Automated Multi-Modal Pipeline for Generating Football Game Summaries

Bibek Khanal
Department of Electronics and Computer Engineering
IOE Thapathali Campus
Kathmandu, Nepal
tha075bct016@tcioe.edu.np

Ashok Budha
Department of Electronics and Computer Engineering
IOE Thapathali Campus
Kathmandu, Nepal
tha075bct013@tcioe.edu.np

Aashish Chapain
Department of Electronics and Computer Engineering
IOE Thapathali Campus
Kathmandu, Nepal
tha075bct002@tcioe.edu.np

Kapil Shrestha
Department of Electronics and Computer Engineering
IOE Thapathali Campus
Kathmandu, Nepal
tha075bct022@tcioe.edu.np

Umesh Kanta Ghimire
Department of Electronics and Computer Engineering
IOE Thapathali Campus
Kathmandu, Nepal
ukg@tcioe.edu.np

***Abstract*— Football, being one of the most globally renowned sports, garners an immense amount of viewership worldwide. In today's fast-paced world, viewers often prefer concise highlights over watching entire games. However, the traditional method of generating text-based highlights is laborious and time-consuming, relying heavily on manual efforts. To overcome these limitations, we present an innovative approach that leverages a multi-modal pipeline for automated football game summary generation. The explosion of artificial intelligence and the increasing availability of sports-related data, including game audio, metadata, and captions, offer a promising opportunity to employ machine learning techniques in creating a unified multi-modal pipeline for summary generation. Our proposed methodology involves two key components: game audio summarization and metadata summarization. For game audio summarization, we employ a selector and rewriter model to generate a comprehensive summary from the audio input. On the other hand, metadata summarization utilizes a naive templating engine in conjunction with the powerful Longformer model to produce a summary based on the captions dataset. By integrating these approaches and employing multiple data modes, we introduce an automated method for generating football game summaries. We evaluate our approach using the Recall Oriented Understudy for Gisting Evaluation (ROUGE) scores and demonstrate that our results are comparable to the current state-of-the-art research. These findings validate the efficacy of our multi-modal approach for automated text summaries and highlight the potential it holds for improving summary generation in the field of football analysis and beyond.**

Keywords—Audio, Metadata, Captions, Deep Learning, Longformer, Metadata, Multimodal, Selector, Rewriter, ROUGE

I. INTRODUCTION

Football gathers a substantial global following, with FIFA reporting that a staggering 1.5 billion individuals witnessed the live broadcast of the 2022 World Cup Final. Similarly, the 2018 World Cup Final, featuring France and Croatia, captivated 1.1 billion television viewers. The fervor for the sport extends well beyond the live matches, as fans and analysts consistently seek avenues to relive the excitement and maintain connections with their cherished teams and players. In this regard, game summaries play a pivotal role by effectively condensing hours of gameplay into concise information, thereby fulfilling this inherent need.

In the realm of football match summarization, traditional practices involve manual efforts by journalists and sports analysts. These professionals invest significant time watching games, extracting key information, and crafting summaries based on their observations. However, advancements in technology have given rise to the demand for automated pipelines capable of generating concise summaries from multimodal inputs. While existing automated approaches rely on data such as play-by-play descriptions, match reports, or social media captions, there is a growing potential to incorporate game audio as a valuable source of information. Football, being a dynamic sport with elements like crowd reactions and commentators' remarks, presents an opportunity to enhance summarization techniques by leveraging the context-rich audio data available.

Our research proposes an innovative approach to generate concise and informative game summaries by leveraging multiple modalities, including captions, metadata, and game audio. By incorporating captions, we establish a narrative backbone that captures textual descriptions of the event. Additionally, the integration of metadata enriches the summaries with contextual information such as team names and game events. To add a layer of dynamism, we analyze game audio, incorporating elements like commentator remarks and crowd reactions into the summaries. Our approach utilizes natural language processing and deep learning techniques to create an automated multi-modal pipeline capable of generating efficient and concise game summaries.

