DyadicTickTacking: Multimodal Analysis of Strategy Adoption and Turn-Taking in Cooperative Joint Action

Ali KhaliliKandjani

Department of Computer Science, Bioengineering, Robotics and System Engineering (DIBRIS)

University of Genova

Supervisor

Prof. Gualtiero Volpe

Co-supervisor: Dr. Nicola Corbellini

In partial fulfillment of the requirements for the degree of

Master of Science in Computer Engineering

July, 2026

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Gualtiero Volpe, for giving me the opportunity to work on this project, for his guidance throughout the research process, and for the thoughtful feedback that shaped this thesis into its final form.

I am deeply grateful to my co-supervisor, Dr. Nicola Corbellini, for the close collaboration, constant support, and shared enthusiasm throughout this work. This thesis would not have been possible without his dedication and support.

I thank the team at Casa Paganini, InfoMus laboratory, for providing the facilities and a stimulating research environment.

I am grateful to all 42 participants who volunteered their time for the experiment.

Finally, I thank my family and friends for their encouragement and support.

To the novice I was, and the experts who scaffolded the way.

Abstract

Performing a cooperative task requires partners to coordinate their actions, but does the presence of an expert change how this process develops over time? We address this question through DyadicTick-Tacking, a cooperative computer-based task where two participants seated side by side each control a cursor on a shared screen, navigating together toward common targets across 160 trials organized in four blocks. We compare 10 Novice-Expert dyads (where one partner received prior solo training) and 10 Novice-Novice dyads to characterize how both task execution strategies and verbal coordination co-evolve with task learning.

To this end, we develop a multimodal analysis pipeline combining unsupervised trajectory clustering for strategy identification, a reference-guided speaker diarization system for automated turn-level annotation, and a temporal statistical framework for detecting divergent coordination dynamics between conditions. Our analysis reveals convergent evidence across both modalities.

At the trajectory level, three emergent movement strategies are identified via density-based clustering, forming a validated skill hierarchy ranging from slow exploratory behavior to fast automated execution. A permutation test reveals that expert-novice dyads synchronize their strategies trial-by-trial significantly above chance, indicating genuine real-time interpersonal coordination, while novice-novice

dyads do not exceed chance-level synchrony despite individually improving over time. In parallel, temporal statistical analysis uncovers divergent turn-taking trajectories: expert-novice dyads progressively reduce speech density and shorten their turns as the expert yields conversational control to the novice, whereas novice-novice dyads show no directional verbal structure across blocks. Both conditions learn the task at equivalent rates, yet expert-novice dyads consolidate to superior performance levels.

Critically, these coordination differences are invisible to static between-group comparisons and emerge only when interaction dynamics are tracked longitudinally across practice blocks. These multimodal findings demonstrate that expert scaffolding is a temporal process, manifesting as a progressive transfer of control across both action and speech, with implications for the design of adaptive collaborative systems that detect and support coordination development in real time.

Contents

1 Introduction **1**

1.1 Motivations 1

1.2 Context of the Study 2

1.3 Objectives and Contributions 4

1.4 Overview of the Thesis 5

2 Background and Related Work **6**

2.1 Joint Action and Interpersonal Coordination 6

2.2 Turn-Taking in Dyadic Interaction 8

2.3 Speaker Diarization 10

2.4 Unsupervised Clustering for Behavioral Analysis 11

2.5 Summary of Research Gaps 12

3 The DyadicTickTacking Paradigm and Experimental Setup **16**

3.1 The DyadicTickTacking Paradigm 16

3.1.1 Task Overview 17

3.1.2 Control Mechanics 17

3.1.3 Proximity-Based Cooperative Coupling 19

3.2 Participants and Experimental Design 19

3.3 Procedure 21

3.4 Data Collection 22

4 Methodology 24

4.1 Analysis Pipeline Overview 24

4.2 Learning Curve Analysis 27

4.3 Trajectory Strategy Analysis 27

4.3.1 Canonical Coordinate Frame Transformation 28

4.3.2 Feature Engineering 28

4.3.3 Block-by-Block Unsupervised Clustering 29

4.3.4 Cluster Alignment Across Blocks 30

4.3.5 Strategy Performance Validation 30

4.3.6 Dyadic Strategy Joint Adoption 30

4.4 Turn-Taking Analysis 31

4.4.1 Speaker Diarization Pipeline 31

4.4.1.1 Stage 1: Neural Segmentation 32

4.4.1.2 Stage 2: Speaker Embedding Extraction 33

4.4.1.3 Stage 3: Agglomerative Clustering 33

4.4.1.4 Stage 4: Local Re-Segmentation 33

4.4.1.5 Stage 5: Reference-Guided Speaker Assignment . 34

4.4.2 Turn-Taking Feature Extraction 36

4.5 Between-Condition Statistical Comparison 38

4.5.1 One-Way ANOVA 38

4.5.2 Mixed-Design ANOVA 38

4.5.3 Post-Hoc Tests and Robustness Checks 38

5 Results and Discussion 40

5.1 Learning Curves 40

5.2 Unsupervised Strategy Profiles 42

5.3 Longitudinal Strategy Evolution 43
5.4 Dyadic Strategy Synchrony 44
5.5 Turn-Taking Dynamics Across Conditions 47
5.6 Prosodic Features Across Conditions 48
5.7 Back-Channel Filtering Results 49
5.8 General Discussion 49

6 Conclusions **52**
6.1 Summary of Findings 52
6.2 Limitations 54
6.3 Future Directions 55

References **61**

List of Figures

3.1 Two snapshots of a single trial: colored dots are the cursors, the circle is the shared target, and lines show example trajectories. . . 18

3.2 Experimental recording setup. Two participants seated side by side at the same workstation, with a conference microphone positioned centrally between them and an OAK-D camera recording from across the table. 23

4.1 Overview of the analysis pipeline. Blue boxes indicate data processing and feature extraction; red boxes indicate statistical testing. The trajectory track (left) and audio track (right) operate independently. 25

5.1 Mean trial completion time per block for NE and NN dyads. Error bars indicate standard error. Both groups improve sharply from Block 1 to Block 2, then plateau at different asymptotes. 41

5.2 Within-dyad performance variability across blocks. Each bar represents the standard deviation of trial completion times computed within each dyad, then averaged across the 10 dyads per condition. Error bars indicate standard error of these per-dyad standard deviations across dyads. 42

5.3 Adoption frequency of strategies in each experimental condition across trial blocks. NE dyads transition to efficient strategies faster and maintain them, while NN dyads regress toward Unskilled by Block 4. 44
5.4 Permutation test results. Grey histograms show the null distribution of synchrony rates from 100,000 shuffled partner assignments. Red lines indicate the observed synchrony. Left: NE dyads synchronize well above chance ($p < 0.001$). Right: NN dyads do not significantly exceed chance ($p = 0.072$). 45
5.5 Overall dyadic strategy synchrony over blocks. NE dyads increase synchrony from Block 1 onward, while NN dyads decline from their initial peak. The lines cross at Block 2, after which NE consistently exceeds NN. 46
5.6 Cluster-specific dyadic synchrony over blocks. Left: NE dyads shift from synchronizing on Cluster 0 (Unskilled) to Cluster 2 (Proficient) by Block 4. Right: NN dyads lose their initial trivial Cluster 0 alignment without developing synchrony on efficient strategies. . 46
5.7 Left: voice activity per block for Expert, Novice (NE), and Novice (NN). Right: average turn duration per block for NE and NN dyads. Error bars indicate standard error. 48

Chapter 1

Introduction

1.1 Motivations

Human cooperation is built on the ability to coordinate actions with others toward shared goals. Whether carrying an object together, playing a musical duet, or navigating a shared workspace, individuals must continuously align their movements, predict each other's intentions, and adapt their strategies in real time [1]. This capacity for joint action is not merely a specialized skill of experts in sports or music; it is a fundamental aspect of everyday social interaction that emerges early in development and underlies much of human collaborative behavior.

