IDIAP RESEARCH REPORT

IMPROVING DOMAIN-SPECIFIC ASR FOR SOCCER BROADCAST COMMENTARY

Philip Awada [a] Petr Motlicek

Idiap-RR-03-2026

AUGUST 2026

[a] Ecole polytechnique fédérale de Lausanne EPFL

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Improving Domain-Specific ASR for Soccer Broadcast Commentary

Fine-Tuning, Prompting, and Shallow Fusion for Entity Recognition

Philip Awada

Supervisor: Dr. Petr Motlicek

School of Engineering, EPFL
Idiap Research Institute

Spring 2026

Abstract

Pre-trained automatic speech recognition (ASR) models such as OpenAI's Whisper achieve excellent performance on clean, read speech but degrade significantly on domain-specific audio. This project investigates this degradation on soccer broadcast commentary, where crowd noise, fast-paced delivery, and domain-specific vocabulary, particularly player and team names, pose challenges. Using the GOAL benchmark and SoccerNet broadcast audio, we fine-tune Whisper Medium on soccer commentary data and explore three inference-time techniques for improving entity recognition without retraining: decoder prompting with match rosters, shallow fusion via sequence biasing, and their combination. The combined approach achieves 88.3% entity detection accuracy, within 2.0 percentage points of the oracle upper bound, and raises unseen entity detection from 27.9% to 67.0%. We find that fine-tuning is a prerequisite for these inference-time techniques: without domain adaptation, prompting and shallow fusion cause severe hallucination. This contrasts with prior results on air traffic control speech, suggesting that domain difficulty determines whether prompting can work without fine-tuning. A companion reproducibility guide details how to replicate all experiments on the EPFL Izar cluster.

1 Introduction

Automatic speech recognition (ASR) has made remarkable progress in recent years, with large-scale models like Whisper [1] achieving near-human performance on standard benchmarks. However, these models are trained predominantly on clean, well-articulated speech and struggle when applied to specialized domains. Soccer broadcast commentary represents a particularly challenging domain: commentators speak rapidly and with varying emotion, crowd noise creates a difficult acoustic environment, and the vocabulary is dense with player names, team names, coaches and stadiums that may not appear frequently in general training data.

This project investigates the gap between general-purpose ASR and domain-specific performance on soccer commentary, and explores techniques to close it. The central research question is: *can inference-time techniques such as prompting and shallow fusion improve entity recognition in domain-specific ASR without retraining the model?* This question is motivated by a practical industry scenario: a broadcasting company deploys a single fine-tuned model and provides a match

roster (player names, team names) a few hours before kickoff, expecting improved transcription accuracy without any additional training.

The project follows a progressive methodology:

1. Establish baselines by evaluating Whisper Medium zero-shot on standard benchmarks (LibriSpeech) and soccer commentary (GOAL benchmark).
2. Fine-tune the model on soccer commentary and analyze entity-specific error patterns.
3. Apply inference-time techniques (decoder prompting and shallow fusion) individually and in combination.
4. Analyze performance on entities seen during training versus unseen entities.
5. Compare results with the original (non-fine-tuned) model to understand the role of domain adaptation.

This work extends the approach of Piaget et al. [2], who explored Whisper prompting for air traffic control (ATC) speech recognition under the same supervisor. While Piaget et al. found that prompting had a bigger impact on non-fine-tuned models, we find the opposite for soccer commentary: fine-tuning is a prerequisite for inference-time techniques to work, as the original model hallucinates entity names when prompted without domain adaptation.

2 Related Work

Whisper. Radford et al. [1] introduced Whisper, an encoder-decoder transformer trained on 680,000 hours of multilingual web audio using weak supervision. The model takes 30-second mel spectrograms as input and generates text tokens autoregressively. It comes in multiple sizes from tiny (39M parameters) to large (1.5B+); we use the Medium variant (769M parameters). Whisper's decoder accepts an optional text prompt of up to 224 tokens that biases the output through transformer cross-attention. Thorbecke et al. [14] demonstrated that Whisper can also serve as a teacher model for knowledge distillation into streaming transducer architectures, highlighting its versatility as a foundation for domain-specific ASR development.

