Live Football Commentary System Providing Background Information

Yuichiro Mori[1,2], Chikara Tanaka[1,2], Aru Maekawa[1], Satoshi Kosugi[1], Tatsuya Ishigaki[2], Kotaro Funakoshi[1], Hiroya Takamura[2], Manabu Okumura[1]
[1]Institute of Science Tokyo,
[2]National Institute of Advanced Industrial Science and Technology (AIST)
{moriy,maekawa,kosugi,funakoshi}@lr.pi.titech.ac.jp, oku@pi.titech.ac.jp
{tanaka.chikara,ishigaki.tatsuya,takamura.hiroya}@aist.go.jp

Abstract

Previous research on sports commentary generation has primarily focused on describing major events in the match. However, real-world commentary often includes comments beyond what is visible in the video content, e.g., "Florentina has acquired him for 7 million euros." For enhancing the viewing experience with such *background information*, we developed an audio commentary system for football matches that generates utterances for the background information, as well as play-by-play commentary. Our system first extracts visual scene information and determines whether it is an appropriate timing to produce an utterance. Then, it decides which type of utterance to generate: play-by-play or background information. In the latter case, the system leverages external knowledge through retrieval-augmented generation.

1 Introduction

Live sports commentary enhances the audience's experience by conveying the excitement and depth of the sports matches including football (a.k.a. soccer). A commentator clarifies what is happening in the match, and also provides real-time analysis, capturing the stadium atmosphere and presenting key events from multiple perspectives to engage the audience in the match (Schaffrath, 2003), as in Table 1. Despite the importance of live commentary, professional commentators are not always available, resulting in most amateur or youth matches being left without live commentary. A promising solution is automatic live commentary generation through natural language generation techniques (Kubo et al., 2013; Taniguchi et al., 2019).

Most existing studies have focused on producing play-by-play commentary that describes visible events in the video, treating this task as a variant of Dense Video Captioning (DVC) (Krishna et al., 2017). However, actual live commentary often contains *background information* (color commentary) such as player profile, historical context, and stats. For example, the last utterance in Table 1 informs the audience that a Brazilian player Firmino scored many goals and had many assists in his previous team, Hoffenheim. Such background information adds more value to the commentary by providing context and insight.

time	live commentary
77.32	*Victim of that somewhat muscular early challenge from Lucas.*
89.97	*Good pressure from Firmino.*
92.60	*Milner breaking into the penalty area.*
96.71	*Here's Firmino again.*
99.01	*Well, he scored plenty of goals and had loads of assists playing in the Bundesliga for Hoffenheim.*

Table 1: An excerpt from a live commentary for the match Liverpool vs. Chelsea in 2015/2016 season of English Premier League. Transcribed from the match video in SoccerNet-v2 (Deliege et al., 2021).

Based on the above observation, we propose a new football commentary system that provides background information as well as play-by-play commentary.[1] Figure 1 shows the overview of our system.[2] Our system consists of three modules: **(i) Video Analysis.** This module first processes an input video to detect the players and the ball shown in the video frames, identifies the type of a play event. **(ii) Spotting.** This module conducts timing identification, which determines when a commentary utterance should be generated, and then predicts the utterance type indicating whether to provide play-by-play commentary or background information. **(iii) Commentary Generation.** This module produces play-by-play or background information commentary following the result of the spotting. To generate the latter, our system uses a retrieval-augmented generation (RAG) frame-

[1]Demo video: https://drive.google.com/file/d/1aneUywfYdrKrrZDSU5gEHdRWXj-eO8jN/view?usp=drive_link

[2]Code repository: https://drive.google.com/drive/folders/1_EqBtLr9YCnRDlB4IS9p69PhnZffTmPx

