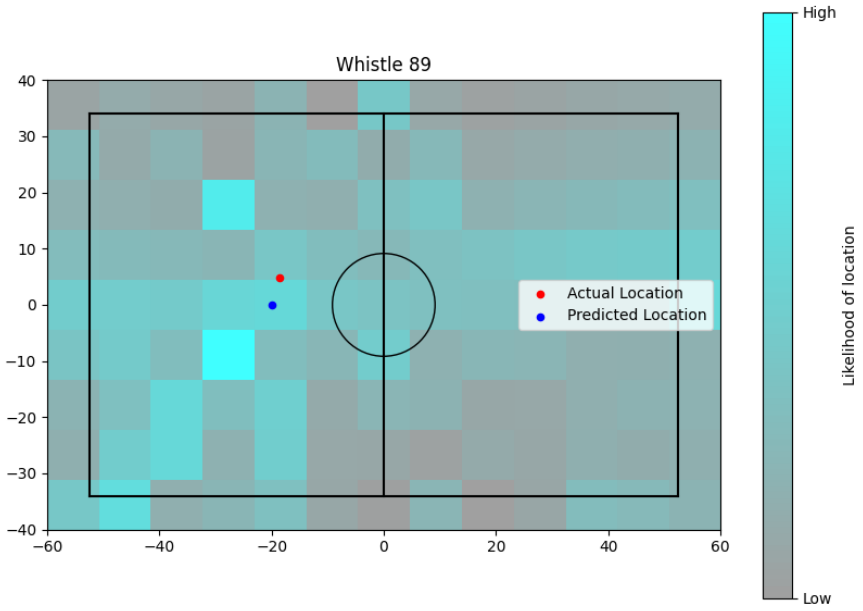
The iterative method substantially improves accuracy but at the cost of high computational demand, taking as long as 40 seconds to localize a single whistle for a 1×1 field grid. To mitigate this complexity, we introduce a further model refinement through an (optional) multi-stage approach. Initially, we determine the approximate location of the whistle using a coarse grid spanning the entire pitch. Subsequently, a more precise localization is conducted using a finer grid focused solely on the area surrounding the initially identified location (Figure 6). This iterative approach can be repeated multiple times as required.

3.1.4 Hyper-parameter Selection

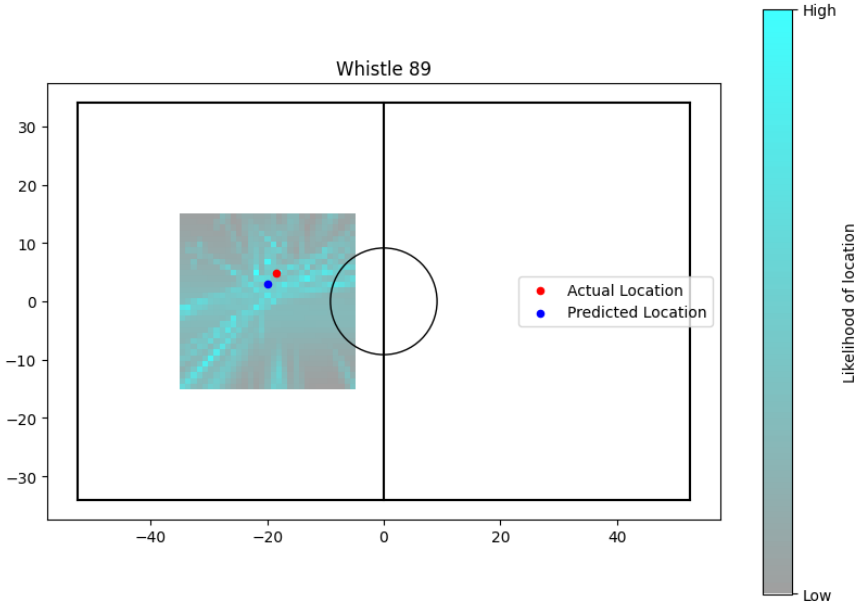
The algorithm relies on two key hyper-parameter selections: the grid size(s) defined and the frequency range considered. Based on extensive testing, we have selected an iterative grid size of [10, 1]m and a frequency range of 3,750 Hz to 4,250 Hz.