Domain-specific ASR degradation. Pre-trained ASR models consistently degrade on domain-specific audio. Piaget et al. [2] demonstrated this for ATC speech, where Whisper's word error rate (WER) rose to 44.6% compared to significantly lower rates on standard benchmarks. Fine-tuning XLSR and Whisper on domain data reduced WER to 20.3% and 17.8% respectively. Prompting with 15 transcript examples further improved WER to 15.9%, though callsign detection accuracy remained low at 42.4%. Importantly, Piaget et al. found that prompting had a larger impact on non-fine-tuned models than on fine-tuned ones, a finding we revisit in this work. Bhattacharjee et al. [11] further showed that minimum-effort adaptation techniques can reduce WER in ATC systems without requiring full retraining, a finding that aligns with our goal of improving entity recognition at inference time. The ATCO2 corpus [12] provides a large-scale benchmark for ATC speech, demonstrating the broader interest in domain-specific ASR evaluation and the challenges posed by specialized vocabularies such as callsigns and waypoint names.

Contextual biasing and shallow fusion. Shallow fusion is a technique that modifies decoder output logits during beam search to boost the probability of specific tokens. Unlike prompting, which provides soft context through the attention mechanism, shallow fusion directly adds a numerical bonus to target token scores at each decoding step. This approach has been used in production ASR systems to handle dynamic vocabularies, for example boosting contact names in voice assistants. Bhattacharjee et al. [8] conducted a systematic comparison of contextual biasing methods for improving rare word detection in ASR, including shallow fusion, deep biasing, and trie-based approaches, finding that biasing consistently improves recognition of domain-specific

terms. Thorbecke et al. [9] proposed a unified trie-based approach that combines global and near-context biasing in a single decoding pass, enabling more efficient entity boosting during inference. Motlicek et al. [10] applied similar techniques to callsign recognition in ATC, leveraging entity lists provided before each session to improve recognition of flight identifiers without retraining. HuggingFace's `SequenceBiasLogitsProcessor` [6] provides a straightforward implementation that works directly with Whisper's generation pipeline, eliminating the need for specialized frameworks such as Icefall [7].

Prompting for ASR. Beyond Whisper's built-in decoder prompting, recent work has explored prompting in the context of large language model (LLM)-based ASR systems. Burdisso et al. [13] investigated prompt sensitivity in LLM-based speech recognition and proposed learnable projection layers to reduce variability across prompt formulations, demonstrating that the interaction between prompts and ASR output is an active area of research.

GOAL benchmark. Qi et al. [3] introduced the GOAL benchmark, containing human-verified transcriptions of 20 SoccerNet broadcast games (of which we use 16 for training and 4 for testing; see Section 3) with entity annotations covering six categories: player, team, coach, referee, stadium, and commentator. The annotations use a tagged text format (e.g., `[player]Sterling[player]`), enabling evaluation of entity-specific recognition accuracy. The benchmark covers games from five European leagues, all with English-language commentary.

3 Data

3.1 Datasets

LibriSpeech. We use the test-clean and test-other splits of LibriSpeech [4] as standard baselines for general ASR performance. This corpus consists of read English audiobook speech and is publicly available via HuggingFace.

SoccerNet. SoccerNet [5] provides broadcast video recordings of professional soccer matches. We downloaded the 20 games covered by the GOAL benchmark (40 half-time video files) at 224p resolution and extracted mono audio at 16 kHz using ffmpeg. Access to SoccerNet video data requires signing a Non-Disclosure Agreement.

GOAL benchmark. GOAL [3] provides human-verified transcriptions of these 20 games (40 half-time JSON files) with segment-level timestamps and entity annotations. Each segment contains an offset, duration, plain-text transcription, and a tagged version with entity type markers using paired tags: `[type]Name[type]`. The benchmark covers 7 English Premier League, 8 Champions League, 2 La Liga, 1 Ligue 1, and 1 Serie A games, all with English-language commentary. Some entity tags in GOAL are malformed (e.g., `Bony[palyer] and`); we filter these out during entity extraction using length, character, and format checks.