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 394–404

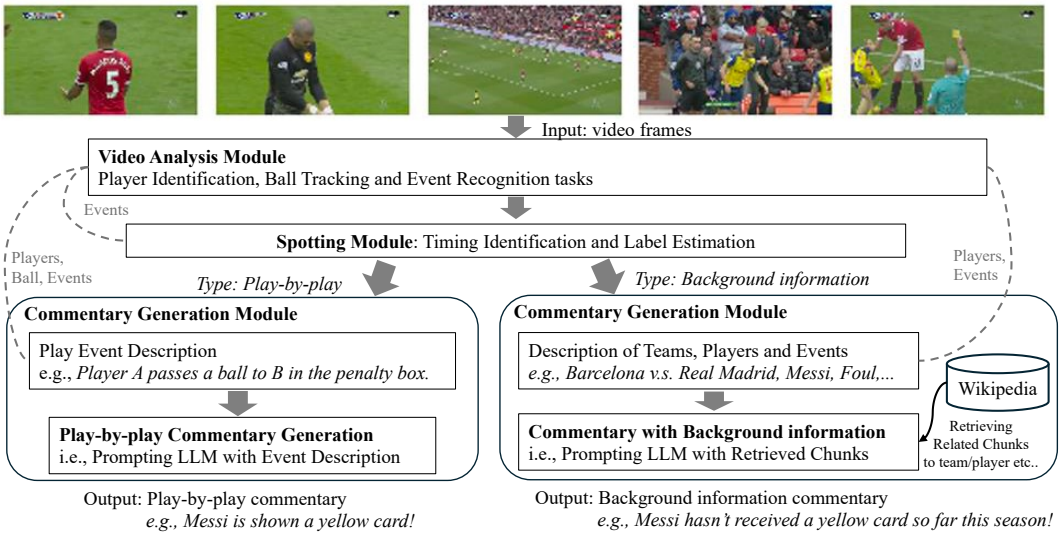

Figure 1: Architecture of our commentary generation system. First, the system analyzes the video to extract visual information. Then, it identifies the timing to start an utterance and predicts the utterance type (play-by-play or background information). Finally, the system generates commentary corresponding to the utterance type.

work (Lewis et al., 2020) to retrieve the relevant information from external knowledge sources such as Wikipedia.

2 Related Work

Text generation for (e-)sports has been explored for football (Tanaka-Ishii et al., 1998; Mkhallati et al., 2023; Kubo et al., 2013; Taniguchi et al., 2019; Oshika et al., 2023; Qi et al., 2023), baseball (Kim and Choi, 2020), and video games (Ishigaki et al., 2021; Wang and Yoshinaga, 2024). These studies are primarily categorized into two streams: 1) generating live text updates that can be read typically on webpages (Mkhallati et al., 2023; Kubo et al., 2013; Taniguchi et al., 2019; Oshika et al., 2023), and 2) generating commentary that can be shown as subtitles or replayed as audio commentary accompanying the video (Qi et al., 2023; Ishigaki et al., 2021; Wang and Yoshinaga, 2024; Kim and Choi, 2020). Our system belongs to the latter.

Different types of modality have been used for commentary generation, including structured data (Taniguchi et al., 2019; Wang and Yoshinaga, 2024), video-based inputs (Mkhallati et al., 2023; Kim and Choi, 2020), or their combination (Ishigaki et al., 2021; Qi et al., 2023). Our system focuses on videos as the primary input.

In sports video-to-text generation, two types of output targets have emerged. One produces a single text per video clip (Qi et al., 2023), while the other generates multiple time-stamped text segments (Mkhallati et al., 2023; Ishigaki et al., 2021), which requires both timing identification and content generation. We focus on the task of jointly identifying timing and generating text.

Generating color commentary or background information has also been explored in some pieces of work. Lee et al. (2014) approached color commentary for baseball as an information retrieval task by representing the game state as a feature vector to retrieve episodes, where the timing for commentary was given in their work. Andrews et al. (2024) attempted to design a system that automatically generates both play-by-play commentary and color commentary for football broadcasts, employing a queue-based approach for (utterance timing, content, and priority). Although their work is most similar to ours, there are significant differences. First, their system did not try to identify the timing of commentary, nor decide the type of the utterance for the timing, i.e., play-by-play or color commentary. Secondly, manual labelling was involved in the video analysis in their work, e.g., player names. Thirdly, the background information used in their system was restricted to game-level stats and season-level stats. In contrast, our system aims for timing identification, utterance type prediction, and a flexible framework for using external knowledge through the RAG framework.

3 Preliminary Analysis

To better understand the characteristics of football commentary, we conducted a preliminary analysis using broadcast videos from SoccerNet-

v2 (Deliege et al., 2021). We first transcribed the audio commentary with WhisperX (Bain et al., 2023), and then used a large language model to automatically label each utterance with a type indicating whether it contains background information (see Appendix A for details). The resulting dataset, which we refer to as the Live Football Commentary with Background Information (LFCBI) Dataset, serves as the basis for our subsequent research. Our manual analysis revealed that approximately 18% of the utterances contain background information. Moreover, as can be seen in Table 6, our analysis showed that the ratio of background information usage is significantly higher in the vicinity of out-of-play events (i.e., events that momentarily interrupt the flow of play, such as fouls, goals, and substitutions). This finding is consistent with the observation that commentators often provide supplementary details during pauses in play to sustain viewer engagement. These insights have guided the design of our system, particularly the timing mechanism for injecting background information into the commentary.

