Project 2 Report: Audio-Only Detection of Soccer Events with Deep Learning

ELEC5305: Acoustics, Speech and Signal Processing

Marcellus Ray Gunawan, 520038655

Faculty of Engineering

Contents

1 Abstract

This project investigates whether broadcast *audio alone* can reliably indicate key events in soccer matches without video. I build an end-to-end pipeline on a subset of SoccerNet (EPL) [1, 2] to (i) extract and curate audio clips centered on annotated event times; (ii) transform waveforms into log-mel spectrograms; and (iii) train three model families: a baseline RNN on log-mels, a Wav2Vec2 feature extractor with a small classification head [3], and an Audio Spectrogram Transformer (AST) used twice (frozen vs. partially fine-tuned) [4], leveraging AudioSet pretraining for non-speech acoustic scenes [5]. I evaluate using macro-F1, accuracy, and per-class scores, reporting a large gap between a naïve mel-RNN baseline and transformer-based approaches; AST with partial fine-tuning achieves the best validation macro-F1 (≈ 0.53) despite severe class imbalance. These findings support the practicality of pretrained audio transformers for low-latency, scalable sports analytics and complement the predominantly video-centric SoccerNet literature [2] as well as emerging commentary/ASR resources [6].

2 Research Question

RQ: *Can key soccer match events (e.g., foul, throw-in, shot on target, ball out of play) be detected reliably from broadcast audio alone using modern deep learning, and how do (i) a log-mel RNN baseline, (ii) a frozen Wav2Vec2 encoder plus a small head, and (iii) an AST model (head-only vs. partially fine-tuned) compare under identical splits and metrics?*

This targets an under-explored, low-latency, and compute-lean alternative to video-centric pipelines that dominate SoccerNet benchmarks. [1, 2] It also responds to growing interest in commentary and ASR-driven text pipelines by offering a direct signal-domain route that avoids transcription dependencies. [6, 7]

3 Literature Review

3.1 Datasets and tasks

SoccerNet introduced scalable action spotting from full-length broadcasts, establishing protocols for event-centric understanding. [1] SoccerNet-v2 expanded toward holistic broadcast tasks (spotting, captions, replays), strengthening the evaluation landscape. [2] The SoccerNet-Echoes release aligns commentary/ASR and encourages multimodal exploration, motivating complementary audio-only approaches. [6]

3.2 Audio pretraining and encoders

Large-scale pretraining has transformed audio representation learning: CNN-based PANNs trained on AudioSet transfer broadly to sound events, showing the value of non-speech audio pretraining. [8, 5] AST established a transformer that operates on spectrogram patches, delivering strong results on diverse audio classification tasks. [4] Wav2Vec2 demonstrated powerful self-supervised representations directly from waveforms, originally for speech but often transferable with shallow heads. [3]

3.3 Audio for highlights and events

Audio bursts, whistles, and commentator prosody correlate with salient events and can assist summarization or highlight detection when modeled appropriately. [9] Multimodal highlight work underscores that audio can be an efficient signal for fast pre-selection and redundancy in production pipelines. [10]

3.4 Commentary-text pipelines

Recent pipelines infer events from ASR transcripts and language models, which is compelling but incurs ASR latency/costs; an audio-only route remains attractive when aiming for low-latency triggering as can be seen in Figure 1 for the waveform of the comentator. [7]