3.2 Train/Test Split

We split the 20 GOAL games into 16 for training and 4 for testing. The 4 test games were selected to provide diversity across leagues, teams, and commentary styles:

Table 1: Test set composition. Four games were selected from three different leagues to ensure diversity in commentary style, team names, and acoustic conditions.

Game	**League**	**Selection rationale**
Liverpool 3–0 Manchester City	EPL	Original baseline game
Arsenal 3–0 Chelsea	EPL	Different teams and speakers
Barcelona 3–0 Bayern Munich	Champions League	Different competition, style
Marseille 2–3 PSG	Ligue 1	French names in English commentary

The training set contains 7,190 segments totaling 15.66 hours of audio. The test set contains 554 segments (1.38 hours), using the first 15 minutes of each test game's two halves. Segments longer than 30 seconds are excluded due to Whisper's fixed-length processing window. Training and test data are completely separated at the game level: no audio from a test game appears in training.

3.3 Entity Extraction and Seen/Unseen Split

Entity names are extracted from GOAL's tagged annotations using regex-based parsing. After filtering malformed tags, we obtain 788 unique entity names in the training set and 231 in the test set. Of the 231 test entities, 155 (accounting for 1,088 mentions) also appear in the training set ("seen" entities) and 76 (accounting for 276 mentions) appear exclusively in the test set ("unseen" entities). Examples of high-frequency unseen entities include Muller (15 mentions), Ozil (13), Payet (16), Gignac (16), Morel (15), and Rafinha (14), many of which are non-English names that may be phonetically challenging for the model.

4 Methods

This section describes the four stages of our experimental pipeline: zero-shot baseline evaluation, domain-specific fine-tuning, and two inference-time techniques (decoder prompting and shallow fusion) that can be applied individually or in combination. The key design principle is that prompting and shallow fusion operate on different parts of the decoder and can therefore be combined without interference.

4.1 Baseline Evaluation

We evaluate Whisper Medium (`openai/whisper-medium`, 769M parameters) in a zero-shot setting on LibriSpeech test-clean, test-other, and the GOAL test set. All evaluations apply HuggingFace's `EnglishTextNormalizer` to both predictions and references before computing WER (see Section 5 for metric definitions). This normalization lowercases text, strips punctuation, expands numbers (e.g., "3" → "three"), and normalizes contractions. Without it, WER is artificially inflated by surface-level formatting differences. WER is computed using the `jiwer` library. The baseline establishes the domain gap between clean speech and soccer commentary, motivating the subsequent fine-tuning and inference-time experiments.

4.2 Fine-Tuning

We fine-tune Whisper Medium on the 16-game training set using HuggingFace's `Seq2SeqTrainer`. Fine-tuning adapts the model's decoder weights to the acoustic characteristics of soccer commentary (crowd noise, emotional delivery) and the domain vocabulary (player names, team names). The hyperparameters are: learning rate 1×10^{-5}, per-device batch size 4 with gradient accumulation over 4 steps (effective batch size 16), 2,000 fine-tuning steps, fp16 mixed precision,

and gradient checkpointing enabled to fit within the V100's 32 GB memory. Fine-tuning takes approximately 1 hour 51 minutes on a single NVIDIA V100 GPU. The fine-tuning loss decreases steadily from 14.68 to 0.27 across the 2,000 steps, indicating successful convergence.

4.3 Decoder Prompting

Whisper's decoder accepts a text prompt of up to 224 tokens via the `processor.get_prompt_ids()` method. The prompt tokens are prepended to the decoder input, and the model attends to both the audio encoding and the prompt through transformer cross-attention. This creates a soft bias: the model's output probabilities shift toward words that appear in the prompt, but the prompt cannot force specific words into the output. The mechanism is analogous to how a human transcriber who has seen the team sheet will be more likely to correctly hear ambiguous names. We evaluate five prompting strategies to isolate different aspects of this mechanism:

1. **N (No prompt):** Baseline with no prompt, establishing the fine-tuned model's default performance.
2. **C (Context):** A short domain description: "football match, crowd noise." Tests whether generic domain context helps.
3. **RT (Random Transcripts):** 15 randomly selected training transcripts concatenated into the 224-token prompt window. Tests whether in-domain text style (without match-specific names) provides a benefit.
4. **PL (Player List):** A per-game roster containing all player names, team names, and coach names extracted from GOAL annotations (approximately 35–40 entities per game). This is the realistic deployment scenario: a roster is available before kickoff.
5. **GT (Oracle):** The actual reference transcript for each segment, serving as an upper bound on what prompting can achieve. This configuration is not realistic (it requires knowing the answer in advance) but establishes the ceiling.

4.4 Shallow Fusion via Sequence Biasing

While prompting biases the decoder through attention (i.e., the model learns to weigh prompt tokens when predicting output), shallow fusion operates at a different level: it directly modifies the logit scores before the softmax at each decoding step. We use HuggingFace's `SequenceBiasLogitsProcessor` [6] to implement this. The procedure is:

1. **Tokenization:** Each entity name is tokenized using the Whisper tokenizer. Single-word names like "Messi" map to a single token, while multi-word or unusual names like "Schweinsteiger" may require four or more tokens.
2. **Bias assignment:** A bias score β is added to the logit of each entity token at every decoding step. Higher β values make the model more likely to output entity tokens, but excessively high values cause the model to insert entity names regardless of the audio content.
3. **Beam search:** We use beam search with $k = 5$ beams, as biasing has minimal effect with greedy decoding. Beam search maintains multiple candidate hypotheses, allowing the bias to influence the final selection among plausible alternatives.

Because prompting and shallow fusion operate on different components of the decoder (attention weights vs. output logits), they are potentially complementary: prompting provides global context about the domain, while shallow fusion provides a targeted numerical boost to specific entity tokens.

We search over bias strengths $\beta \in \{0.5, 1.0, 1.5, 2.0, 3.0, 5.0\}$, finding that $\beta = 3.0$ is optimal. Values above 5.0 cause the model to degenerate, inserting entity names throughout the output regardless of the audio content.

4.5 Combined Approach

We combine player list prompting (PL) with shallow fusion at $\beta = 3.0$ and beam size $k = 5$. Both techniques receive the *exact same* entity list per game, extracted from GOAL annotations. This is critical for fair comparison: an early version of our experiments used different entity lists for prompting ($\sim$22 entities) and shallow fusion ($\sim$35–40 entities), leading to misleading comparisons. After correcting this discrepancy, all three approaches use identical entity inputs. The combined approach tests whether the two mechanisms provide additive gains or whether their benefits overlap.

5 Evaluation Metrics

Word Error Rate (WER). The standard ASR metric, defined as:

$$\text{WER} = \frac{S + D + I}{N}$$

where S, D, I are the number of substitutions, deletions, and insertions respectively, and N is the number of words in the reference. WER can exceed 100% when the model inserts more words than exist in the reference. All WER values are computed after applying `EnglishTextNormalizer` to both predictions and references.

Entity WER. To isolate entity recognition difficulty, we compute WER on the subset of segments that contain at least one entity mention in the reference transcript. This is computed by first identifying all test segments whose reference text contains at least one tagged entity, then computing standard WER only over that subset. This measures overall transcription quality in entity-rich segments, rather than entity recognition specifically.

Entity Detection Accuracy. A binary per-entity metric that measures whether the model successfully outputs entity names regardless of their exact position in the transcript. For each of the 1,374 entity mentions in the test set, we check whether the entity name appears anywhere in the ASR output for that segment after normalization. We report two variants:

- *Full match:* All words of a multi-word entity must be found in the output. For example, "Thiago Silva" requires both "thiago" and "silva" to appear.
- *Partial match:* At least one word of the entity must appear.

The metric is averaged across all entity mentions. Unlike entity WER, which measures transcription quality of entity-containing segments, entity detection accuracy asks a simpler binary question: *did the entity name appear in the output or not?*

In the results tables, we use the following abbreviations: **Ent. WER** for entity WER, **Ent. Det. Acc.** for entity detection accuracy (full match), and **Partial Acc.** for entity detection accuracy (partial match).