II. BACKGROUND AND RELATED WORK

In recent years, the application of machine learning techniques has greatly facilitated the task of text summarization [1]–[3]. Text summarization aims to concisely capture the main ideas of a document while reducing its length. Sports summarization, as a specialized form of text summarization, presents unique challenges that set it apart from general text summarization. Specifically, the use of repetitive language by commentators to describe significant events, along with the distinct reactions of home and away crowds to similar events, adds complexity to the task [4].

The generation of summaries from sports game audio poses an additional hurdle due to the dynamic nature of the audio data. The presence of dynamic and challenging-to-isolate noises further exacerbates the difficulty. Notably, machine learning techniques have proven to be more effective than traditional methods such as Low Pass filtering [5], Wiener Filtering [6], and Time-Frequency Block thresholding [7] for audio denoising [8] [9].

With the advancements in machine learning and its application to document summarization and audio denoising, several approaches have emerged for sports game summarization. Zhang et al. [4] proposed an automatic generation method for sports news articles using live text commentary. They leveraged real-time updates from the commentary and transformed them into comprehensive news articles. Huang et al. [10] took a distinct approach to football game summarization by employing live game text and introducing a scorer and rewriter model to generate text summaries. Wang et al. [11] introduced a knowledge-enhanced approach that integrated player and team information with live text commentary. Additionally, they introduced an enhanced dataset, K-SportsSum, which included a broader range of game-related information. In SportsSum 2.0. [12] they addressed the noise-related issues that affected summarization performance. Furthermore, a reranked-enhanced summarizer was introduced to enhance the fluency and expressiveness of the generated news summaries.

Moreover, Van Der Lee et al. [13] presented the PASS framework, a data-to-text system capable of generating two sports reports in Dutch for each game, catering to both the home and away teams. The framework is designed to produce language that elicits an expected emotional response from individuals reporting on events they are emotionally invested in. The data used for summary generation within the PASS framework adheres to a specific, well-defined format.

Additionally, Gautam et al. [14] proposed a multimodal approach for football game summarization, leveraging diverse inputs such as game audio, captions, and game metadata. They devised a unified pipeline capable of generating three distinct types of summaries from these heterogeneous data sources.

These advancements in machine learning techniques and their applications in document summarization, audio denoising, and sports game summarization contribute to the broader field of natural language processing and hold promise for further improvements in summarization quality and accuracy.

III. DATA

Our research paper incorporates data from diverse sources, namely SportsSum, K-SportsSum, SoccerNet, and Goal. These datasets encompass a wide range of modalities and data types, including game audio, metadata, captions, and ground truth summaries. Due to the absence of a comprehensive dataset containing all modalities, we employed multiple datasets to ensure coverage and inclusiveness in our analysis.

A. SPORTSSUM

SportsSum [10] consists of a total of 5428 soccer game commentaries with corresponding news scrapped from online sources in Chinese. We have used an open-source dataset created by Gautam et al. [14]. The open-source dataset consists of soccer game commentaries with corresponding news in English.

The dataset consists of game commentaries in JSON data format with fields such as id (unique id of the event), s1 (goal scored by the home team), s2 (goal scored by away team), t (time in minutes of the event) and m (message describing the details of the events). The corresponding news file was provided in .txt file format.

B. K-SPORTSSUM

The K-SportsSum [11] dataset encompasses 7854 sports game summaries and live commentaries, supplemented by a knowledge corpus comprising background information on 523 sports teams and over 14,000 sports players, all in Chinese. Initially introduced in Chinese, the K-SportsSum dataset was subsequently translated into English by Gautam et al. [14]. For our research, we utilized the English-translated variant of the dataset to fulfill our objectives.

C. SOCCERNET-V2

The SoccerNet-V2 dataset comprises match videos of 250 games, accompanied by encoded metadata. Each match is segmented into two parts: the first and second half. As our research objective focused solely on games in English, videos of matches in other languages were excluded from our analysis. Notably, the dataset encompasses videos from six distinct football leagues. The accompanying metadata, provided in JSON format, offers comprehensive information regarding 17 different game events.