Grid Size

Different configurations were considered, including single-phase sizes of [1]m, [2.5]m, [5]m, and [10]m, and various multi-stage approaches. Run times were relatively consistent across both matches. However, the pattern of median errors deviated (Figure 7). Whereas accuracy improved with granularity in the Brazil vs. Croatia match as expected,



(a) Step 1: Coarse localization with 10x10m grid.

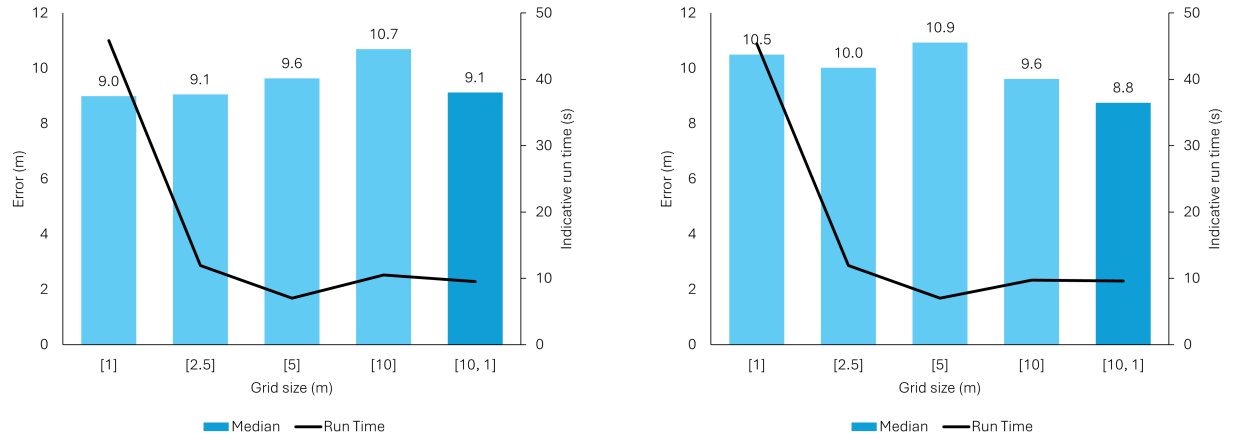


(b) Step 2: Fine localization with 1x1m grid.

Figure 6: Implementation of two-step localization approach for Whistle 89 of Brazil vs. France.

this pattern was not observed for England vs. France. Future work may consider extending this investigation to more matches to further explore this issue.

We select a two-phase [10, 1]m grid size. In both matches, this achieved the best or near-best accuracy, at approximately $5\times$ the speed of the [1]m grid. Other multi-phase grid sizes were tested, including two-phase grids with different components and three- and four-phase grids. These approaches did not yield consistently improved results across both accuracy and computational dimensions, and therefore were not selected.