4 System Architecture

We describe the architecture of our system. Its overview is shown in Figure 1.

The system generates commentary for a match video through the following three stages: `(i) Video Analysis`: The system detects and tracks players and the ball in the video and extracts necessary information, such as the list of players and ball coordinates. `(ii) Spotting`: Based on the end time of the previous utterance, the system predicts the timing for the next utterance and decides whether to provide background information or play-by-play commentary. `(iii) Commentary Generation`: The system generates play-by-play commentary as well as commentary for background information.

4.1 Video Analysis

By using player identification and ball tracking techniques, the system extracts each player's position and team affiliation, and jersey number, as well as the ball's trajectory from the input video.

Player Identification: We adopted the player tracking tool proposed by Somers et al. (2024), which combines multi-object tracking, team classification, and OCR-based jersey number recognition, to extract the location, team affiliation, and jersey number of players in the video. Note that the jersey number may be missing. When a jersey number was successfully recognized, we automatically assigned the player's name by using a mapping between jersey numbers and player names, provided by SoccerNet-Caption (Mkhallati et al., 2023).

Ball Tracking: Our system first employed yolov8[3] (Jocher et al., 2023) to detect ball candidates. Next, using Dijkstra's algorithm, we constructed a smooth trajectory that reflects information from previous frames.[4] Specifically, for each candidate, a cost was computed as the weighted sum of the distance and confidence from the ball candidate in the previous frame, and the trajectory with the minimum cost was selected. Subsequently, missing segments up to 25 frames (1 second) were linearly interpolated using coordinates from adjacent frames, and the ball coordinates were merged with the player identification results. Since there are cases where the broadcast switches to a camera that zooms in on players, ball position data may contain missing values.

Play Event Recognition: To generate play-by-play commentary, we first performed dense event detection using T-DEED (Xarles et al., 2024), the first-place model from the 2024 BAS-Challenge (Cioppa et al., 2024). This system detects Ball-Action events (e.g., pass, drive, out, etc.) in the video and extracts both their occurrence times and spatial locations (e.g., left top corner, left top midfield).

4.2 Spotting

This module predicts the timing of the next utterance and predicts the utterance type: play-by-play commentary or background information.

Timing Identification: In our timing identification, given the end time of the last utterance, we determined the start time for the next utterance. A simple sampling-based method was employed.[5] Specifically, we estimated the empirical distribution over the utterance intervals,[6] as shown in Figure 3, and sampled an interval from the estimated distribution.

[3]To capture even balls blurred by motion, the confidence threshold was set relatively low (0.3).

[4]Inspired by https://www.kaggle.com/competitions/dfl-bundesliga-data-shootout/discussion/360097.

[5]Prior work (Ishigaki et al., 2021; Mkhallati et al., 2023) mentioned that predicting utterance timing is challenging.

[6]When estimating the empirical distribution, intervals longer than 4 seconds were excluded to avoid unnatural gaps.

Utterance Type Prediction: At the identified utterance timing obtained above, the system decides whether to generate play-by-play commentary or background information. Based on the analysis of events presented in Appendix A, our system adopted the following rule-based method. If an out-of-play event occurs within the 15 seconds preceding the identified utterance timing, our system selected background information with 50% probability.[7] Otherwise, at other timings, a Bernoulli distribution with probability p was sampled to decide whether to generate background information. In this study, p was set to the proportion of background information in the overall commentary (18%).

4.3 Commentary Generation

Depending on the outputs from the video analysis and spotting, the system generated either of the two types of commentary: play-by-play or background information.

4.3.1 Play-by-play Commentary

First, using the results obtained in Section 4.1, we created a `play event description` by filling in a template with items (action, involved player(s), team, ball position). Here, the player closest to the center of the bounding box of the ball was regarded as the player involved in the play. For example, if the involved player is identified as *Aubameyang* through the recognized jersey number and the team classification, the play event description becomes "Aubameyang passed from Left bottom corner." If the jersey number was not recognized, the description is "An Arsenal player passed from Left bottom corner."

Next, we converted this play event description into a play-by-play commentary in a commentary style using gpt-4o (OpenAI et al., 2024). For instance, "An Arsenal player passed from Left bottom corner." might be transformed into "*The Arsenal player delivered a stunning pass from the bottom left corner with precision and flair!*", thus yielding text with expressive and emotive language.

4.3.2 Background Information Commentary

Leveraging the RAG framework, we extracted content related to the video matching the situation from external knowledge, and fed it into a large language model to generate commentary that includes background information.