D. GOAL

The GOAL dataset, introduced in the paper by Wang et al. [15] titled "GOAL: Towards Benchmarking Few-Shot Sports Game Summarization" consists of 103 pairs of commentary and news articles. The average word count for commentary articles is 2724.9, while news articles have an average length of 476.3 words. Notably, GOAL provides an additional set of 2160 unlabeled commentary documents; however, these documents were not utilized in our project. Remarkably, this dataset is unique in that it is the sole football-related dataset that includes pairs of news and commentary in the English language. The data extraction process involved sourcing information from the goal.com website, which serves as a repository for football match-related data. It is important to note that the GOAL dataset differs from other Chinese datasets in that it does not include temporal information associated with the events. Due to its smaller size and use of English, the GOAL dataset offers precise information without undergoing potential loss during translation from Chinese to English.

E. SOCCERNET EXTENSION DATASET

The SoccerNet extension dataset comprises a meticulously curated collection of summaries from 100 games. This dataset was manually created by our team, leveraging game audio and metadata available on SoccerNet. While SoccerNet provides comprehensive game-related information such as commentary, statistics, and reports, it does not include pre-existing summaries. Therefore, to evaluate the performance of our audio summarization model and metadata template, we generated the game summaries

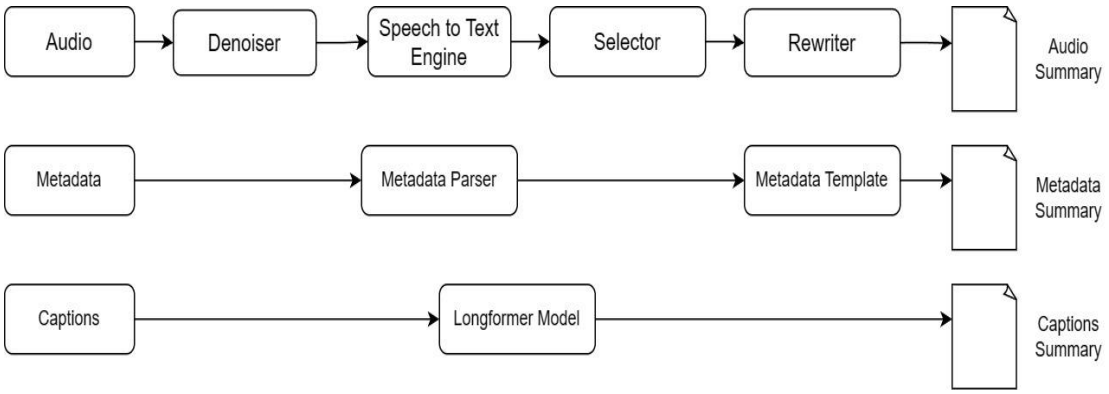

Fig. 1. Pipeline For Summary Generating Football Game Summaries using Audio, Metadata and Captions

using this dataset. The source of the original game data is espn.com, a highly regarded sports-based website renowned for publishing match commentary, statistics, reports, and other pertinent information.

IV. IMPLEMENTATION DETAILS

A. Caption Summarization with LongFormer

To generate summaries from captions, we employed the Longformer model, a state-of-the-art architecture for text summarization. For the fine-tuning process, we utilized a dataset extracted from the Goal, SportsSum, and K-SportsSum datasets, consisting of 12,000 caption and summary pairs. To accommodate the characteristics of our dataset, we made modifications to the Longformer model [16], configuring it to accept up to 5,120 input tokens, while setting the output token size to 1,024.

One notable feature of the Longformer model is its implementation of windowed self-attention, which substantially reduces the computational complexity involved. By transforming the quadratic complexity of self-attention in relation to the input length into a linear complexity, the model can effectively handle longer documents without sacrificing contextual information. This improvement in self-attention allows for a more comprehensive summarization of lengthy captions, minimizing the loss of essential details.

By employing the Longformer model with adapted configurations, we achieved enhanced summarization performance and successfully generated concise summaries from the captions in our dataset.