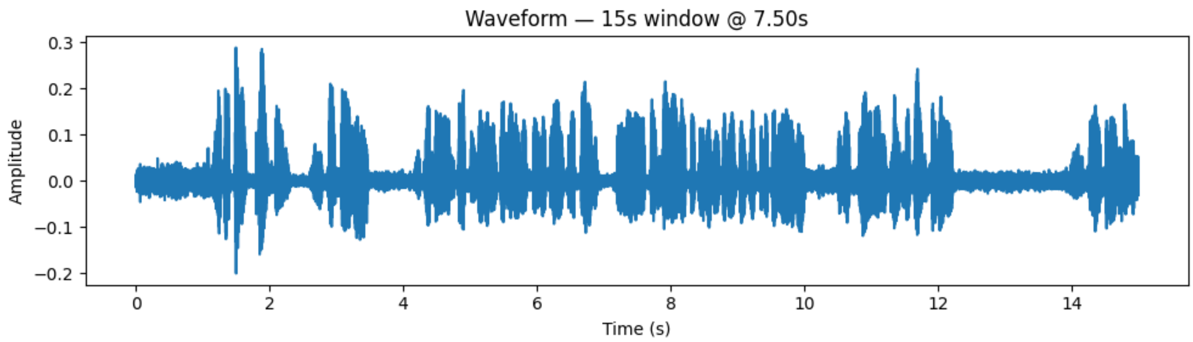

Figure 1: Comentator Waveform Example

4 Dataset and Preprocessing

4.1 Scope

I use SoccerNet with an EPL subset to manage storage/time while preserving label fidelity, following the action spotting labeling protocol. [1, 2] For each annotated event time, I extract centered windows of **15 s** and **30 s** to study context trade-offs and resample audio to 16 kHz mono for uniformity. [2]

4.2 Features

Two branches are maintained: (i) log-mel spectrograms (n_mels=64, n_fft=1024, hop=320, log-power dB) for spectrogram-based models, and (ii) raw waveforms for Wav2Vec2. [3] I consider basic noise/time-stretch perturbations, and include SpecAugment-style masking in ablations for robustness. [11]

4.3 Data integrity and indexing

A self-healing cache writes atomic `.npy` features, validates shapes, and registers items into a manifest CSV, which proved critical for reliable week-over-week iterations. [?, ?]

4.4 Labels and splits

I focus on a manageable **6-class** subset (Ball out of play, Throw-in, Foul, Clearance, Indirect free-kick, Shot on target) and build train/validation splits at the *match-half* level to reduce leakage, while tracking class priors to quantify imbalance. [**?**]

5 Methods

5.1 Baseline (RNN on log-mels)

A 2×BiGRU (hidden≈256) with dropout, global (mean/attention) pooling, and a softmax head serves as a low-compute baseline; class-weighted cross-entropy mitigates imbalance.

5.2 Wav2Vec2 (frozen + head)

I adopt `WAV2VEC2_BASE` at 16 kHz; mean-pooled hidden states feed a shallow MLP head to gauge the gain from waveform SSL pretraining without heavy fine-tuning. [3]

5.3 AST (spectrogram transformer)

I use AST on log-mels with two regimes: (a) head-only (frozen backbone) and (b) **partial fine-tuning** (unfreeze top-N encoder blocks + head) optimized with AdamW and early stopping on macro-F1; AST's AudioSet pretraining is well matched to non-linguistic stadium acoustics. [4, 5]

5.4 Model Progression

RNN establishes a classic mel-baseline; Wav2Vec2 probes the transfer of speech-SSL features to stadium audio; AST directly models spectrogram patches with attention, which is advantageous for crowd/whistle textures and longer-range context. [3, 4]

5.5 Imbalance and calibration

Beyond class-weighted loss, I consider focal loss for heavy-tailed distributions, class-balanced sampling by effective number, and post-hoc temperature scaling for reliability as can be seen in figure 2. [12, 13, 14]