A particularly interesting case arises when collaborating partners differ in expertise. In educational, clinical, and workplace settings, novices frequently learn by working alongside more experienced individuals. The more skilled partner can scaffold the learning process, providing structure, guidance, and feedback that facilitates the novice's acquisition of new competencies [2]. This scaffolding often operates through both explicit verbal instruction and implicit behavioral modeling: the expert demonstrates efficient strategies while gradually transferring responsibility to the novice as competence develops. This process, first theorized by Vygotsky as operating within the learner's zone of proximal development [3], has been extensively studied in pedagogical contexts but has received less attention in the domain of real-time cooperative task execution.

Understanding how expertise asymmetry shapes the development of cooperative coordination has broad practical implications. In **collaborative training**, insights into how experts structure interaction could inform the design of pair-based learning programs that maximize skill transfer. In **rehabilitation**, where patients often practice motor tasks alongside a therapist, understanding the temporal dynamics of scaffolding could help optimize therapeutic protocols. In the design of **adaptive interactive systems**, such as collaborative robots or intelligent tutoring systems, characterizing the multimodal signatures of successful scaffolding could enable technology that detects coordination patterns and adjusts its support in real time, mimicking the natural behavior of a human expert.

Yet achieving these goals requires first understanding the underlying dynamics of expert-novice coordination. This raises a fundamental question: how does the presence of an expert partner shape not just immediate task performance, but the temporal trajectory of both task coordination and verbal communication as learning unfolds? This thesis addresses this question by developing a multimodal analysis framework that jointly investigates trajectory-level task strategies and turn-taking dynamics in a cooperative task, comparing dyads where one partner has prior expertise with dyads where both partners are novices.

1.2 Context of the Study

This work is situated within the interdisciplinary field of Social Signal Processing (SSP), which investigates how social and communicative phenomena can be computationally modeled from behavioral signals such as speech, gesture, gaze, and body movement [4]. Within SSP, a central theme is the study of interpersonal coordination: how interacting individuals dynamically influence each other's behavior, cognition, and timing during collaborative activities.

One of the most extensively studied dimensions of interpersonal coordination is turn-taking, the mechanism by which speakers alternate in conversation. Turn-taking is not merely a linguistic phenomenon but a multimodal one, regulated through prosody, gaze, gesture, and timing cues [5]. The temporal properties of speaker transitions, including gaps, overlaps, and response latencies [5], have been shown to reflect the quality and efficiency of dyadic interaction [6]. Beyond speech, coordination in joint action also manifests in the strategies partners adopt for task execution, though few computational approaches have been developed to capture how these strategies emerge and align between partners over time.

Recent work by Haraped et al. [7] investigated how differential expertise affects multimodal coordination in dyads performing a joint visual task. Their findings revealed that although more knowledgeable participants dominated the dialogue by talking and pointing more, the underlying coupling dynamics remained symmetric overall. Asymmetric coordination emerged only when data were disaggregated by conversational turns: the momentary speaker led the joint focus of attention regardless of expertise level. This suggests that expertise effects on coordination are not static but temporally structured, becoming visible only when the dynamics of interaction are analyzed at a finer temporal granularity.

However, despite progress in computational modeling of turn-taking and expertise-driven coordination, several open challenges remain. First, most studies examine either task coordination or verbal interaction in isolation; a recent systematic review of spontaneous interpersonal coordination confirms that the field remains fragmented, with existing work focusing predominantly on motor synchrony while speech-based coordination dynamics are studied in separate research traditions [4; 8; 9]. Second, existing studies of expertise-driven coordination tend to focus on single interactions, providing a snapshot of coordination at one point in time rather than tracking how it develops longitudinally over repeated practice; notably, all 19 studies reviewed by Ayache et al. examine single sessions [9], and recent expert-novice work similarly captures only one interaction [7]. Third, while scaffolding theory provides a rich conceptual framework for understanding expert-novice interaction [2; 10], the specific mechanisms through which an expert's scaffolding manifests behaviorally, both in action and in speech, have not been systematically characterized using computational methods over extended learning periods.

1.3 Objectives and Contributions

This thesis aims to investigate how expertise asymmetry shapes the longitudinal development of cooperative coordination across both task execution and verbal communication. Specifically, we address three research questions:

1. Does coupling with an expert improve task performance, and if so, does it accelerate learning or shift the performance level?
2. Do participants adopt distinct task strategies over time, and how does expert presence shape the emergence and joint adoption of these strategies?
3. How do turn-taking behaviors co-evolve with task learning, and do the two experimental conditions develop different verbal coordination patterns across practice blocks?

To answer these questions, we develop a multimodal analysis pipeline comprising unsupervised trajectory clustering for strategy identification, a reference-guided speaker diarization system for automated turn-level annotation, and a temporal statistical framework for detecting divergent coordination dynamics between conditions.

The results of this investigation are presented and discussed in Chapter 5, with implications for both theoretical understanding of scaffolding and the design of adaptive collaborative systems.

1.4 Overview of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 provides the theoretical and technical background, covering joint action theory, turn-taking in conversation, speaker diarization methods, and unsupervised clustering approaches. Chapter 3 describes the DyadicTickTacking paradigm, the experimental protocol, participant characteristics, and the multimodal data collection setup. Chapter 4 presents the methodology in detail, including the trajectory feature engineering and clustering pipeline, the speaker diarization system with local re-segmentation, the turn-taking feature extraction, and the statistical analysis framework. Chapter 5 reports and discusses the results across learning curves, unsupervised strategy profiles, longitudinal strategy evolution, dyadic strategy synchrony, and turn-taking dynamics. Finally, the Conclusions chapter summarizes the main findings, discusses limitations, and outlines directions for future work.

Chapter 2

Background and Related Work

Summary

This chapter provides the theoretical foundations and reviews prior work relevant to this thesis. We first introduce joint action and scaffolding theory, then review computational approaches to turn-taking analysis and speaker diarization, and discuss unsupervised methods for behavioral strategy identification. Finally, we identify research gaps by examining how recent studies have addressed similar goals and where they fall short, motivating the contributions of this thesis.