B. Summaries From Metadata Parser

In order to convert metadata into coherent and intelligible sentences, we developed a template engine. This engine enables the transformation of structured metadata into natural language sentences that accurately convey the corresponding information.

Sample Metadata template for SoccerNet-V2 dataset

```
{
        "gameTime" : "[ half number ] – [ Time ]",
        "label" : “[ action type ]",
        "position" : "[ time in ms ]",
        "team" : "[ home / away / not applicable ]",
        "visibility" : "[ visible / not shown ]"
}
```

From the sample metadata the template engine generates static sentences like "Team <home> throws in the ball". The value for <home> is extracted from the metadata.

C. Summary Generation From Game Audio Commentary

1) Denoising using Deezer Spleeter

Generating accurate summaries from game audio commentary presents a significant challenge due to the dynamic and difficult-to-isolate noises inherent in the audio data. To mitigate this issue, we employed Spleeter, an open-source library developed by Deezer, renowned for its noise removal capabilities [17]. Spleeter facilitated the separation of noise from the game audio, thereby enhancing the quality of the audio input for subsequent summarization processes. By leveraging Spleeter's advanced noise removal algorithms, we successfully attenuated the impact of background noise, ultimately improving the overall performance and reliability of our summarization system. The integration of Spleeter as a noise reduction tool underscores our commitment to employing cutting-edge technologies to address the intricate challenges associated with game audio summarization.

2) Speech to Text (STT)

The initial step in the process of generating text summaries from the audio commentary dataset involves the crucial task of converting speech into text using advanced Speech to Text (STT) technology. STT enables the machine or software program to accurately transcribe spoken words into written or displayed text, finding widespread applications in areas such as voice recognition software, automated transcription services, and virtual assistants. This technology is pivotal in various tasks, including dictation, audio transcription, and real-time captioning for live events.

In our research, we employed the highly sophisticated Azure Speech to Text (STT) API to transform the audio generated by the Deezer Spleeter module into noise-free text data. Azure Speech to Text, an integral component of the Azure Cognitive Services suite provided by Microsoft, offers a suite of advanced features and powerful APIs specifically designed for generating precise and reliable text transcripts from diverse audio sources. This robust service boasts an extensive language and accent support, while also allowing for customizable recognition of domain-specific vocabularies and terminologies, thereby ensuring utmost accuracy and adaptability in the generation of text summaries.

3) Selector and Rewriter Model

The selector and rewriter model originally introduced by Huang et al. [10] constitute essential components within our comprehensive summarization framework. The selector model functions as a fundamental classifier responsible for discerning the importance of individual sentences within the input text. In our specific implementation, we adopted a multinomial naive Bayes algorithm to power the selector model. Leveraging this algorithm, the selector model efficiently analyzes the intrinsic features and characteristics of each sentence, enabling it to accurately classify whether a sentence carries significance or not.

Working in conjunction with the selector model, the rewriter model incorporates a BART [18] architecture. By leveraging the advanced capabilities of the BART model, the rewriter model takes as input a sentence identified as important by the selector model from the transcribed commentary. Subsequently, it effectively transforms this selected sentence into a summary-like sentence that captures the core essence and vital information contained within the original text. Through this iterative process, the rewriter model adeptly distills the key elements from the selected sentence, skillfully generating a concise and informative summary sentence.

4) Dataset Curation For Selector and Rewriter

To construct datasets for training the selector and rewriter models, we utilized the SportsSum [10] and K-SportsSum [11] datasets. The creation of the selector model dataset requires annotated labels that indicate the importance of commentary sentences and their associated summary sentences. Although the summary sentences lack explicit timeline information, certain sentences begin with phrases such as "In the n^{th} minute" as highlighted by Huang et al. [10] analyzing these sentences, we extracted the timeline information, determining the specific value of "n" for each news sentence.