We used Wikipedia articles as the external knowledge source. Since we focus on player information, we collected a total of 3,257 Wikipedia pages concerning players.[8] The query for document retrieval contains the following:

- Detailed match information (e.g., match schedule, competing teams).
- The list of players recognized by the player identification module in the 2-second period preceding the identified utterance timing.
- Recent commentary within the past 60 seconds.
- Event information in the 15-second period preceding the identified utterance timing.

All information used to construct the query, as well as the extracted related documents, were included in the prompt to the large language model with instructions to generate commentary in a live commentary style. Details of the query and the prompt are provided in Appendix C.

For implementation, we used LangChain (Chase, 2022). The external knowledge was chunked every 1,000 characters, with a 100-character overlap between adjacent chunks. During search, instead of relying solely on word matching, text embeddings (text-embedding-ada-002 (OpenAI)) were used to capture semantic relatedness flexibly. During document retrieval, up to 10 chunks with cosine similarity above 0.7 to the query embedding were retrieved in the descending order of the similarity. Finally, the background information commentary was generated by feeding these retrieved documents into gpt-4o (OpenAI et al., 2024).

4.4 Interface and System Workflow

The system is executed through a web browser-based interface by specifying the match information (match identifier, start and end times), and outputs a video containing automatically generated audio commentary and subtitles. The process is divided into two stages. (i) Offline processing runs the video analysis module to extract players, a ball, and events. (ii) Online processing invokes the spotting and commentary generation modules on demand when the GUI request arrives. It operates once per requested segment and is therefore not a continuous streaming pipeline. As part of the on-

[7] For kick-off, if the identified timing is within 15 seconds before or after, background information was always selected.

[8] Wikipedia contains numerous interesting pieces of information for general soccer viewers, such as unique records and notable statistics.

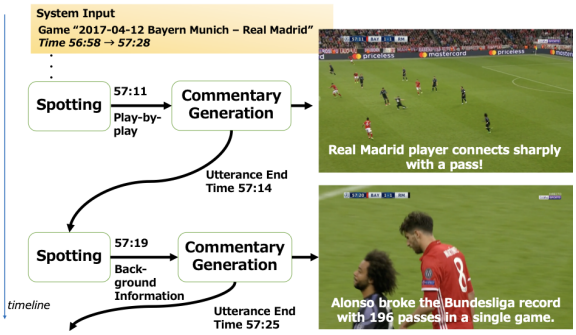

Figure 2: System workflow.

line processing stage, and as shown in Figure 2, the system generates commentary within the specified match time range through the following steps:

1. Initialize it by setting the end time of the previous utterance to the match start time, and use the match end time as the loop termination condition.
2. Predict the next utterance timing and type.
3. Generate play-by-play or background information commentary depending on the type.
4. Calculate the end time of the generated commentary based on its length (assuming a speaking rate is 200 words per minute), and add it to the comment history.
5. Repeat steps 2–4 until the end time is reached.
6. Output the generated commentary as an SRT subtitle file and synthesize the text into speech using OpenAI TTS,[9] overlaying it on the video to produce the final output.

5 Evaluation

To verify whether our system can effectively provide live commentary, we conducted the following three evaluations, focusing on the new feature of our system, background information commentary: (i) Automatic evaluation of the timing identification and utterance type prediction. (ii) Human evaluation of the commentary that includes background information. (iii) Inference time profiling to identify bottlenecks and quantify the gap with the real-time operation.

5.1 Evaluation for Spotting

5.1.1 Baselines

For timing identification, we compare the following two methods:

- `Our System`: A method that identifies the utterance timing by sampling from an empirical distribution over utterance intervals.
- `Baseline (Fixed Interval)`: A method that identifies the utterance timing by using a fixed interval equal to the average silence time between consecutive utterances (2.14 seconds).

For utterance type prediction, we compare the following two methods:

- `Our System (50% probability for out-of-play)`: If an out-of-play event occurs within 15 seconds prior to the identified timing, background information is selected with 50% probability; otherwise, background information is generated with a probability of 18%.
- `Baseline (Ignoring out-of-play)`: This method always generates background information with a probability of 18%.

5.1.2 Evaluation Metrics

The evaluation used the LFCBI Dataset (68,919 instances from the test set of the train:valid:test split) containing utterance start and end times and utterance types. For timing identification, we used the mean squared error (MSE) between the identified and the ground-truth utterance start times.

For utterance type prediction, we used Precision, Recall, and F_1. These metrics were calculated for background information. We report results for two settings when performing utterance type prediction: (1) using the identified utterance timing from our system (`Pred`) and (2) using the ground-truth utterance timing (`Gold`). These metrics were computed by averaging the results obtained for 10 different random seeds.