6 Results

6.1 Baseline Performance

Table 2 shows Whisper Medium's zero-shot performance across three evaluation sets of increasing difficulty.

Table 2: Whisper Medium zero-shot WER across datasets, showing progressive degradation from clean read speech to domain-specific soccer commentary. No fine-tuning or prompting is applied.

Dataset	**Conditions**	**WER**
LibriSpeech test-clean	Clean read audiobooks	2.43%
LibriSpeech test-other	Noisier read audiobooks	6.40%
GOAL (4-game test set)	Soccer broadcast commentary	17.09%

Whisper degrades approximately 7× on soccer commentary compared to clean speech (17.09% vs. 2.43%), confirming a substantial domain gap that motivates domain-specific fine-tuning.

6.2 Fine-Tuning and Error Analysis

Fine-tuning Whisper Medium on 16 games of soccer commentary reduces WER from 17.09% to 13.53% (this is the value that we're going to be interested in for the upcoming parts), a 3.56% absolute improvement (approximately 21% relative). Table 3 breaks down errors by whether they affect entity words (player names, team names, etc.) or non-entity words.

Table 3: Error breakdown on the fine-tuned model, separating entity words (player, team, coach, referee, stadium, commentator names) from non-entity words. "Errors (S+D+I)" counts the total substitutions, deletions, and insertions attributed to each word category after alignment. WER is computed per category as errors divided by word count.

Category	**Words**	**Errors (S+D+I)**	**WER**
Entity words (player, team, coach, etc.)	1,734 (11.6%)	392 (19.5%)	22.61%
Non-entity words	13,155 (88.4%)	1,221 (80.5%)	9.28%
Overall	14,889	1,613	13.53%

The entity WER of 22.61% is computed by aligning the ASR output against the reference for each segment, identifying which alignment errors fall on entity words, and dividing the number of entity-word errors by the total number of entity words. Entity names are thus 2.4× harder to transcribe than non-entity words (22.61% vs. 9.28%). Despite constituting only 11.6% of all words, entities account for 19.5% of all errors, confirming entity recognition as the primary challenge and motivating the inference-time techniques explored below. The most common entity errors are player name substitutions, where phonetically similar names are confused. The Marseille vs. PSG game (Ligue 1) appears hardest, likely because French player names are spoken within English commentary.

6.3 Prompting

Table 4 presents results for the five prompting strategies on the fine-tuned model across the full 554-segment test set (1,374 entity mentions).

Table 4: Prompting results on the fine-tuned model, comparing five strategies. WER = word error rate on all segments. Ent. WER = WER computed only on the subset of segments containing at least one entity. Ent. Det. Acc. = entity detection accuracy, i.e., the fraction of entity mentions (out of 1,374) whose name appears in the ASR output (full match). Partial Acc. = same metric but requiring only one word of multi-word entities to match.

	Description	**WER**	**Ent. WER**	**Ent. Det. Acc.**	**Partial Acc.**
N	No prompt (baseline)	13.53%	13.76%	75.8%	80.3%
C	Context description	14.87%	15.11%	76.5%	80.9%
RT	15 random transcripts	13.04%	13.19%	76.9%	81.3%
PL	**Player list (roster)**	**12.17%**	**12.23%**	**84.8%**	**89.6%**
GT	Oracle (ground truth)	12.17%	12.59%	90.3%	92.1%

The player list prompt (PL) is the best realistic strategy, improving entity detection accuracy from 75.8% to 84.8% (+9.0 percentage points). Its WER of 12.17% matches the oracle, indicating that with a complete roster, prompting is as effective for overall transcription quality. The context prompt (C) slightly hurts WER, consistent with Piaget et al.'s finding for ATC [2] that generic domain descriptions provide little benefit. Random transcripts (RT) show a marginal WER change that is likely not statistically significant, and barely affect entity detection, since they contain player names from other games rather than the current match roster. The oracle prompt (GT) sets the entity detection ceiling at 90.3%, representing the maximum achievable through the attention-based prompting mechanism.