To map news sentences to corresponding commentary sentences, the following steps were undertaken:

- For each news sentence, denoted as r^i, we extracted the relevant timeline information, h^i, from potential commentary sentences.
- We established a set $C(i) = \{c_k, c_{k+1}, c_{k+2}, \ldots\}$ where c_j represents the commentary sentence with timeline information $t_j \in [h_i, h_{i+3}]$ for $k \leq j \leq k + l$.
- BERTScore was computed between the news sentence r_i and all commentary sentences in C(i). The commentary sentence $c_j \in C(i)$ with the highest score was considered the mapping for news sentence r_i.

This mapping process yielded a set of paired commentary and news sentences, forming the training data for our rewriter model. For training the selector model, sentences were labeled as '1' for the sentence with the highest BERTScore [19] and '0' for the remaining sentences, creating a labeled dataset for the selector model.

To train the rewriter model, the labeled pairs of commentary sentences were paired with their corresponding news sentences. It is important to note that the corresponding commentary sentence could only be generated from sentences containing the phrase "In the n^{th} minute."

5) Training Selector and Rewriter Models

The selector model employs a multinomial Naïve Bayes classifier to determine the importance of a given sentence. The features used in the multinomial Naïve Bayes algorithm were generated using the tf-idf (Term Frequency-Inverse Document Frequency) feature extraction technique. This approach enables the model to assess the significance of each sentence based on its unique characteristics and contextual relevance.

$$\mathrm{TF(t)} = \frac{Number\ of\ times\ term\ (t)\ appears\ in\ a\ document}{total\ no\ of\ terms\ in\ the\ document}$$

$$IDF_i = log\left(\frac{n\ (total\ no\ of\ documents)}{df_i\ (no.of\ documents\ containing\ term\ i)}\right)$$

Multinomial Naïve Bayes predicts the class with the highest posterior probability by multiplying the prior probability of the class with the conditional probabilities of the features given the class, assuming feature independence.

$$P(C_i|D) = p(C_i) * (\prod P(w_j|C_i) * f(w_j, D))$$

Here, $p(C_i)$ is the class prior probability, $P(w_j|C_i)$ is the likelihood probabilities of each word for each class, $f(w_j, D)$ is the frequency of word w_j in document D.

The training process involves training a Selector model followed by training a rewriter model. The dataset generation process outlined above produces pairs of commentary and news sentences. The Selector model classifies these commentary sentences as important or non-important. The important labeled commentary sentences are then fed into the rewriter model. The rewriter model is a fine-tuned version of BART [18], a pretrained language model developed by Facebook. BART utilizes a transformer-based architecture and employs a denoising auto-regressive decoder approach. It reconstructs original sentences from masked versions, enabling it to learn robust relationships between words and phrases. BART's large corpus of denoised text data enhances its ability to capture meaningful representations. Notably, BART is easier to fine-tune compared to other advanced language models like BERT, requiring fewer parameters and boasting a simpler architecture. BART has also demonstrated superior performance in text summarization tasks.

V. RESULT AND ANALYSIS

A. Performance Evaluation

In this study, we investigated the potential of audio summarization using a Selector and Rewriter model. The model achieved a Rouge score of 0.0963, indicating room for improvement in the quality of the generated audio summaries. However, it is noteworthy that our findings unveiled

promising results, highlighting the possibility of generating meaningful summaries from audio sources.

We also explored the generation of summaries from metadata using a naive metadata template. These summaries achieved a Rouge score of 0.1994, demonstrating the effectiveness of utilizing static information present in the metadata. It is important to acknowledge that the metadata only contains static information and lacks the dynamic game information, resulting in the generation of static summaries.

To evaluate the quality of the audio and metadata-based summaries, we encountered the challenge of lacking ground truth summaries for comparison. To address this limitation, we utilized the SoccerNet extension dataset, which served as a comprehensive set of ground truth summaries for our evaluation. By utilizing this dataset, we were able to assess the performance of the audio and metadata-generated summaries effectively. This approach allowed us to overcome the absence of direct ground truth and provide a robust evaluation framework for our summarization models.