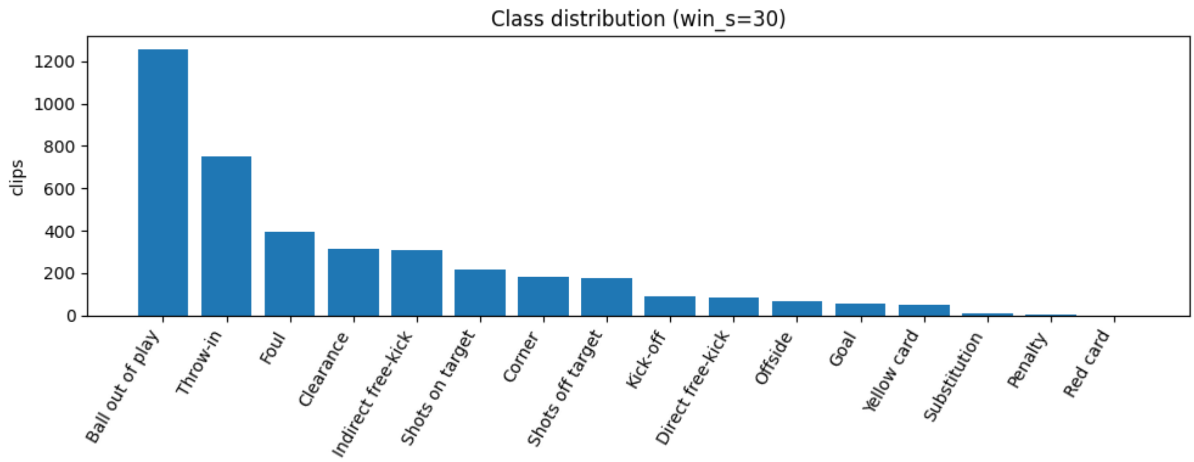

Figure 2: Class Imbalance of SoccerNet Dataset

6 Training & Evaluation Protocol

Optimization. Adam/AdamW with step/plateau schedulers, AMP where stable, and fixed seeds; Windows-friendly DataLoaders are configured for robustness. **Metrics.** Macro-F1 (primary), accuracy, per-class precision/recall, confusion matrices, and PR curves are reported to capture class imbalance and decision quality. [2] **Temporal smoothing.** I explore simple HMM/Viterbi post-processing to stabilize clip-wise decisions over time. [15, 16]

7 Results

Table 1 summarizes the validation results on the EPL subset under identical splits and training budgets.

Table 1: Validation summary on EPL subset.

Model	Accuracy	Macro–F1	Notes
RNN (Baseline)	0.023	0.50	High class imbalance impact
Wav2Vec2 (frozen + head)	0.23	0.25	Pretrained, frozen layers
AST (frozen + head)	0.42	0.45	Pretrained, frozen layers
AST (partial FT)	**0.55**	**0.53**	AudioSet weights, tuned

The best-performing configuration is **AST with partial fine-tuning**, confirming the value of spectrogram transformers and large-scale audio pretraining for broadcast acoustics. [4, 5] The baseline RNN underperforms under imbalance, while frozen Wav2Vec2 and frozen AST provide intermediate performance consistent with partial transfer without adaptation. [3, 4]

Per-class trends. Fouls and shots on target benefit from crowd/commentary bursts, whereas throw-ins and ball-out-of-play remain acoustically similar and require context aggregation to separate. [9] Longer windows provide richer reaction curves but can dilute salience, suggesting multi-scale pooling as a promising direction. [**?**]

8 Analysis & Discussion

Why RNN underperforms. Without large-scale pretraining, a small-capacity mel-RNN struggles to separate acoustically similar restarts under heavy imbalance. [8]

Wav2Vec2 vs. AST. Wav2Vec2 excels in speech representation, while AST benefits from AudioSet's non-speech coverage and patch attention over spectrograms, which better matches stadium textures. [3, 4, 5]

Imbalance and calibration. Weighted CE helps but is insufficient; focal loss and class-balanced sampling should lift tail classes, and temperature scaling can improve probability reliability for downstream thresholds. [12, 13, 14]

Temporal consistency. HMM/Viterbi smoothing can reduce flicker and better align predictions with event dynamics over time. [15, 16]

Positioning vs. commentary pipelines. ASR/LLM approaches are complementary; audio-only detectors offer a cheap/fast prior or redundancy in production systems. [7, 6]