2.1 Joint Action and Interpersonal Coordination

Joint action refers to any form of social interaction in which two or more individuals coordinate their actions in space and time to bring about a change in the environment [1]. Successful joint action depends on three core abilities: sharing representations of the task and each other's roles, predicting what the partner will do next, and integrating predicted effects of one's own and the other's actions into a unified plan. These mechanisms allow individuals to act in anticipation of their partner rather than merely reacting, enabling the fluid coordination observed in everyday activities such as jointly carrying an object [1] or playing music in ensemble, where performers must predict each other's timing with millisecond precision [11].

A central distinction in joint action research is between symmetric and asymmetric coordination [1]. In symmetric joint action, both partners contribute equally and hold equivalent roles, as in the mirror game paradigm where two individuals improvise movements together without a designated leader [12]. In asymmetric joint action, partners differ in skill, knowledge, or assigned responsibility [7]. The latter case is particularly relevant to learning contexts, where a more competent partner can structure the interaction to facilitate the less experienced partner's development.

The concept of scaffolding, introduced by Wood, Bruner and Ross [10], describes the process by which a more skilled individual provides temporary structured support that enables a novice to accomplish tasks beyond their current independent ability. The scaffold is not fixed: it is calibrated to the learner's current competence and progressively withdrawn as the learner gains autonomy. This concept builds on Vygotsky's zone of proximal development (ZPD) [3], which defines the space between what a learner can do independently and what they can achieve with guidance. Scaffolding operates within this zone, providing just enough support to bridge the gap without removing the learner's active participation.

While scaffolding has been extensively studied in educational settings, including classroom instruction [13], tutoring [10], and parent-child interaction [3], its manifestation in real-time cooperative task execution remains less explored from a computational perspective. In particular, how scaffolding operates simultaneously through verbal coordination (what partners say and when) and through action coordination (what strategies partners adopt and how they align) has not been systematically characterized over extended learning periods.

Recent work has begun to address how expertise asymmetry affects coordination in dyadic tasks. Haraped et al. [7] studied expert-novice dyads performing a joint visual task and found that more knowledgeable participants dominated the dialogue by talking and pointing more, yet the underlying attentional coupling dynamics remained symmetric overall. Asymmetric coordination emerged only when data were disaggregated by conversational turns, with the momentary speaker leading the joint focus of attention regardless of expertise level. This suggests that expertise effects are temporally structured rather than static, motivating longitudinal approaches that track coordination over repeated interactions.

2.2 Turn-Taking in Dyadic Interaction

Turn-taking is the fundamental mechanism by which speakers alternate in conversation. First formalized by Sacks, Schegloff and Jefferson [14], the turn-taking system governs how participants coordinate the smooth transfer of speaking rights, minimizing both gaps (silence between turns) and overlaps (simultaneous speech). At each potential transition point, speakers use a combination of linguistic, prosodic, and nonverbal cues to signal whether they intend to continue or yield the floor.

Heldner and Edlund [5] provided a comprehensive analysis of the temporal properties of turn transitions, documenting the distributions of gaps, overlaps, and within-turn pauses across different conversational contexts. Their work established that turn-taking timing is remarkably precise: most transitions occur within a few hundred milliseconds, indicating that listeners begin planning their response well before the current speaker finishes. The 500 ms threshold has become a standard criterion for distinguishing between within-turn pauses and between-turn gaps in computational work.

From a computational perspective, turn-taking features have been used as indicators of interaction quality, dominance, and coordination. Jayagopi et al. [15] demonstrated that nonverbal activity cues including speaking time, turn duration, and turn rate could predict perceived dominance in group conversations. They introduced the concept of Total Speaking Turns without Short Utterances (TST-woSU), filtering out back-channel responses shorter than 1 second, which proved more robust for dominance classification than raw turn counts. Raman et al. [6] extended this line of work to perceived conversation quality in spontaneous interactions, defining speaking turns as continuous segments separated by at least 500 ms of silence and demonstrating that turn-taking patterns predict how participants perceive the quality of their interaction.

Sequential structure beyond individual turns also carries coordination information. Modeling the speaker sequence as a first-order Markov chain reveals transition probabilities between speakers: high alternation ratios indicate rapid back-and-forth exchange, while high self-transition probabilities indicate floor-holding behavior. The conditional entropy of the transition matrix quantifies how predictable or structured the turn-taking pattern is, distinguishing rigid alternation from more flexible, asymmetric conversational dynamics.

The growing interest in computational modeling of dyadic turn-taking is evidenced by recent community benchmarks such as the ACII-DaiKon challenge [16], which specifically targets turn-taking prediction and rapport trajectory modeling in naturalistic dyadic conversations across multiple languages.

2.3 Speaker Diarization

Speaker diarization is the task of partitioning an audio recording into segments according to speaker identity, answering the question ”who spoke when?” without prior knowledge of the number of speakers or their voice characteristics [17]. In the context of dyadic interaction analysis, reliable diarization is a prerequisite for computing per-speaker turn-taking features, speaking ratios, and sequential structure metrics.

Modern diarization pipelines typically follow a four-stage architecture. First, a voice activity detection (VAD) module identifies speech regions. Second, the audio is segmented into short windows, and speaker embeddings are extracted from each window using neural models trained to map speech to a fixed-dimensional vector space where same-speaker segments cluster together. Third, these embeddings are grouped using clustering algorithms (agglomerative, spectral, or k-means) to assign speaker identities. Fourth, the resulting segments are post-processed to merge short gaps and resolve inconsistencies.

The pyannote.audio framework [18] implements this architecture using a neural segmentation model with a sliding window approach, producing frame-level speaker activity probabilities. Speaker embeddings are extracted using models such as WeSpeaker ResNet34-LM [19], which produces 256-dimensional representations trained on the VoxCeleb2 dataset (5994 speakers). Agglomerative hierarchical clustering with centroid linkage groups embeddings into speaker clusters, with the number of speakers either specified or estimated from the data.

Several challenges affect diarization quality in naturalistic recordings. Overlapping speech, where both speakers talk simultaneously, degrades embedding quality because standard extraction assumes one active speaker. Short segments provide insufficient audio for reliable embedding computation. Most critically for long recordings, the scale-dependent behavior of agglomerative clustering means that centroid positions shift as more data is processed, potentially causing speaker transitions near the decision boundary to be incorrectly merged. These challenges motivate the pipeline extensions developed in this thesis.

2.4 Unsupervised Clustering for Behavioral Analysis

Unsupervised clustering methods allow researchers to discover natural groupings in behavioral data without requiring predefined categories or labeled examples. In the context of motor learning and strategy identification, clustering can reveal distinct movement strategies that participants adopt, track how strategy use evolves over time, and quantify coordination between partners.

Density-based clustering methods are particularly suitable for behavioral data because they do not require specifying the number of clusters in advance, can handle clusters of irregular shape and varying density, and explicitly identify outlier points rather than forcing them into inappropriate groups. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [20] extends the classical DBSCAN algorithm by building a cluster hierarchy across all density levels and extracting the most stable clusters. The single parameter `min_samples` controls the minimum neighborhood density required to form a cluster, providing a balance between sensitivity and robustness.