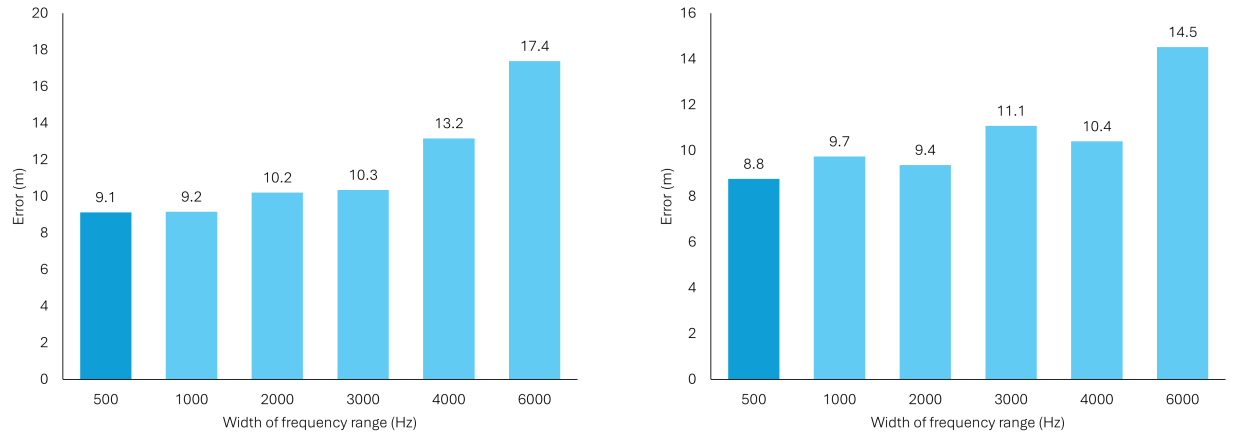


(a) Brazil vs. Croatia

(b) England vs. France

Figure 7: Median error and average run time for different grid size configurations across all whistles from Brazil vs. Croatia and England vs. France. Baseline selected model shown in darker blue.

Frequency Range



(a) Brazil vs. Croatia

(b) England vs. France

Figure 8: Median error for different frequency ranges across all whistles from Brazil vs. Croatia and England vs. France. Baseline selected model shown in darker blue.

Different frequency ranges, centered around 4,000 Hz were considered (Figure 8). Across both matches, the best performance was achieved using a range of 500 Hz, corresponding to a frequency band of 3,750 Hz to 4,250 Hz.

3.2 Results

Table 2: Summary statistics for performance metrics. All in meters.

Match	Median	75th Percentile	90th Percentile
Brazil vs. Croatia	9.1	13.9	21.3
England vs. France	8.8	22.6	31.8

The results indicate that sound data does contain signal to reasonably localize referees, with a median error of under 10 meters in both matches: 9.1 meters for Brazil vs. Croatia and 8.8 meters for England vs. France. Across the two matches, 32.9% and 31.1% of whistles are localized within 5 meters of the true referee's position and 56.4% and 56.6% within 10 meters respectively.

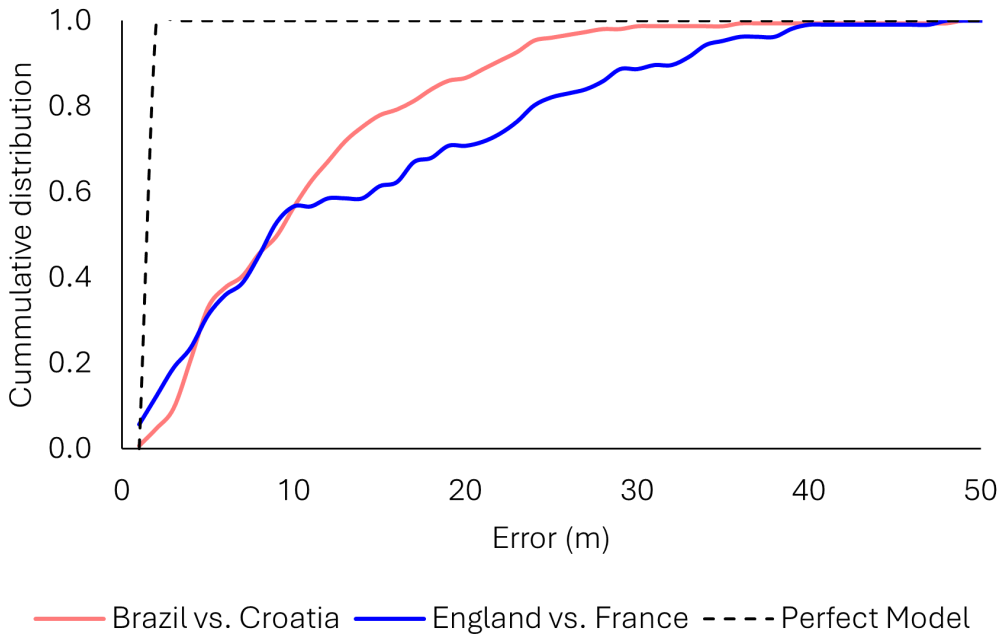


Figure 9: Cumulative distribution of errors for Brazil vs. Croatia and England vs. France.

However, the distribution of errors exhibits a pronounced skew, particularly evident in the England vs. France match, which suggests significant variability in performance across whistles within a match (Figure 9). While the majority of whistles can be localized within 10 meters of the true position, errors can occur up to 50 meters in magnitude for isolated instances. A visual depiction of these errors, on the scale of a football pitch, is provided in Appendix A.1.