6.4 Shallow Fusion

Table 5 shows results on the fine-tuned model for different bias strengths. An important baseline consideration: the no-prompt (N) result uses greedy decoding, while all shallow fusion experiments use beam search with $k = 5$. Beam search alone (without bias) yields WER = 14.31% and entity detection accuracy = 77.7%, which is the proper comparison point for shallow fusion.

Table 5: Shallow fusion results at various bias strengths β, all using beam search with $k = 5$ beams on the fine-tuned model. The greedy baseline (first row) uses no beam search and no bias. The beam baseline ($\beta = 0$) isolates the effect of beam search alone. Ent. Det. Acc. = fraction of entity mentions whose name appears in the ASR output (full match, out of 1,374 mentions). The optimal bias ($\beta = 3.0$) is marked in bold.

Bias β	**WER**	**Ent. Det. Acc.**
n/a (greedy)	13.53%	75.8%
0 (beam=5)	14.31%	77.7%
0.5	14.08%	79.0%
1.0	14.02%	79.5%
1.5	14.00%	79.9%
2.0	13.93%	80.5%
3.0	**13.84%**	**80.9%**
5.0	15.06%	80.2%

Bias $\beta = 3.0$ is optimal, improving entity detection accuracy from 77.7% to 80.9% relative to the beam search baseline while also improving WER from 14.31% to 13.84%. Beyond $\beta = 5.0$, the model degenerates rapidly: at $\beta = 10.0$, the model repeatedly inserts entity names throughout

the output regardless of the audio content, producing WER above 900%. This demonstrates that the bias must be carefully calibrated to avoid overwhelming the acoustic signal.

6.5 Combined Approach

Table 6 presents the headline result: combining prompting and shallow fusion.

Table 6: Combined prompting and shallow fusion results on the fine-tuned model. Each row represents a different inference-time configuration applied to the *same* fine-tuned model (no retraining). "Prompt" indicates whether the player list (PL) prompt was used. "Beam" is the number of beam search hypotheses. "Bias" is the shallow fusion strength β. Ent. Det. Acc. = entity detection accuracy (full match, out of 1,374 mentions). The oracle row uses the ground truth transcript as prompt, establishing an upper bound.

Configuration	**Prompt**	**Beam**	**Bias**	**WER**	**Ent. Det. Acc.**
N (greedy)	–	1	–	13.53%	75.8%
N (beam=5)	–	5	–	14.31%	77.7%
PL (greedy)	✓	1	–	12.17%	84.8%
PL (beam=5)	✓	5	–	11.91%	85.0%
SF3 (beam=5)	–	5	3.0	13.84%	80.9%
PL+SF3 (beam=5)	✓	5	3.0	**12.49%**	**88.3%**
GT (oracle)	oracle	1	–	12.17%	90.3%

The combined approach achieves 88.3% entity detection accuracy, only 2.0 percentage points below the oracle of 90.3%. This confirms that prompting and shallow fusion are complementary, not redundant: PL alone gives 84.8%, SF alone gives 80.9%, and their combination yields 88.3%. The best overall WER (11.91%) is achieved by PL with beam search but without SF, as shallow fusion slightly trades WER for entity detection. Crucially, all improvements use the same fine-tuned model with no retraining; only the inference-time configuration changes.

6.6 Seen vs. Unseen Entity Analysis

Table 7 breaks down entity detection accuracy by whether entities appeared in the training data. Of the 231 unique entities in the test set, 155 were also present in the 16 training games ("seen," totaling 1,088 mentions) and 76 appeared exclusively in the test set ("unseen," totaling 276 mentions).

Table 7: Entity detection accuracy (full match) on the fine-tuned model, broken down by training exposure and entity token length. "Seen" entities (155 unique, 1,088 mentions) appeared in at least one of the 16 training games. "Unseen" entities (76 unique, 276 mentions) appear only in the test set. "1-token" = single-word entities (e.g., "Messi"). "2+ tokens" = multi-word entities (e.g., "Thiago Silva"). All three configurations use the same fine-tuned model.