When clustering is applied independently to sequential time periods (such as blocks of trials in a learning experiment), cluster labels are arbitrary across periods: a cluster labeled "0" in one block may correspond to a cluster labeled "2" in the next block. Longitudinal tracking therefore requires alignment of cluster identities across time. The Hungarian algorithm (linear sum assignment) [21] provides an optimal solution by computing the minimum-cost bipartite matching between cluster centroids of consecutive blocks, ensuring that equivalent strategies maintain consistent labels throughout the session.

Feature engineering for trajectory data typically involves transforming raw positional coordinates into interpretable metrics that capture movement quality, efficiency, and control strategy. Common approaches include computing path efficiency, deviation from ideal paths, speed profiles, and correction frequency. Canonical coordinate frame transformations that normalize for task geometry (target angle, distance) ensure that features reflect movement strategy rather than spatial configuration.

Applications of unsupervised clustering in behavioral analysis span motor learning (identifying skill-level transitions through movement pattern classification [22]), gait analysis (detecting walking pattern changes in clinical populations [23]), and sports science (classifying tactical strategies from player tracking data [24]). In dyadic contexts, comparing cluster assignments between partners enables quantification of strategy synchrony, the degree to which partners adopt the same behavioral mode on the same trial.

2.5 Summary of Research Gaps

The literature reviewed above reveals several converging gaps that this thesis aims to address. In what follows, we identify each gap by describing how existing related work has approached similar goals and where it falls short.

The Multimodal Gap

A recent systematic review of spontaneous interpersonal coordination [9] confirms that the field remains fragmented: all 19 studies reviewed focus exclusively on motor coordination (body sway, head movement, postural synchrony), while speech-based coordination dynamics are studied in entirely separate research traditions [4]. Vink et al. [8] studied movement synchrony in dyadic interaction using nonlinear time series methods, but without incorporating any speech analysis. Conversely, Raman et al. [6] demonstrated that turn-taking features predict perceived conversation quality, but without measuring task-level coordination. Onishi et al. [25] used multimodal deep learning to predict turn-taking events in expert-novice conversations, demonstrating that expertise asymmetry creates detectable patterns in vocal and visual behavior; however, their work frames turn-taking as a static classification problem rather than tracking how patterns evolve with learning, and does not incorporate task-level strategy analysis. Even Haraped et al. [7], who analyzed both gaze and speech in expert-novice dyads using Cross Recurrence Quantification Analysis, treated these channels independently rather than examining their co-evolution over time. No existing framework jointly analyzes task-level strategy adoption and speech coordination dynamics within the same experiment, limiting our understanding of how action and verbal coordination develop together.

The Longitudinal Gap

Existing computational studies of expertise-driven coordination consistently examine single interactions. Haraped et al. [7] captured one session per dyad, finding that expertise effects only emerge when temporal dynamics are disaggregated by conversational turns. Ayache et al. [9] reviewed a decade of interpersonal coordination research and all 19 selected studies were single-session. Onishi et al. [25] trained models on individual conversations from the NoXi database without examining how expert-novice turn-taking dynamics develop across repeated interactions. Vesper et al. [26] demonstrated that partners make themselves more predictable to facilitate coordination, but measured this within a single block rather than tracking how such adaptations develop over repeated practice. How expert presence shapes the developmental trajectory of coordination, how patterns emerge, consolidate, and potentially regress across extended practice, remains largely uncharacterized. Single-session studies may systematically underestimate the role of expertise in shaping coordination dynamics.

The Computational Scaffolding Gap

While scaffolding theory provides a rich conceptual framework [3; 10], computational characterization of its behavioral signatures is lacking. Fawcett and Garton [2] studied peer collaboration and expertise effects on problem-solving performance, demonstrating that working with a more skilled partner improves outcomes, but their analysis was observational and did not computationally track how scaffolding unfolds over time. Plank et al. [27] developed an automated pipeline for extracting speech and turn-taking parameters from dyadic conversations, demonstrating that computational turn-taking analysis can distinguish between populations; however, they applied it to diagnostic classification rather than to characterizing coordination dynamics or scaffolding progress. Van de Pol et al. [13] reviewed a decade of scaffolding research in teacher-student interaction, identifying key strategies such as modeling, hints, and fading, yet noted that effectiveness studies remain scarce and no computational framework exists to automatically detect scaffolding progress from behavioral data. The specific mechanisms through which an expert's scaffolding manifests behaviorally, simultaneously in task strategies and in verbal coordination, have not been systematically characterized using computational methods over extended learning periods. Identifying measurable multimodal signatures of scaffolding progress would enable both scientific understanding and engineering applications in adaptive system design.

This thesis addresses all three gaps by combining unsupervised trajectory clustering with automated speaker diarization and temporal statistical analysis within a single cooperative paradigm, tracking how both movement strategies and turn-taking patterns co-evolve across four practice blocks in expert-novice versus novice-novice dyads.

Chapter 3

The DyadicTickTacking Paradigm and Experimental Setup

Summary

This chapter describes the experimental platform, the participant sample, and the data collection protocol. We first introduce the DyadicTickTacking paradigm and its control mechanics, then detail the two experimental conditions, the session procedure, and the multimodal data recorded.

3.1 The DyadicTickTacking Paradigm

The experiment was conducted using an adapted version of the cooperative motor paradigm introduced by Rocchesso et al. [28]. The platform was designed to study how two co-located participants coordinate their actions in a shared task that requires both individual skill and interpersonal cooperation.

3.1.1 Task Overview

Two participants, seated side by side at the same workstation, each controlled an independent cursor represented as a colored dot on a shared full-screen display. The objective of each trial was for both participants to navigate their cursors to a common circular target as quickly as possible. A trial was considered complete only when both dots had reached the target zone, requiring joint success rather than individual achievement.

Targets were presented at a fixed distance of 350 pixels from the screen center, placed at one of eight evenly spaced angular positions: 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°. Each target location was visited an equal number of times across the session in a pseudorandomized order that was fixed across all participants to ensure comparability.

3.1.2 Control Mechanics

At trial onset, both dots were repositioned to the screen center and assigned an initial movement direction drawn from a predefined sequence, fixed across participants. The two dots were initialized with horizontally mirrored trajectories, ensuring that participants started each trial facing distinct directional challenges.

Cursor direction and speed were controlled through *key-press duplets*: ordered pairs of consecutive keypresses produced within a maximum temporal window of 5 seconds. Each participant was provided with a three-key keyboard and instructed to use only the leftmost (L) and center (R) keys. This gave each participant access to four ordered two-key combinations, each mapped to a distinct cardinal direction of motion:

- **LL** (Left-Left): movement to the left (negative v_x)
- **RR** (Right-Right): movement to the right (positive v_x)

- **LR** (Left-Right): movement downward (negative v_y)
- **RL** (Right-Left): movement upward (positive v_y)

Each duplet updated exactly one component of the dot's velocity vector, either horizontal (v_x) or vertical (v_y), while leaving the orthogonal component unchanged. Identical-key duplets (LL, RR) governed the horizontal axis, while mixed-key duplets (LR, RL) governed the vertical axis. The magnitude of the updated velocity component was set inversely proportional to the Inter-Onset Interval (IOI) between the two keypresses, with a lower bound of 0.2 seconds to prevent excessively high velocities. Because the two velocity components were updated independently and persisted until explicitly overridden by a subsequent duplet, diagonal motion toward the target emerged naturally from the temporal interleaving of duplets across both axes.