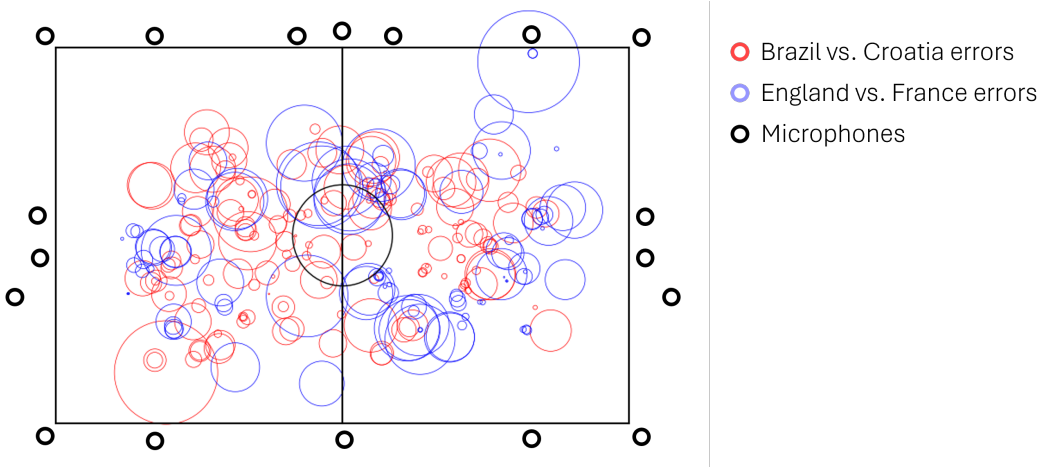
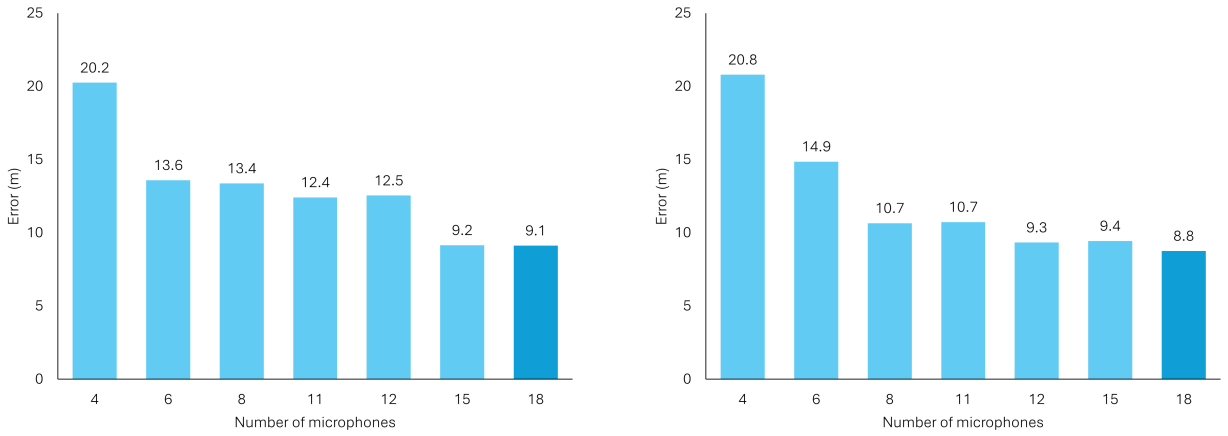


Figure 10: Errors by actual referee location for Brazil vs. Croatia and England vs. France.

Figure 10 examines the correlation between the referee's actual position and the magnitude of localization errors to determine if there are any consistent patterns that might elucidate the pronounced tail observed in the England vs. France match. The analysis reveals that errors tend to be minimized when the referee is positioned near the center of the pitch. Intriguingly, during the England vs. France match, the referee's presence near the center circle is noticeably reduced, which could potentially account for the observed disparities in error distributions. Moreover, it is important to highlight that errors significantly increase near the coordinates (0, 20), in proximity to the dugouts. We speculate that this may be attributed to elevated background noise from players and coaches, however further work would be needed to validate this hypothesis.