5.1.3 Results

For timing identification, the MSE of our system was 15.50, while the baseline achieved an MSE of 19.86. This result suggests that our simple sampling-based method from the empirical distribution effectively captures the natural tempo of the commentary, resulting in relatively more accurate timing prediction than using fixed intervals.

Table 2 shows the evaluation results for utterance type prediction. As seen in the table, our system achieves higher Precision, Recall, and F_1 scores than the baseline. Furthermore, using the ground-truth timing (`Gold`) slightly improves the F_1 score, suggesting that timing identification errors adversely affect utterance type prediction.

[9] https://platform.openai.com/docs/guides/text-to-speech

Method	Timing	Precision	Recall	F_1
Baseline	(1) Pred	18.76	17.07	17.88
Baseline	(2) Gold	18.32	17.45	17.87
Our System	(1) Pred	19.64	26.83	22.67
Our System	(2) Gold	19.87	27.08	22.93

Table 2: Automatic evaluation results of utterance type prediction for Baseline (ignoring out-of-play) and Our System (50% probability for out-of-play). For timing, Pred uses the identified utterance timing from our system, while Gold uses the ground-truth timing.

However, even with the ground-truth timing, an F_1 score of 22.93% indicates that the performance is not yet sufficient. In real commentary, a commentator deeply understands the match and predicts that the match will not see a major change, and thus provides background information accordingly. Achieving higher utterance type prediction performance may require a more refined modeling of the match progression.

5.2 Human Evaluation of Background Information Commentary

5.2.1 Compared Models

We evaluated the generated commentary from the following four approaches:

- `(Baseline)` A baseline method where gpt-4o generates background information on its own without any document retrieval.
- `(Our system)` Our system that uses both the player identification module and the RAG framework.
- `(Our system w/ GP)` Our system with the ground-truth list of players (GP; Gold Players) from the frame used in place of the output from our player identification module.
- `(Our system w/ GP & GD)` Our system with the ground-truth players (GP) and the ground-truth background information documents (GD).

Comparing these approaches allows us to evaluate the effectiveness of the RAG framework and to understand the potential of our system when individual modules operate ideally versus its current limitations.

5.2.2 Evaluation Methods

We employed human evaluation with three evaluators; two regularly watch football matches and one watches international matches about once a year. The evaluators rated each generated commentary on three aspects with a three-point scale: relevance, usefulness, and an overall evaluation (see Appendix D for details). A total of 20 instances were evaluated. The utterance timings for the evaluated instances were taken from the start times of the commentary labeled as background information in the LFCBI Dataset. During the evaluation, real commentary containing background information was not provided as a reference example, as our evaluation does not focus on similarity to a single reference. This setting is based on the idea that, although only one correct example might be available, many acceptable alternatives exist, and we aim to prevent evaluators from being biased by a single reference. For Our system w/ GP, the ground-truth player list (GP) is taken from the labels included in SoccerNet-v3 (Cioppa et al., 2022). Gold Documents (GD) are excerpts from the articles gathered from the internet by referring to the commentary labeled as background information in the LFCBI Dataset.

Method	Relevance	Usefulness	Overall
Baseline	1.68	2.15	1.87
Our system	1.73	2.26	1.95
Our system w/ GP	2.05	2.36	2.11
Our system w/ GP & GD	2.55	2.40	2.36

Table 3: Human evaluation on background information commentary.

5.2.3 Results

Table 3 shows the average evaluation scores for each approach.[10] As shown in Table 3, the baseline scored the lowest in relevance, usefulness, and overall ratings. In contrast, our system outperformed the baseline, suggesting that incorporating the RAG framework helps provide commentary with content that is related to the video context and engaging. Moreover, our system w/ GP scored higher than our system in all evaluation metrics, indicating that the accuracy of the player identification directly affects the quality of the generated commentary. Furthermore, our system w/ GP & GD achieved the highest scores in all metrics, with relevance reaching 2.55 and usefulness 2.40. This result shows that, in addition to accurate player information, having richer and more appropriate external knowledge enables more effective provision of background information.

[10]The average inter-evaluator weighted kappa coefficient (Fleiss and Cohen, 1973) was 0.61 for Relevance, 0.23 for Usefulness, and 0.25 for Overall, respectively.

Component	Time
Offline: Video Analysis	
Player-identification subtotal	5:26
Object detection	0:50
Tracking	0:42
Jersey OCR	3:54
Ball tracking[11]	0:04
Play-event recognition	0:25
Offline total	5:55
Online: Commentary Generation	
Retrieval with GPT-4o	0:10
TTS synthesis (200 wpm)	0:17
Online total	0:27

Table 4: Inference time of whole processes (min:s). In the retrieval with GPT-4o, we aggregate three background information calls for the 30 s clip.