This control scheme imposed a non-trivial motor challenge: participants needed to learn which key combinations produced which directions, and to develop temporal precision in their keypresses to control speed effectively. Figure 3.1 illustrates two snapshots of a single trial.

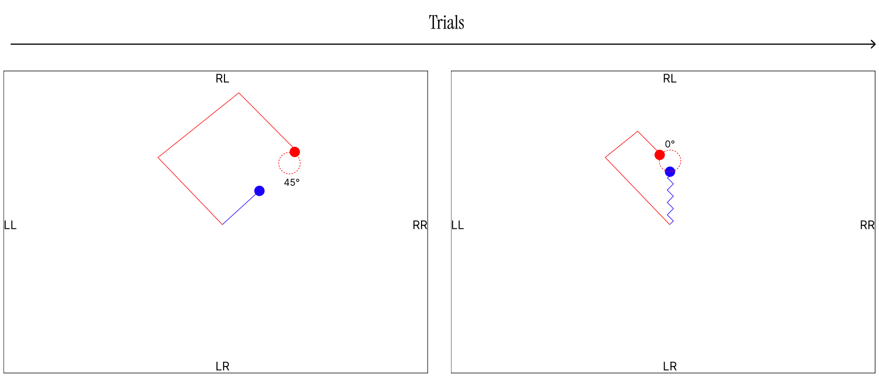

Figure 3.1: Two snapshots of a single trial: colored dots are the cursors, the circle is the shared target, and lines show example trajectories.

3.1.3 Proximity-Based Cooperative Coupling

To introduce a cooperative incentive that required interpersonal coordination rather than independent parallel play, a proximity-based velocity coupling mechanism was imposed between the two cursors. Whenever the Euclidean distance between the two dots exceeded a threshold of 150 pixels, both velocity vectors were attenuated by a factor of 0.7 per update cycle (applied every 20 milliseconds), with a minimum speed floor preserved to prevent complete arrest of motion.

This mechanic penalized excessive spatial separation between the two cursors, implicitly incentivizing participants to maintain proximity and coordinate their movement strategies. Participants in both conditions were informed about the coupling mechanism prior to the task, but did not experience it during the familiarization phase. The expert in the NE condition had prior proficiency with the control mechanics from Day 1 solo training (which did not include the coupling), but experienced the cooperative constraint for the first time alongside their novice partner during the Day 2 trials.

The coupling mechanism created a fundamental tension in the task: each participant needed to make progress toward the target (requiring independent movement) while simultaneously staying close to their partner (requiring coordination). This tension is what makes the task a genuine joint action rather than two parallel individual tasks performed on the same screen.

3.2 Participants and Experimental Design

A total of 42 healthy adults with normal or corrected-to-normal vision (age range: 24–44 years; $\mu = 28.4$, $\sigma = 5.1$) were recruited from university student networks and their contacts, and organized into 21 dyads. One dyad withdrew before completing the experiment and was excluded from all analyses, yielding a final sample of 40 participants across 20 dyads. Participants came from diverse cultural backgrounds, including Spanish, Italian, Iranian, Moroccan, Pakistani, Cameroonian, Indonesian, and Syrian nationals. A subset of dyads (86%) were previously acquainted prior to participation. Dyads sharing a common native language were free to communicate in that language; the remainder (3 NE dyads and 2 NN dyads) communicated in English.

Dyads were assigned to one of two between-dyad experimental conditions, with 10 dyads per condition:

- **Novice–Expert (NE) condition:** one member of the dyad was designated as the Expert based on availability to attend an additional session. On Day 1, this participant completed the task individually in a solo, non-cooperative mode (without the proximity-based coupling) in order to acquire task proficiency. The following day (Day 2), the Expert returned with their partner (the Novice), who had no prior exposure to the task, and they performed the cooperative version together.

- **Novice–Novice (NN) condition:** both members of the dyad attended only Day 2 and performed the cooperative task together without any prior individual training. Neither participant had previous experience with the paradigm.

This design allows direct comparison between dyads where one partner brings prior task knowledge and dyads where both partners learn the cooperative task from scratch simultaneously. The study was approved by the Ethics Committee for University Research (CERA) of the University of Genoa. All participants provided written informed consent prior to participation. They were informed that participation was voluntary, carried no academic or work advantage, and that they could withdraw at any time without consequence. Data were collected in anonymized form using alphanumeric identification codes; no personal information is linked to experimental data. Audio and video recordings are stored on password-protected drives within the InfoMus laboratory at DIBRIS, accessible only to the research team.

3.3 Procedure

Each session began with a 5-minute free-roaming familiarization phase in which participants could explore the control mechanics without time pressure or performance goals. During this phase, the four edges of the screen displayed the key combination required to move in that respective direction, and two arrows at the screen center indicated the cursor's current movement vector as a combination of horizontal and vertical components. This visual scaffolding was removed once the main experimental trials began.

In dyadic sessions, the familiarization phase was conducted without the proximity-based coupling active, allowing participants to acclimate to the shared display and to each other's presence before the cooperative constraints were applied. For the NE group, the Day 1 familiarization was performed individually; the Day 2 familiarization was performed cooperatively.

Following familiarization, participants completed 160 trials divided into four blocks of 40 trials each. Individual trials were separated by a 3-second inter-trial interval, during which the cursors were reset to the screen center. Blocks were separated by a 3-minute rest period, during which participants were free to talk, stretch, or rest. Individual training sessions for NE experts on Day 1 took on average 47.1 minutes ($\sigma = 6.6$; range: 37–58 minutes). Cooperative sessions on Day 2, common to both NE and NN conditions, took on average 81.2 minutes ($\sigma = 22.2$; range: 47–126 minutes).

3.4 Data Collection

Multimodal data were collected throughout each cooperative session, comprising audio, video, and trajectory recordings.

Audio. Stereo audio was captured at 44.1 kHz using a conference microphone positioned centrally between the two participants on the table surface. This placement ensured balanced capture of both speakers' voices while minimizing ambient noise. In recordings where persistent background speech was present, spectral noise reduction was applied as a preprocessing step prior to diarization: a noise profile was captured from a section containing only the interfering voice, then subtracted from the full audio using Sound eXchange (SoX), a command-line audio processing utility, with a reduction factor of 0.65 to attenuate the background while preserving the main speakers' quality. The audio recordings were then used for speaker diarization and turn-taking analysis.

Video. Video was recorded at 52 frames per second using a Luxonis OAK-D camera programmed for automated capture. The camera was positioned across the table from the participants to record an upper-body frontal view of both seated individuals. Video data was collected for potential future analysis of gesture and facial expressions, though it was not analyzed in the present work.