3.3 Sensitivity Analyses

3.3.1 Number of Microphones



(a) Brazil vs. Croatia

(b) England vs. France

Figure 11: Median error for different number of microphones considered across all whistles from Brazil vs. Croatia and England vs. France. Baseline selected model shown in darker blue.

Although less expensive than optical tracking, utilizing microphones still incurs financial costs, and additional microphones take time and effort to deploy and ensure connectivity. To evaluate the feasibility of the proposed model, we assessed its performance with varying subsets of available microphones, specifically with configurations of 4, 6, 8, 11, 12, 15, and 18 microphones. The configuration for each subset was optimized to maximize either the spacing between or the coverage area of the microphones. As depicted in Figure 11, there is a clear trend showing that accuracy improves as the number of microphones increases, though with diminishing returns. Notably, deploying as few as 6 microphones yields reasonably good results, which is promising given that lower league matches typically operate with 6 to 8 microphones.

3.3.2 Microphone Failure

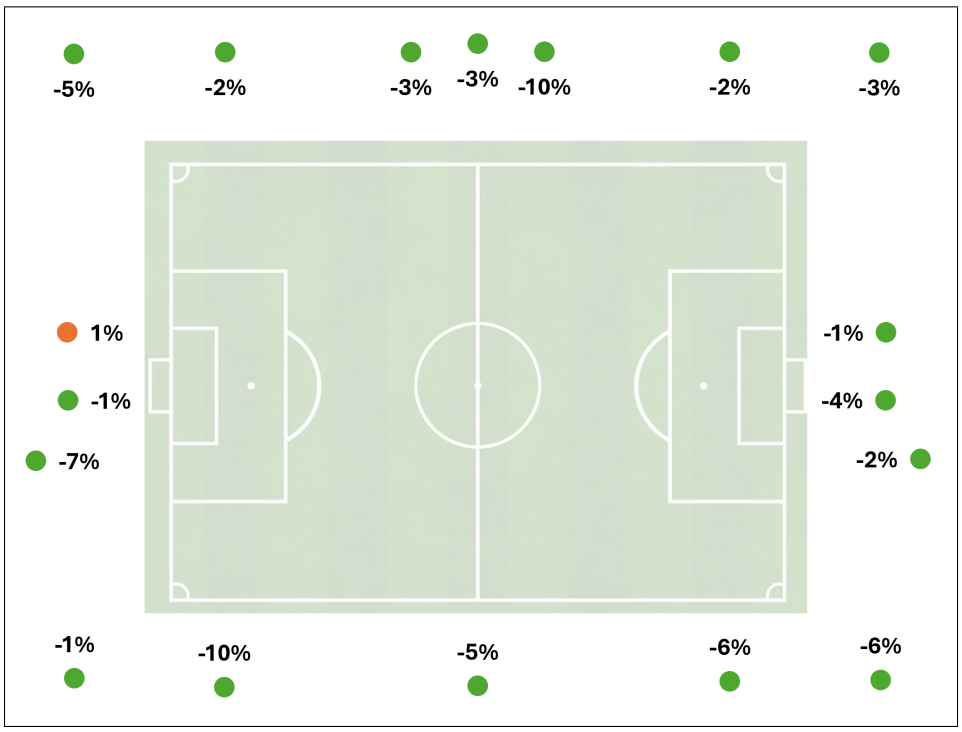
While relatively rare, microphone "failures," resulting from technical malfunctions or physical interference from players do occur. To assess the robustness of our model, we conducted a series of tests, each assuming the failure of a different microphone from the beginning of the match. In each scenario, the model relied on the remaining 17 microphones for the duration of the match. The effects of each microphone's failure on the accuracy for each match are depicted in Figure 12. Here, positive values represent an increase in the median localization errors (worse performance) for the match corresponding to the "failure" of a specific microphone.

Interestingly, the results from the Brazil vs. Croatia match suggest that the failure of almost any single microphone could actually enhance the overall performance, with only one exception. However, due to the limited number of whistles analyzed and the modest size of these improvements, these findings are likely not statistically significant. Nonetheless, it is encouraging to see that performance likely remains robust under various test conditions.