Moreover, we investigated the generation of summaries from captions using a Longformer model. The summaries generated from captions achieved a Rouge score of 0.5546, showcasing the highest Rouge score among the different summarization methods employed in this study. The use of captions allowed us to capture additional information associated with the game, contributing to more comprehensive and informative summaries.

These results highlight the potential of utilizing various sources of information for football game summarization. By combining audio, metadata, and captions, we can improve the quality and comprehensiveness of the generated summaries. Furthermore, the approaches and techniques employed in this study are not limited to football but can be applied to various sports games, making it a versatile solution for summarization in different sporting domains.

TABLE I. CLASSIFICATION REPORT FOR THE SELECTOR MODEL

Labels	**Classification Report**		
	Precision	***Recall***	***F1-score***
Important	0.71	0.64	0.68
Non-Important	0.67	0.74	0.71
Accuracy	0.69		

TABLE II. ROUGE SCORES FOR VARIOUS DATA MODALITIES

Modes	**datasets**	**ROUGE Scores**		
		ROUGE-1	***ROUGE-2***	***ROUGE-L***
Audio	SoccerNet	0.963	0.168	0.703
Metadata	SoccerNet	0.199	0.213	0.132
captions	SportsSum, K-SportsSum, Goal	0.554	0.287	0.278

B. Limitations

Our research paper acknowledges several limitations that should be considered:

- Quality of Generated Summaries: The presence of significant noise in the audio data poses challenges to accurate transcription, leading to potential inaccuracies and inconsistencies in the generated summaries. The performance of the Selector model, although showing promise, is not yet optimal and requires further refinement.
- Limitations of Available Data Sources: The static nature of the metadata restricts the capture of dynamic and evolving game information, resulting in static and templated summaries. Relying solely on pre-generated captions may limit the flexibility and adaptability of the summarization process.
- Scarcity of Comprehensive Datasets: There is a lack of comprehensive datasets that incorporate multiple modalities, such as audio, video, and text, in a unified format. The absence of such datasets hinders the development and evaluation of more holistic summarization models.
- Challenges in Multimodal Integration: Integrating multiple modalities, such as audio, visual, and textual information, is an ongoing area of research. Further exploration is needed to effectively leverage multiple modalities and exploit their complementary nature for more robust and informative summaries.
- Data Alignment: The reliance on pre-existing datasets for training and evaluation may limit the generalizability of our findings to the sports domain. These datasets may not fully align with the specific requirements and nuances of sports summarization tasks. Future research should prioritize the creation of dedicated and comprehensive datasets specifically tailored to sports summarization, encompassing a wide range of sports events and diverse commentary styles.

VI. CONCLUSION

Our research focuses on utilizing deep learning techniques to generate game summaries of football matches through the analysis of audio, metadata, and captions. The summarization process involves leveraging three distinct data types to generate comprehensive and informative summaries. Our approach, particularly the utilization of metadata and captions, has achieved results on par with the state-of-the-art methods in the field, demonstrating its effectiveness and competitive performance. Additionally, we have explored the untapped potential of audio-based summarization, showcasing promising results that highlight the possibilities and opportunities for leveraging audio data in the summarization process. Furthermore, we have addressed limitations in the existing datasets by improving their quality, resolving encoding issues, and reducing noise. This research also contributes to the expansion of available datasets in the sports summarization domain, fostering further advancements and enabling future research in this area. Our findings contribute to the ongoing development of sports summarization techniques and lay a solid foundation for future exploration.

ACKNOWLEDGMENT

We extend our heartfelt appreciation to the Department of Electronics and Computer Engineering for granting us the invaluable opportunity to pursue research in this field and for their unwavering support throughout the duration of this project. Their guidance, motivation, and provision of

necessary resources have played a pivotal role in the successful execution of our research. We are deeply grateful for their continuous encouragement and assistance, which have been instrumental in our exploration and advancement of this topic.

REFERENCES

[1] J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, "Automatic Text Summarization Using a Machine Learning Approach," in *Advances in Artificial Intelligence*, G. Bittencourt and G. L. Ramalho, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2002, pp. 205–215. doi: 10.1007/3-540-36127-8_20.