Trajectory data. Cursor position data (x, y coordinates for both dots) were logged by the game software at each update cycle (every 20 ms), along with timestamps of all keypresses. These trajectory logs constitute the primary data source for the movement strategy analysis presented in Chapter 4.

Figure 3.2 illustrates the physical recording setup during a cooperative session.

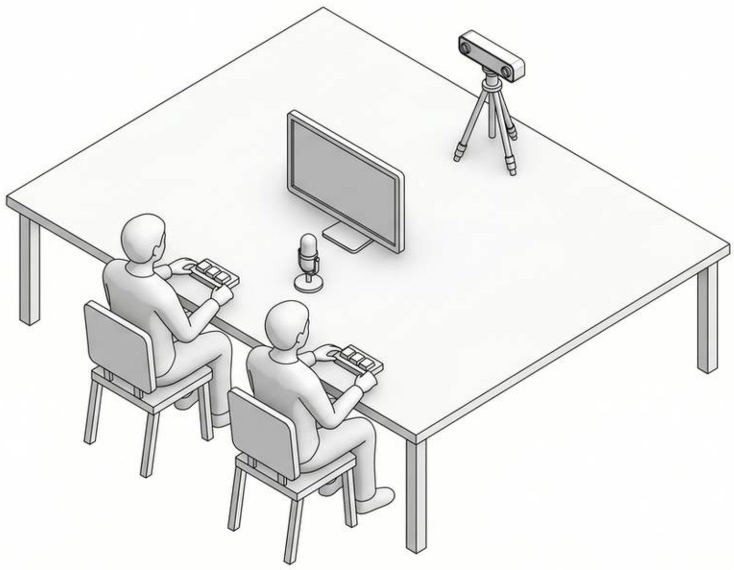

Figure 3.2: Experimental recording setup. Two participants seated side by side at the same workstation, with a conference microphone positioned centrally between them and an OAK-D camera recording from across the table.

Chapter 4

Methodology

Summary

This chapter presents the full analysis methodology. The pipeline processes multimodal data along two parallel tracks: trajectory data is transformed, clustered, and evaluated for dyadic strategy coordination, while audio data is processed through speaker diarization and turn-taking feature extraction. Each track employs its own statistical tests suited to the nature of the data and the research questions addressed. We first provide an overview of the complete pipeline, then detail each component in the order it is applied.

4.1 Analysis Pipeline Overview

The analysis framework processes multimodal data from each dyadic session along two parallel tracks: a trajectory track that addresses the first two research questions (whether expert presence improves task performance, and whether participants adopt distinct strategies whose evolution and joint adoption differ between conditions), and an audio track that addresses the third (whether turn-taking behaviors co-evolve with task learning differently across conditions). Additionally, trial completion times are analyzed directly as learning curves to establish baseline performance differences between conditions. Each track applies statistical methods appropriate to its data structure and research question. Figure 4.1 illustrates the complete pipeline.

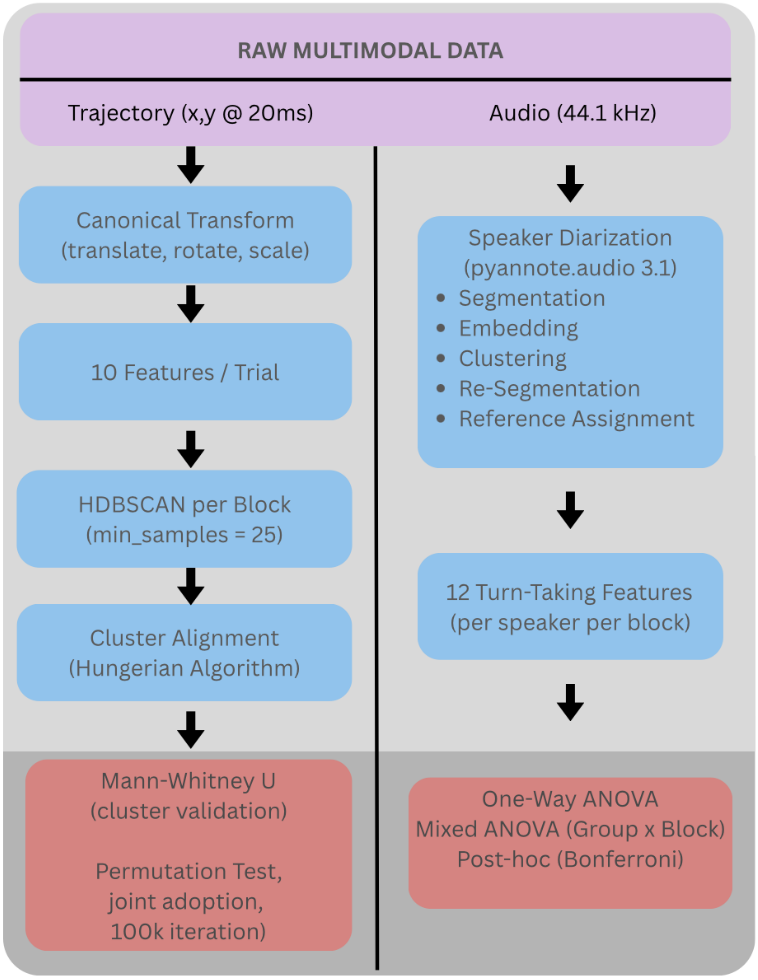

Figure 4.1: Overview of the analysis pipeline. Blue boxes indicate data processing and feature extraction; red boxes indicate statistical testing. The trajectory track (left) and audio track (right) operate independently.

The trajectory track takes raw cursor position logs (x, y coordinates sampled every 20 ms) and transforms them into a canonical coordinate frame that removes target angle and distance confounds. From these normalized trajectories, 10 features are engineered per trial to characterize path geometry, correction behavior, forward progress, and kinematics. HDBSCAN clustering is then applied independently to each block of 40 trials, revealing emergent movement strategies. Clusters are aligned across blocks using the Hungarian algorithm to enable longitudinal tracking. The resulting strategy labels are validated against execution time using one-sided Mann-Whitney U tests and used to quantify dyadic joint adoption through a within-condition permutation test (100,000 iterations).

The audio track takes raw stereo recordings (44.1 kHz) and passes them through a five-stage speaker diarization pipeline built on pyannote.audio 3.1, which produces speaker-labeled segments with consistent identities. From these segments, 12 turn-taking features are extracted per speaker per block. The turn-taking features are then analyzed using one-way ANOVA on aggregated values, mixed-design ANOVA (Group $\times$ Block) to capture temporal divergence between NE and NN conditions, Bonferroni-corrected post-hoc tests, and non-parametric slope comparisons as robustness checks.

Trial-level learning curves are analyzed separately using Wilcoxon signed-rank tests for within-group block-to-block transitions and Mann-Whitney U tests for between-group comparisons at each block.

The remainder of this chapter details each component: Section 4.2 describes the learning curve analysis, Section 4.3 presents the trajectory strategy pipeline, Section 4.4 covers the speaker diarization system and turn-taking feature extraction, and Section 4.5 describes the statistical framework applied to the turn-taking features.