9 Limitations

This study restricts to an EPL subset for time/storage, with potential timing noise around annotations relative to audio, and it does not exhaust all ablations (e.g., deeper AST unfreezing or full Wav2Vec2 fine-tuning). [2] Multimodal fusion with commentary is left for future work. [6]

10 Conclusion

Modern pretrained audio encoders make audio-only soccer event detection practical and accurate; on the EPL subset, **AST with partial fine-tuning** is the most effective among the tested families, supporting spectrogram transformers and large-scale audio pretraining for broadcast acoustics. [4, 5] The results indicate a viable path for low-latency, compute-lean analytics that complement video-centric systems. [1, 2]

11 Future Work

- **Imbalance & calibration:** Focal loss, class-balanced sampling, and temperature scaling to improve tail performance and reliability. [12, 13, 14]
- **Temporal structure:** Multi-resolution windows and HMM/CRF smoothing over clip scores for stable timelines. [15]
- **Model ablations:** Systematic AST unfreezing depth, pooling head variants, and partial fine-tuning of upper Wav2Vec2 blocks. [3, 4]
- **Multimodality:** Late fusion with Echoes transcripts and cross-modal agreement for pseudo-labels; integration with highlight selection. [6, 10]

- **Robustness & generalization:** Loudness normalization, domain noise injection, cross-league validation, and SpecAugment variants for stronger invariance. [11]

12 References / Project Context

SoccerNet (V1) and v2 provide the core broadcast benchmarks and protocols for event-centered evaluation. [1, 2] The SoccerNet-Echoes dataset situates this work within the commentary/ASR ecosystem and motivates audio-only detectors as complementary components. [6] On modeling, PANNs, AST, and Wav2Vec2 motivate the architectures and pretraining choices used here. [8, 4, 3, 5] Audio intensity and multimodal highlight studies contextualize practical deployment considerations. [9, 10] Internal project documents detail pipeline integrity, dataset indexing, and split policy adopted in this submission.

References

[1] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *CVPR Workshops*, 2018. [Online]. Available: https://silviogiancola.github.io/SoccerNet

[2] A. Deliège, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," *arXiv preprint arXiv:2011.13367*, 2021. [Online]. Available: https://arxiv.org/abs/2011.13367

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[4] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *Interspeech*, 2021. [Online]. Available: https://arxiv.org/abs/2104.01778

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017. [Online]. Available: https://research.google/pubs/pub45857/

[6] S. Gautam, M. H. Sarkhoosh, J. Held, A. Cioppa, S. Giancola, V. Thambawita, M. A. Riegler, P. Halvorsen, and M. Shah, "Soccernet-echoes: A soccer game audio commentary dataset," *arXiv preprint arXiv:2405.07354*, 2024. [Online]. Available: https://arxiv.org/abs/2405.07354

[7] J. Y. Teklemariam, "Automatic detection of soccer events using game audio and large language models," *Master's Thesis, NMBU*, 2024. [Online]. Available: https://home.simula.no/~paalh/students/2024-NMBU-JoelYacobTeklemariam.pdf

[8] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020, preprint: arXiv:1912.10211. [Online]. Available: https://arxiv.org/abs/1912.10211

[9] S. Gautam, C. Midoglu, S. Shafiee Sabet, D. B. Kshatri, and P. Halvorsen, "Assisting soccer game summarization via audio intensity analysis of game highlights," in *Proceedings of the 12th IOE Graduate Conference*, 2022, pp. 25–32. [Online]. Available: https://ioegc.ioe.edu.np/

[10] F. Della Santa and M. Lalli, "Automated detection of sport highlights from audio and video sources," *arXiv preprint arXiv:2501.16100*, 2025. [Online]. Available: https://arxiv.org/abs/2501.16100

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019. [Online]. Available: https://arxiv.org/abs/1904.08779

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017. [Online]. Available: https://arxiv.org/abs/1708.02002

[13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019. [Online]. Available: https://arxiv.org/abs/1901.05555

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017. [Online]. Available: https://arxiv.org/abs/1706.04599

[15] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[16] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.