[2] N. S. Shirwandkar and S. Kulkarni, "Extractive Text Summarization Using Deep Learning," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Aug. 2018, pp. 1–5. doi: 10.1109/ICCUBEA.2018.8697465.

[3] S. Song, H. Huang, and T. Ruan, "Abstractive text summarization using LSTM-CNN based deep learning," *Multimed Tools Appl*, vol. 78, no. 1, pp. 857–875, Jan. 2019, doi: 10.1007/s11042-018-5749-3.

[4] J. Zhang, J. Yao, and X. Wan, "Towards Constructing Sports News from Live Text Commentary," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1361–1371. doi: 10.18653/v1/P16-1129.

[5] A. O. M. Salih, "Audio Noise Reduction Using Low Pass Filters," *Open Access Library Journal*, vol. 4, no. 11, Art. no. 11, Nov. 2017, doi: 10.4236/oalib.1103709.

[6] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006, doi: 10.1109/TSA.2005.860851.

[7] G. Yu, S. Mallat, and E. Bacry, "Audio Denoising by Time-Frequency Block Thresholding," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1830–1839, May 2008, doi: 10.1109/TSP.2007.912893.

[8] D. Liu, P. Smaragdis, and M. Kim, "Experiments on Deep Learning for Speech Denoising".

[9] S. J. Mohammed and N. Radhika, "Audio Denoising Using Deep Neural Networks," in *Intelligent Data Communication Technologies and Internet of Things*, D. J. Hemanth, D. Pelusi, and C. Vuppalapati, Eds., in Lecture Notes on Data Engineering and Communications Technologies. Singapore: Springer Nature, 2022, pp. 33–47. doi: 10.1007/978-981-16-7610-9_3.

[10] K.-H. Huang, C. Li, and K.-W. Chang, "Generating Sports News from Live Commentary: A Chinese Dataset for Sports Game Summarization," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 609–615. Accessed: Jun. 24, 2023. [Online]. Available: https://aclanthology.org/2020.aacl-main.61

[11] J. Wang *et al.*, "Knowledge Enhanced Sports Game Summarization." arXiv, Nov. 24, 2021. Accessed: Jun. 24, 2023. [Online]. Available: http://arxiv.org/abs/2111.12535

[12] J. Wang *et al.*, "SportsSum2.0: Generating High-Quality Sports News from Live Text Commentary." arXiv, Oct. 12, 2021. doi: 10.48550/arXiv.2110.05750.

[13] C. van der Lee, E. Krahmer, and S. Wubben, "PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences," in *Proceedings of the 10th International Conference on Natural Language Generation*, Santiago de Compostela, Spain: Association for Computational Linguistics, Sep. 2017, pp. 95–104. doi: 10.18653/v1/W17-3513.

[14] S. Gautam, C. Midoglu, S. Shafiee Sabet, D. B. Kshatri, and P. Halvorsen, "Soccer Game Summarization using Audio Commentary, Metadata, and Captions," in *Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos*, Lisboa Portugal: ACM, Oct. 2022, pp. 13–22. doi: 10.1145/3552463.3557019.

[15] J. Wang, T. Zhang, and H. Shi, "GOAL: Towards Benchmarking Few-Shot Sports Game Summarization." arXiv, Jul. 18, 2022. doi: 10.48550/arXiv.2207.08635.

[16] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer." arXiv, Dec. 02, 2020. Accessed: Jun. 24, 2023. [Online]. Available: http://arxiv.org/abs/2004.05150

[17] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "SPLEETER: A FAST AND STATE-OF-THE ART MUSIC SOURCE SEPARATION TOOL WITH PRE-TRAINED MODELS," 2019.

[18] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." arXiv, Oct. 29, 2019. Accessed: Jun. 24, 2023. [Online]. Available: http://arxiv.org/abs/1910.13461

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT." arXiv, Feb. 24, 2020. Accessed: Jun. 24, 2023. [Online]. Available: http://arxiv.org/abs/1904.09675