		Entity Det. Accuracy			**By Token Length**	
	WER	**All**	**Seen**	**Unseen**	**1-token**	**2+ tokens**
N (baseline)	13.53%	75.8%	88.0%	27.9%	78.1%	68.2%
PL (prompt)	12.17%	84.9%	92.6%	54.3%	88.3%	73.6%
PL+SF3	**12.49%**	**88.0%**	**93.4%**	**67.0%**	**91.9%**	**75.5%**

Unseen entities are dramatically harder than seen ones: at baseline, the model detects only

27.9% of unseen entities compared to 88.0% of seen entities. The combined approach (PL+SF3) raises unseen entity detection to 67.0%, an improvement of 39.1 percentage points over the baseline, while seen entity detection reaches 93.4%. This directly validates the industry deployment scenario: a single fine-tuned model receiving a match roster at inference time can significantly improve recognition of players the model has never encountered during fine-tuning.

Entity length also affects performance: single-token entities (e.g., "Messi", "Neymar") achieve 91.9% detection with PL+SF3, while multi-token entities (e.g., "Thiago Silva", "De Bruyne") reach 75.5% for full match but 95.0% for partial match. The model typically recognizes at least one word of multi-word names; it is the requirement of matching *all* words that reduces the full-match score.

6.7 Comparison with the Non-Fine-Tuned Model

To investigate the role of fine-tuning, we run the same three configurations (N, PL, PL+SF3) on the original Whisper Medium without any domain adaptation. This comparison is motivated by Piaget et al.'s [2] finding that prompting had a bigger impact on non-fine-tuned models for ATC speech.

Table 8: Side-by-side comparison of the fine-tuned model and the original (non-fine-tuned) Whisper Medium under identical inference-time configurations. Entity Det. Acc. = entity detection accuracy (full match, out of 1,374 mentions). WER values above 100% indicate the model inserts more words than exist in the reference.

	WER		**Entity Det. Accuracy**	
Config	**Fine-tuned**	**Original**	**Fine-tuned**	**Original**
N (baseline)	13.53%	17.09%	75.8%	63.9%
PL (prompt)	12.17%	55.9%	84.8%	66.6%
PL+SF3	12.49%	282.7%	88.3%	67.1%

On the original model, prompting causes WER to increase from 17.09% to 55.9%, and adding shallow fusion raises it further to 282.7%, meaning the model inserts approximately 2.8× more words than exist in the reference. The model hallucinates by inserting player names throughout the transcript where they do not belong. While entity detection accuracy does improve slightly on the original model (63.9% → 67.1%), the overall transcription becomes unusable.

In contrast, the fine-tuned model handles the same prompting gracefully: WER *improves* from 13.53% to 12.17%. Fine-tuning appears to give the model robustness to incorporate roster information without hallucinating.

This contrasts with Piaget et al.'s ATC findings, where prompting helped more on non-fine-tuned models. The difference likely stems from domain difficulty: ATC speech is relatively clean and formulaic, while soccer commentary involves crowd noise, emotional and variable-speed delivery, and a higher density of unusual proper nouns. In the harder domain, the original model is too fragile to benefit from inference-time techniques: it needs the stability provided by domain-specific fine-tuning first.

7 Conclusion

This project demonstrates that pre-trained ASR models degrade significantly on soccer broadcast commentary (17.09% WER vs. 2.43% on clean speech) and that a combination of fine-tuning

and inference-time techniques can substantially improve entity recognition without retraining. Our key findings are:

1. **Fine-tuning** Whisper Medium on 15.66 hours of soccer commentary reduces WER by 21% relative (17.09% → 13.53%). Entity names remain 2.4× harder than non-entity words (22.61% vs. 9.28% WER).
2. **Decoder prompting** with match rosters is the single most effective inference-time technique, improving entity detection accuracy from 75.8% to 84.8% and matching the oracle WER (12.17%).
3. **Shallow fusion** provides complementary gains, and the **combined approach** reaches 88.3% entity detection accuracy (within 2.0 percentage points of the 90.3% oracle) without any retraining.
4. For **unseen entities**, the combined approach raises detection from 27.9% to 67.0%, validating the industry deployment scenario where a match roster is provided at inference time.
5. **Fine-tuning is a prerequisite** for inference-time techniques in this domain: without it, prompting and shallow fusion cause severe hallucination (WER $>$ 282%). This contrasts with prior ATC results where prompting helped non-fine-tuned models, suggesting that domain difficulty determines whether prompting can work without domain adaptation.