5.3 Inference Time

We measure inference time for the two–stage architecture described in Section 4.4: online and offline processing stages. Table 4 shows the inference time of the most time-consuming components on a single NVIDIA TITAN 24GB. The offline processing is dominated by jersey-number OCR within the player-identification stage, reproducing the bottleneck reported by Somers et al. (2024). In the online stage, most of the latency comes from retrieval with GPT-4o and from TTS synthesis. Both components must be accelerated to achieve real-time performance.

6 Conclusion

We have constructed a football commentary system that provides background information as well. Our evaluation indicates that the system can present background information related to the video through the use of external knowledge, even though limitations in the player identification module sometimes lead to the insertion of information that is less relevant to the match situation.

Limitations

Real-Time Performance The Video Analysis Module and Commentary Generation Module used in this study do not operate at real-time speeds; therefore, commentary for the demo video was generated in advance. To support real-time generation, it will be necessary to develop lightweight models for player identification and to speed up both document extraction and commentary generation using large language models.

[11] Ball positions are estimated by reusing the player tracking tool's output, so the additional computation is minimal.

Further Diversification of Background Information Although this study primarily focuses on players' background knowledge and statistics, the actual scope of background information in live commentary should be broader. For example, topics such as the history of coaches and referees, team backgrounds, stadium characteristics, as well as explanations of rules and tactical analysis, are all part of the commentary landscape (Tanaka-Ishii et al., 1998). It will be necessary to systematically categorize these topics and provide content optimized according to the viewers' preferences and interests.

Spotting Module Design The rule-based spotting module ignores game dynamics and may be vulnerable to distribution shift. Future work will investigate event-driven approaches and learned policies to better anticipate play transitions, while carefully validating their robustness.

Evaluation Scope Each demonstration video for evaluation covers only a 30s segment. The factual accuracy of background information is unmeasured. The human study involves three raters, which can be increased for better reliability. These issues require further improvement.

Ethics Statement

We outline the following potential ethical concerns: (i) Human evaluation was conducted in Japan; annotators were compensated above the local minimum wage. (ii) Audio commentary is synthesized with the OpenAI TTS engine; no real commentators' voices are cloned or imitated. (iii) Background information provided by our system may contain inaccuracies; future work will add fact-checking and a user feedback channel for corrections.

Acknowledgments

This work was supported by a grant from the New Energy and Industrial Technology Development Organization (NEDO) under the project (JPNP20006).

References

Peter Andrews, Oda Elise Nordberg, Njål Borch, Frode Guribye, and Morten Fjeld. 2024. Designing for automated sports commentary systems. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, IMX '24, page 75–93,

New York, NY, USA. Association for Computing Machinery.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain. Software available from GitHub.

Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2022. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific Data*, 9(1):355. Published: 2022/06/21.

Anthony Cioppa, Silvio Giancola, Vladimir Somers, Victor Joos, Floriane Magera, Jan Held, Seyed Abolfazl Ghasemzadeh, Xin Zhou, Karolina Seweryn, Mateusz Kowalczyk, Zuzanna Mróz, Szymon Łukasik, Michał Hałoń, Hassan Mkhallati, Adrien Deliège, Carlos Hinojosa, Karen Sanchez, Amir M. Mansourian, Pierre Miralles, and 65 others. 2024. Soccernet 2024 challenges results. *Preprint*, arXiv:2409.10587.

Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. 2021. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4508–4519.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. Ultralytics yolo. Released on 2023-01-10. Licensed under AGPL-3.0. Repository: https://github.com/ultralytics/ultralytics.

Byeong Jo Kim and Yong Suk Choi. 2020. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1056–1065.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2013. Generating live sports updates from twitter by finding good reporters. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 527–534. IEEE.

Greg Lee, Vadim Bulitko, and Elliot A. Ludvig. 2014. Automated story selection for color commentary in sports. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(2):144–155.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5073–5084.

OpenAI. New and improved embedding model: text-embedding-ada-002. https://openai.com/index/new-and-improved-embedding-model/text-embedding-ada-002. Accessed: 2025-01-02.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and 1 others. 2024. Gpt-4o system card. https://doi.org/10.48550/arxiv.2410.21276. *Preprint*, arXiv:2410.21276. [arXiv:2410.21276].