4.2 Learning Curve Analysis

Task performance is quantified as the average trial completion time per block (40 trials per block, 4 blocks per session). Completion time is defined as the interval from trial onset (when both dots are placed at the screen center) to trial completion (when both dots have entered the target zone).

To investigate where learning occurs within the session, we apply Wilcoxon signed-rank tests on within-group block-to-block transitions ($B_1 \to B_2$, $B_2 \to B_3$, $B_3 \to B_4$, and $B_1 \to B_4$). This non-parametric paired test is appropriate because each dyad is measured at all four blocks, and the small sample size (10 dyads per condition) does not guarantee normality of the difference scores.

Between-group differences at each block are assessed using the Mann-Whitney U test, which compares the distributions of dyad-averaged completion times between NE and NN conditions without assuming equal variances or normal distributions.

Together, these analyses establish whether the two conditions learn at different speeds (different slopes of improvement), plateau at different performance levels (different asymptotes), or both.

4.3 Trajectory Strategy Analysis

This section describes the pipeline for identifying and tracking movement strategies from raw cursor trajectories, from coordinate normalization through feature extraction, clustering, validation, and dyadic synchrony analysis.

4.3.1 Canonical Coordinate Frame Transformation

Raw cursor trajectories cannot be compared directly across trials because targets appear at different angular positions. A trajectory moving toward a target at 0° (rightward) looks entirely different from one moving toward 135° (upper-left), even if the participant applied the same control strategy. To eliminate this spatial confound, we project each trajectory into a canonical coordinate frame where the target angle and distance are normalized away.

Given a trajectory of N points $\{(x_i, y_i)\}_{i=1}^{N}$, let $\mathbf{d} = \mathbf{p}_{\text{target}} - \mathbf{p}_1$ be the vector from the start position to the target, $D = \|\mathbf{d}\|_2$ the ideal distance (350 pixels), and $\theta = \text{atan2}(d_y, d_x)$ the target angle. The canonical coordinates $\mathbf{p}_i^c = [x_i^c, y_i^c]^\top$ are computed via:

$$\begin{bmatrix} x_i^c \\ y_i^c \end{bmatrix} = \frac{1}{D} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x_i - x_1 \\ y_i - y_1 \end{bmatrix} \tag{4.1}$$

This applies translation (start to origin), rotation (target direction aligned with positive x-axis), and scaling (target placed at unit distance). After transformation, x_i^c represents forward progress toward the target (0 at start, 1 at target) and y_i^c represents lateral deviation from the ideal straight-line path. All features computed from these canonical coordinates are inherently invariant to target angle and distance.

4.3.2 Feature Engineering

From the canonical coordinates, we compute 10 trajectory features per trial organized along four dimensions.

Path Geometry (4 features): (1) *Path Efficiency*, $\text{PE} = 1/\sum_{i=1}^{N-1} \|\mathbf{p}_{i+1}^c - \mathbf{p}_i^c\|_2$, measuring path compactness where 1.0 indicates a perfectly straight trajectory; (2) *Maximum Deviation*, $\text{MaxDev} = \max_i |y_i^c|$, capturing peak lateral error; (3) *Closure*, the ratio of end-to-end displacement to total path length, where high values indicate direct movement; and (4) *DTW Distance*, the Dynamic Time Warping distance between the trajectory and a reference straight line from $(0, 0)$ to $(1, 0)$, normalized by N, capturing overall shape dissimilarity independent of speed.

Corrections and Strategy (2 features): (5) *Diagonal Direction Changes*, the rate of direction switches on the raw screen axes normalized by total steps, indicating frequency of corrective inputs; (6) *Lateral Oscillations*, the number of zero-crossings of y_i^c, representing side-to-side weaving.

Progress and Overshoot (2 features): (7) *Progress Linearity*, the R^2 of x_i^c regressed over time, where values near 1.0 indicate steady forward progress; (8) *Overshoot*, $\max(0, \max_i x_i^c - 1.0)$, measuring how far past the target the trajectory extends.

Kinematics (2 features): (9) *Average Speed*, the mean step-to-step velocity in canonical coordinates ($\Delta t = 20$ ms); (10) *Time to Peak Speed*, the normalized timestamp of maximum smoothed velocity, where early peaks indicate ballistic launches and late peaks indicate hesitant movement.

4.3.3 Block-by-Block Unsupervised Clustering

To capture how strategies evolve across the session while preserving transient learning phases, clustering is performed independently on each chronological block of 40 trials. This prevents later blocks from dominating the cluster structure and washing out early exploratory behavior.

All 10 features are z-score standardized (zero mean, unit variance) within each block prior to clustering. We apply Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [20] with `min_samples` = 25. HDBSCAN was chosen because it does not require specifying the number of clusters in advance, can identify clusters of varying density, and explicitly labels low-density points as noise ($C = -1$) rather than forcing them into inappropriate clusters.

4.3.4 Cluster Alignment Across Blocks

Because HDBSCAN is applied independently to each block, cluster labels are arbitrary across blocks. To enable longitudinal tracking, we align labels using bipartite centroid matching. For each pair of consecutive blocks $(t-1, t)$, we compute the centroid of each cluster and construct a cost matrix of Euclidean distances between new and reference centroids. The Hungarian linear sum assignment algorithm [21] solves the optimal one-to-one mapping:

$$\min_{\mathbf{X}} \sum_i \sum_j \mathrm{Cost}(i, j) \cdot X_{i,j} \tag{4.2}$$

where $\mathrm{Cost}(i, j)$ is the Euclidean distance between centroids and $X_{i,j} \in \{0, 1\}$ is the assignment variable. After alignment, equivalent strategies carry consistent labels across all four blocks.

4.3.5 Strategy Performance Validation

To confirm that emergent clusters represent genuinely different skill levels, we validate the cluster hierarchy against task execution time using one-sided Mann-Whitney U tests. Specifically, we test whether execution time monotonically decreases across the expertise levels: $T_{\text{Unskilled}} > T_{\text{Proficient}}$ and $T_{\text{Proficient}} > T_{\text{Masterful}}$.

4.3.6 Dyadic Strategy Joint Adoption

We quantify strategy joint adoption within each dyad on a trial-by-trial basis. A strategy at trial t is defined as jointly adopted if both partners belong to the same non-noise cluster:

$$\mathrm{JA}(t) = \mathbb{I}\big(C_{\text{user1}}(t) = C_{\text{user2}}(t) \;\wedge\; C_{\text{user1}}(t) \neq -1\big) \tag{4.3}$$

where $\mathbb{I}$ is the indicator function and $C_{\text{user}}(t)$ is the aligned cluster label. The overall joint adoption rate for a condition is the mean of $\mathrm{JA}(t)$ across all trials.