These results have practical implications for live sports broadcasting, where accurate transcription of player and team names is critical for automated subtitling, real-time match reports, and content indexing. The inference-time approach demonstrated here allows a single fine-tuned model to serve across different matches by simply updating the entity list before each game.

Future work could explore larger Whisper variants (e.g., Whisper Large), audio preprocessing to reduce crowd noise before ASR, more sophisticated contextual biasing via trie-based approaches [9] or external language models (e.g., through the K2/Icefall framework), hallucination detection and filtering, and extension to multilingual commentary. A companion reproducibility guide provides full instructions for replicating these experiments on the EPFL Izar cluster.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, 2023.

[2] J. Piaget, A. Prasad, and P. Motlicek. Improving ASR and callsign detection in air traffic control speech using Whisper prompting. Technical Report Idiap-RR-04-2025, Idiap Research Institute, 2025.

[3] J. Qi, M. Xu, Z. Li, S. Zhan, and J. Li. GOAL: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of CIKM*, 2023.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, 2015.

[5] A. Deliege, 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. In *CVPR Workshops*, 2021.

[6] HuggingFace. Text generation utilities. https://huggingface.co/docs/transformers/main_classes/text_generation, 2024.

[7] K2-FSA. Icefall: A speech processing toolkit. https://github.com/k2-fsa/icefall, 2023.

[8] M. Bhattacharjee, I. Nigmatulina, A. Prasad, P. Rangappa, S. Madikeri, P. Motlicek, H. Helmke, and M. Kleinert. Contextual biasing methods for improving rare word detection in automatic speech recognition. In *Proceedings of ICASSP*, 2024.

[9] I. Thorbecke, E. Villatoro-Tello, J. Zuluaga-Gomez, S. Kumar, S. Burdisso, P. Rangappa, A. Carofilis, S. Madikeri, P. Motlicek, K. Pandia D S, K. Hacioglu, and A. Stolcke. Unifying global and near-context biasing in a single trie pass. In *Text, Speech, and Dialogue (TSD)*, Springer, 2025.

[10] P. Motlicek, S. Kumar, D. Khalil, A. Prasad, and C. Schüpbach. Leveraging untranscribed data for end-to-end speech and callsign recognition in air-traffic communication. In *SESAR Innovation Days*, 2025.

[11] M. Bhattacharjee, P. Motlicek, S. Madikeri, H. Helmke, O. Ohneiser, M. Kleinert, and H. Ehr. Minimum effort adaptation of automatic speech recognition system in air traffic management. In *European Journal of Transport and Infrastructure Research*, 24(4):133–153, 2025.

[12] J. Zuluaga-Gomez, K. Vesely, I. Szoke, A. Blatt, P. Motlicek, M. Kocour, K. Choukri, I. Nigmatulina, C. Cevenini, A. Tart, J. Cernocky, and D. Klakow. ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications. In *Journal of Data-centric Machine Learning Research*, 2024.

[13] S. Burdisso, E. Villatoro-Tello, S. Kumar, S. Madikeri, A. Carofilis, P. Rangappa, M. K E, K. Hacioglu, P. Motlicek, and A. Stolcke. Reducing prompt sensitivity in LLM-based speech recognition through learnable projection. In *Proceedings of ICASSP*, 2026.

[14] I. Thorbecke, J. Zuluaga-Gomez, E. Villatoro-Tello, S. Kumar, P. Rangappa, S. Burdisso, P. Motlicek, K. Pandia D S, and A. Ganapathiraju. Fast streaming transducer ASR prototyping via knowledge distillation with Whisper. In *Findings of EMNLP*, pages 16747–16762, 2024.