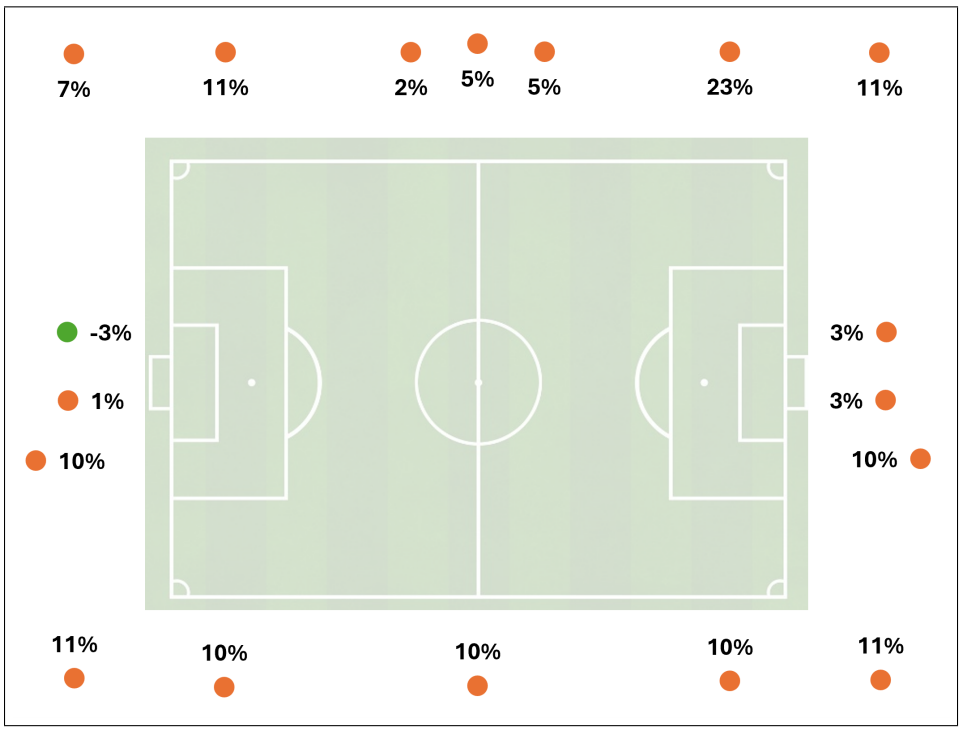
Conversely, in the England vs. France match, the results show a stark contrast, with the failure of most microphones leading to a performance decline of approximately 10% (or about 0.9 meters). This variance in outcomes between the two matches may be attributed to the positional differences of the referee during the games. In the Brazil vs. Croatia match, the referee predominantly stays near the center of the pitch, where no single microphone is crucial for most whistles. In contrast, in the England vs. France match, the referee's proximity to the pitch boundaries makes the microphone closest to the action significantly more important, and its failure results in a more substantial loss of information.

3.3.3 Microphone Placement Error

According to Salsa, microphone coordinates were measured only once at the start of the World Cup and were assumed to maintain their positions unchanged throughout the tournament. However, it is improbable that microphones were positioned identically for each match.



(a) Brazil vs. Croatia



(b) England vs. France

Figure 12: Median error for different microphone failures across all whistles from Brazil vs. Croatia and England vs. France. Percentages reflect the change in the median error across all whistles given the “failure" of that microphone.