Masashi Oshika, Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2023. Transformer-based live update generation for soccer matches from microblog posts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10100–10106, Singapore. Association for Computational Linguistics.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, and Jie Tang. 2023. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 5391–5395, New York, NY, USA. Association for Computing Machinery.

M. Schaffrath. 2003. Mehr als 1:0! bedeutung des live-kommentars bei fußballübertragungen– eine explorative fallstudie [more than 1:0! the importance of live commentary on football matches – an exploratory case study]. *Medien und Kommunikationswissenschaft*, 51.

Vladimir Somers, Victor Joos, Silvio Giancola, Anthony Cioppa, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe De Vleeschouwer. 2024. SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In *2024 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*.

Kumiko Tanaka-Ishii, Koiti Hasida, and Itsuki Noda. 1998. Reactive content selection in the generation of real-time soccer commentary. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating live soccer-match commentary from play data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7096–7103.

Zihan Wang and Naoki Yoshinaga. 2024. Commentary generation from data records of multiplayer strategy esports game. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 263–271, Mexico City, Mexico. Association for Computational Linguistics.

Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. 2024. T-deed: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3410–3419.

A Dataset Analysis

In this section, we investigate the differences in the distributions of silence durations and the variations in the usage rates of commentary containing background information around various events in the LFCBI Dataset. Our analysis reveals that there is little difference between the silence duration distributions for play-by-play commentary and background information commentary and that the usage rate of background information is significantly higher following specific events.

A.1 Dataset Overview

Table 5 shows the components of the data contained in the LFCBI Dataset. There are 338,034 instances in this dataset. Each instance consists of match information, an English transcription of the commentary generated by an automatic speech recognition model, along with the corresponding utterance intervals and the types automatically labeled by a large language model. Note that the types are binary: play-by-play commentary and background information.

Item	Example
Match Information	{date: 2015-08-29, time: 19-30, team_1: Bayern Munich, team_2: Bayer Leverkusen, score: 3 - 0}
Utterance Interval	14:06 – 14:22
Comment	It's a game that we brought you here on BT Sport, and it was a stunning performance from Roger Schmid's side to see off the Italians from 1-0 down in the first leg.
Type	background information

Table 5: An example from the LFCBI Dataset. The match information includes the match date, start time, and team names. Comments that include background information are labeled as background information, whereas those that do not are annotated as play-by-play commentary.

A.2 Distribution of Silence Durations

In this section, we compare the distributions of silence durations to capture the natural intervals and timing tendencies observed in actual live commentary to reflect these tendencies in our system's utterance timing identification. Figure 3 shows the distribution of silence durations measured from the end of the preceding utterance. As the figure shows, there is little difference between the silence duration distributions for play-by-play commentary and for background information commentary.

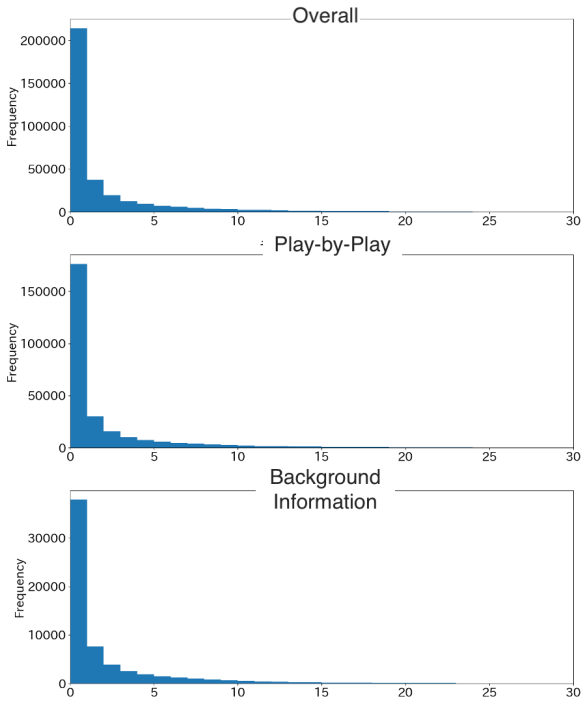

Figure 3: Distribution of silence durations (up to a maximum of 30 seconds). The horizontal axis represents the silence durations binned in one-second intervals, and the vertical axis shows the frequency. The top panel shows the overall distribution; the middle panel shows the distribution for utterances labeled as play-by-play commentary following the preceding silence; and the bottom panel shows the distribution for utterances labeled as background information following the preceding silence.

This suggests that even when inserting background information, commentators do not require significantly longer pauses; they tend to maintain a similar rhythm as when delivering play-by-play commentary.