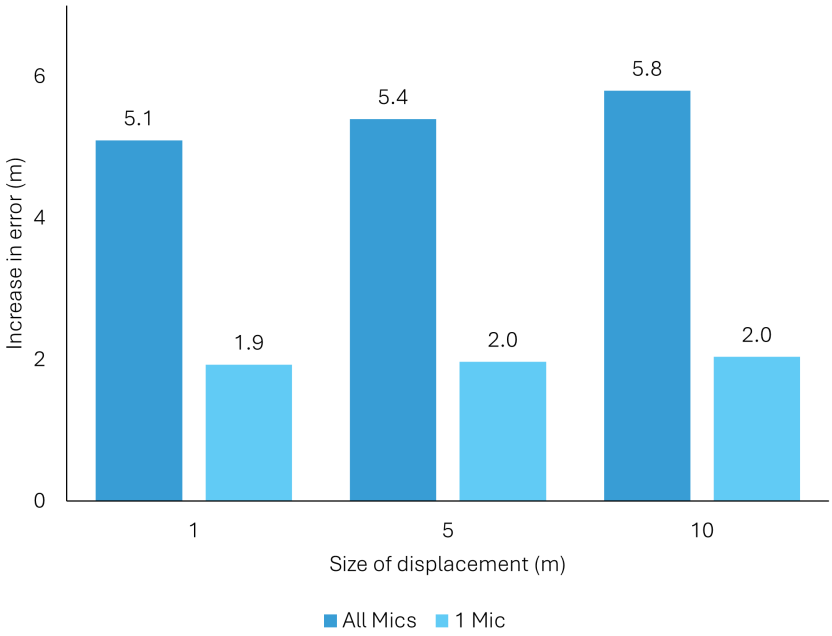


Figure 13: Increase in error from synthetic microphone displacements.

To evaluate the influence of errors in the positional accuracy of microphones, a theoretical study was undertaken. In this study, we simulate scenarios where the origin of the referee's whistle is predetermined, while the positions of the microphones are subjected to random perturbations. We investigate various scales of these perturbations and categorize them into two distinct types:

1. **All microphones:** All microphones undergo random displacements within a specified range. This condition is intended to simulate errors that may occur during the setup and positioning of the microphones.
2. **One microphone:** Only one microphone is randomly displaced within the same defined range. This scenario represents potential in-game shifts of a microphone due to interactions with players or contact with the ball.

As illustrated in Figure 13, there is a noticeable escalation in errors corresponding to the magnitude of displacement. This error amplification is more severe when all microphones are perturbed. In contrast, the error remains minimal when only a single microphone is displaced. These findings highlight the critical role of precise microphone placement, suggesting that FIFA should prioritize the initial placement and stability of microphones in future competitions.

3.4 Discussion

The results from our study underscore the viability of using sound data to localize referees in football matches, evidenced by the relatively modest median errors observed: 9.1 meters in the Brazil vs. Croatia match and 8.8 meters in the England vs. France match. Moreover, a significant portion of whistle localizations—over 30% in each match—occurred within 5 meters of the true referee's position, with more than half within 10 meters. These findings are promising, and highlight the potential for sound data to serve as an accessible alternative to more costly optical tracking technologies.

However, the distribution of errors shows significant variability, indicating that while sound data can be useful, it also has limitations in consistency across different scenarios. The pronounced error skew in certain matches highlights the need for improved microphone placement strategies to ensure more reliable data collection. Addressing these challenges will enhance the accuracy and reliability of sound-based tracking systems, making them a more viable tool for broad-scale implementation across various levels of football competition. Future work could explore three potential use cases: using the tool in its current capacity as a low-cost validation tool for existing video-based technologies, developing the algorithm further to serve as a robust and accurate standalone tracking tool, or expanding the model to a multi-modal context (e.g., integration with available broadcast video footage) for even further enhanced accuracy.

4 Final Remarks

4.1 Conclusion

This project has delved into the opportunities and limitations of using sound data in football. Recognizing the existing availability of audio data, which eliminates the need for additional hardware, and the low cost of capture and storage, sound data offers a promising avenue to democratize data access in football. Our investigations have yielded valuable insights. The Kick & Drum Classifier consistently demonstrates high accuracy, reaching 100% out-of-sample performance, suggesting that when coupled with an appropriate model framework and diverse environments, sound can accurately identify different events within a stadium. Concurrently, the Referee Localizer demonstrated the ability to predict the referee's position within a 10-meter radius for the majority of whistles. While not perfect, this result highlights that there is potential for sound to be used within a localization context, which can have profound impacts for optical tracking, semi-automated offside technologies, and referee analysis.

These outcomes encourage further research to improve upon the models built and into the broader application of audio data within football.

4.2 Next Steps

Despite the encouraging results, further efforts are essential to enhance the robustness and generalizability of our findings. These include:

- **Data collection:** Extensive collection of more raw audio data is required, including more diverse samples of kicks, drums, and whistles from different matches. This will improve the generalizability of the models developed.
- **Microphone placement accuracy:** The referee localizer is highly sensitive to the accuracy of microphone placement. In future competitions, FIFA should place greater emphasis recording their positions and ensuring they have not been moved in the course of play.
- **Advanced machine learning models:** Exploring machine learning-based solutions for referee localization, analogous to our kick and drum classifier, could further refine accuracy. However, this would require substantially larger volumes of data.

4.3 Future Explorations

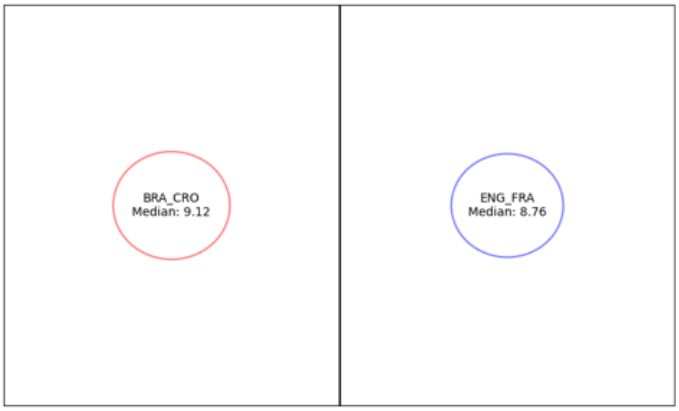
While focusing on kick and drum classification and referee localization, we also explored additional potential ideas for applications of sound data:

- **Event detection:** Investigating major match events such as goals and the start/end of each half through envelope transformers applied to sound data.
- **Fan sentiment analysis:** Utilizing rolling averages of sound amplitudes to analyze fan reactions and overall stadium atmosphere.
- **Tactical intensity:** Developing concepts like "tactical intensity" to correlate match tactics with audio data, potentially revealing team strategies and game dynamics.
- **Innovative broadcasting:** Considering alternative microphone placements to enhance broadcast experiences, similar to specialized NFL broadcasts aimed at younger audiences.
- **Referee training and analysis:** Using audio data to train and assess referees' performance by analyzing whistle sounds, movement, and fan reactions.

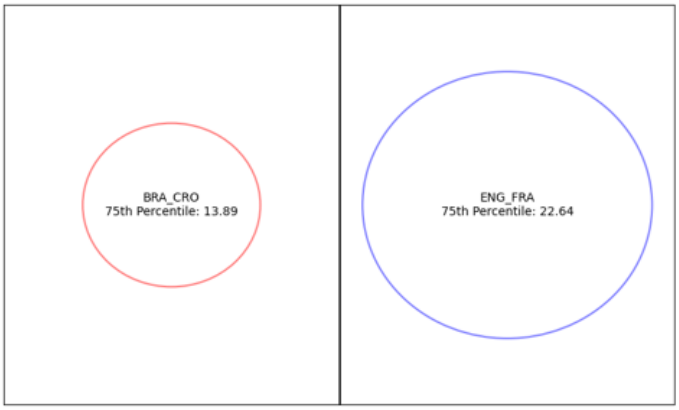
These initiatives could significantly expand the utility of sound data in football, offering novel insights and enhancing both fan experiences and game management, and may be explored with future work.

A Appendix

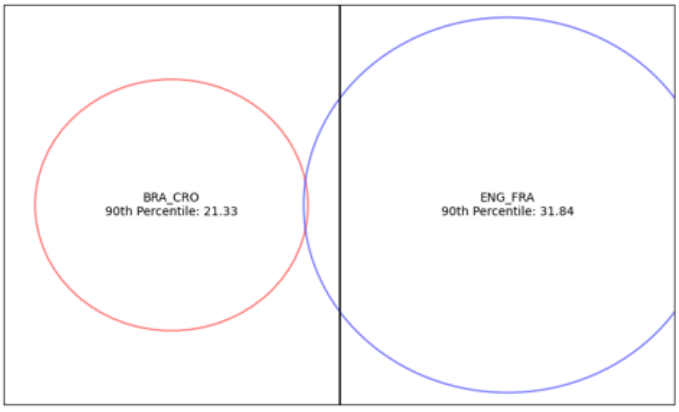
A.1 Distribution of Referee Localization Errors



(a) Median error.



(b) 75th percentile error.



(c) 90th percentile error.

Figure 14: Indicative size of referee localization errors on scale of football pitch.