A.3 Relationship between Event Types and Utterance Timing

SoccerNet-v2 provides annotations for event labels[12] along with their occurrence times. We utilize these annotations by associating commentary utterances whose start times fall within 15 seconds before or after an event timestamp with the corresponding event. This 15-second window is chosen under the assumption that commentators first refer to an event immediately after it occurs and then provide background information. Our analysis shows, as summarized in Table 6, that the usage rate of background information in the vicin-

[12]These include events such as corner kicks, yellow card presentations, goals, substitutions, etc.

Event	15 sec before	15 sec after
Ball out of play	15.31	21.33
Clearance	23.26	20.88
Corner	23.20	17.01
Direct free-kick	21.97	18.76
Foul	16.45	24.41
Goal	14.39	32.06
Indirect free-kick	24.05	18.56
Kick-off	27.21	27.05
Offside	16.58	26.88
Penalty	34.18	46.03
Red card	30.34	30.51
Shots off target	12.88	18.58
Shots on target	13.26	22.58
Substitution	22.35	25.06
Throw-in	20.93	17.90
Yellow card	25.60	31.50
Yellow → red card	22.64	36.59

Table 6: Usage ratio of commentary containing background information within 15 seconds before and after events. Bold entries indicate out-of-play events and Kick-off.

ity of events tends to be higher than the overall usage rate of 18%. In particular, events that temporarily interrupt the flow of the match (e.g., `Foul`, `Goal`, `Penalty`, `Red card`, `Yellow card`, `Yellow → red card`, `Substitution`, `Offside`, `Ball out of play`) exhibit a markedly higher usage ratio of background information after the event. This suggests that commentators tend to provide reflective or supplementary information at moments when the play has momentarily paused. Here, we call such events *out-of-play events*.

B Performance of Player Identification

In the task of tracking players within a video, Somers et al. (2024) not only introduced a baseline method but also proposed a new evaluation metric for the player tracking task called **GS-HOTA**. GS-HOTA measures the ability to accurately estimate and continuously track, in sports such as soccer, the players (and referees) on the pitch, including their positions, roles (e.g., field player, goalkeeper, referee), jersey numbers, and team affiliations. The baseline method reported by Somers et al. achieved a GS-HOTA score of 22.26%, which can be interpreted as indicating that approximately 22.26% of the frames have all players correctly identified. Although it would be ideal for our system to perfectly recognize every player in each frame, it is capable of generating the minimum necessary background information even when player identification is imperfect. Therefore, for our use case, the player identification can be considered sufficiently functional even if it is not flawless[13].

C Prompt for Background Information Commentary

After executing the player identification and completing document extraction, the following is an example of a prompt used to generate commentary that includes background information with a large language model.

```
You are a professional color commentator for a live
     broadcast of soccer.
Using the documents below,
provide just one comment with a fact, such as player
      records or team statistics, relevant to the
     current soccer match.
The comment should be short,  clear, accurate, and
     suitable for live commentary.
The game date will be given as YYYY-MM-DD. Do not
     use information dated after this.
This comment should be natural comments following
     the previous comments given to the prompt.
===documents
Julian Weigl
Weigl with Benfica in 2021
Personal information Date of birth: 8 September 1995
      (age 29) Place of birth: Bad Aibling, Germany
     Height: 1.86 m (6 ft 1 in) Position(s):
     Defensive midfielder Team information
........
===
Recent Event: Foul
Game: germany_bundesliga germany_bundesliga
     2016-09-10 RB Leipzig vs Dortmund
Players shown in this frame: Halstenberg M. from RB
     Leipzig, Piszczek L. from Dortmund, Weigl J.
     from Dortmund
Previous comments: Losermann. Pause against Piszczek
     . Heizenberg is with him from behind. Now it's
     too late. Castro came back today. Pause.
     Heizenberg.

Comment:
```

The text enclosed between "===documents" and "===" consists of the documents obtained from document extraction. Note that during document extraction, the text between "===" and "Comment:" is used directly as the search query.

D Details of the Human Evaluation

Below, we outline the evaluation criteria and the standards for each rating dimension. **Relevance:** To what extent does the content of the utterance provide information related to the events (e.g., goals, fouls, passes, etc.) or the players shown in the video; **Usefulness:** How useful is the content of the utterance for the general viewer in terms of providing new knowledge; **Overall:** Considering both relevance and usefulness, to what extent does the utterance enhance the viewer's interest in the match or the players.

[13]Note that the videos used by Somers et al. for computing GS-HOTA were captured with a single camera, whereas our videos include multiple camera switches. Consequently, it should be noted that the baseline method might not achieve a GS-HOTA of 22.06% on our videos.