To determine whether observed joint adoption reflects genuine interpersonal coupling or is simply an artifact of both partners independently converging on the dominant strategy, we conduct a within-condition permutation test with $n = 100{,}000$ iterations. For each iteration, we shuffle partner assignments across dyads within the same condition, preserving each individual's cluster sequence but randomly re-pairing them with a different partner. This destroys dyadic coupling while preserving individual strategy trajectories and overall cluster distributions. The p-value is computed as the proportion of permuted values that equal or exceed the observed joint adoption rate.

4.4 Turn-Taking Analysis

This section describes the pipeline for extracting turn-taking features from raw audio recordings. It consists of three components: a speaker diarization system that produces speaker-labeled segments, turn-taking feature extraction, and prosodic feature extraction.

4.4.1 Speaker Diarization Pipeline

Speaker diarization is the task of determining "who spoke when" in an audio recording [17]. The standard pyannote.audio pipeline produces speaker-labeled segments, but the labels it assigns are arbitrary: they are determined by the internal order of the clustering algorithm and carry no information about speaker identity. There is no guarantee which label corresponds to which real participant, and the assignment can flip between different runs or parameter configurations of the same recording.

Computing turn-taking features such as speaking ratio or voice activity per speaker requires knowing which segments belong to the expert and which to the novice throughout the entire session. Furthermore, when two speakers have similar vocal characteristics, standard clustering can misassign segments entirely, attributing speech from one participant to the other without any mechanism for detection or correction.

To address these limitations, we developed a five-stage pipeline that augments the standard pyannote.audio architecture with two additional components: a local re-segmentation step that recovers speaker transitions lost during clustering, and a reference-guided assignment stage that resolves arbitrary cluster labels into consistent, named speaker identities. The pipeline takes as input the raw audio recording and two short reference voice samples (one per speaker, 12–22 seconds of clean solo speech), and produces as output a sequence of time-stamped segments, each labeled with a verified speaker identity and an associated confidence score quantifying the certainty of the assignment.

4.4.1.1 Stage 1: Neural Segmentation

A neural segmentation model processes the audio using a 10-second sliding window with a 1-second hop size. For each frame (≈ 17 ms), the model outputs independent probabilities for up to three local speakers being active. These probabilities are not mutually exclusive: multiple speakers can be active simultaneously (overlap), or none can be active (silence). The output is a 3D array of shape (num_windows $\times$ frames_per_window $\times$ 3).

4.4.1.2 Stage 2: Speaker Embedding Extraction

For each window, a 256-dimensional speaker embedding is extracted using the WeSpeaker ResNet34-LM model [19], trained on the VoxCeleb2 dataset with 5994 speakers. Embeddings are computed exclusively from frames where the target speaker is active and the other speakers are silent. This overlap-aware extraction produces more discriminative embeddings than standard fixed-window approaches, because the embedding reflects only one speaker's voice without contamination from the other.

4.4.1.3 Stage 3: Agglomerative Clustering

All extracted embeddings are grouped into speaker clusters using agglomerative hierarchical clustering with centroid linkage (UPGMC). Starting with every embedding as its own cluster, the algorithm iteratively merges the two clusters whose centroids are closest:

$$(C_i^*, C_j^*) = \arg\min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2, \quad \boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x} \tag{4.4}$$

where $\boldsymbol{\mu}_i$ is the centroid (mean embedding) of cluster C_i. Merging continues until exactly $K = 2$ clusters remain, reflecting the dyadic design. The remaining hyperparameters were selected via systematic evaluation across 10 configurations on a held-out 5-minute clip: clustering threshold $\delta = 0.25$, minimum cluster size $= 5$, and minimum intra-speaker gap $\Delta = 0.0$ s.

4.4.1.4 Stage 4: Local Re-Segmentation

A critical challenge arises from the scale-dependent behavior of agglomerative clustering. When processing a full 55-minute session, the cluster centroids are computed from approximately 2000 embeddings. A speaker transition that falls near the cluster boundary may land on the wrong side simply because the centroid shifted with the larger data volume. The segmentation model detected the transition correctly, but the clustering step silently absorbed it into a single merged segment.

To recover these lost transitions, we introduce a local re-segmentation step. For every pipeline output segment exceeding 2 seconds in length:

1. Extract the corresponding audio chunk from the full waveform
2. Run the segmentation model on that isolated chunk
3. Store frame-level probabilities across all overlapping windows, then average them
4. Binarize the averaged probabilities at threshold 0.5 to determine speaker activity per frame
5. Detect turn boundaries wherever the binary pattern changes between consecutive frames
6. Filter boundaries that are closer than 0.5 seconds to segment edges or to each other
7. Split the original segment at the detected boundaries

Running the segmentation model on an isolated chunk eliminates centroid drift interference, as the model operates solely on local acoustic evidence without being affected by clustering decisions from the rest of the file.

4.4.1.5 Stage 5: Reference-Guided Speaker Assignment

Each segment is assigned a speaker identity by comparing its embedding to pre-enrolled reference samples (12–22 seconds of clean solo speech per speaker) via cosine similarity. This applies to all segments regardless of whether they were re-segmented in Stage 4: short segments (≤ 2 s) that bypassed re-segmentation and sub-segments produced by re-segmentation are both processed identically in this stage.

Reference samples were created by manually identifying sections of each recording where only one speaker was clearly active (no overlapping speech, no background noise). These sections were extracted from the raw audio files using precise start and end timestamps, producing one clean solo speech clip per speaker per recording. Sample durations ranged from 12 to 22 seconds, chosen to be long enough for stable embedding computation while including varied intonation (questions, statements, exclamations) rather than monotone speech. The same embedding model (WeSpeaker ResNet34-LM with `window="whole"`) was then applied to each reference clip, producing one 256-dimensional reference embedding per speaker that serves as the enrollment template for all subsequent segment comparisons.

$$\hat{s} = \arg\max_{k \in \{1,2\}} \frac{\mathbf{e}_{\text{seg}} \cdot \mathbf{r}_k}{\|\mathbf{e}_{\text{seg}}\| \cdot \|\mathbf{r}_k\|} \tag{4.5}$$

where $\hat{s}$ is the assigned speaker identity, $\mathbf{e}_{\text{seg}}$ is the segment embedding, and $\mathbf{r}_k$ is the reference embedding for speaker k. The segment is assigned to whichever reference it is most similar to. A confidence score is computed as $|\text{sim}_1 - \text{sim}_2|$: large differences indicate clear assignment, while values near zero indicate ambiguous segments.

Finally, all segments are sorted chronologically and consecutive same-speaker segments separated by gaps shorter than 500 ms (or overlapping in time) are merged into single turns, with confidence scores combined as a duration-weighted average.

4.4.2 Turn-Taking Feature Extraction

From the diarization output, we extract turn-taking features at the dyad level per experimental block (B_1–B_4), excluding the three inter-block rest periods (3 minutes each). The break periods contain free conversation unrelated to the task and are analyzed separately where relevant. We define a *turn* as a maximal contiguous speech segment attributed to one speaker, bounded by silence $\geq$ 500 ms or a partner transition [6]. This 500 ms threshold follows established conventions in conversational analysis, where pauses shorter than this duration are considered within-turn hesitations rather than turn boundaries [5].

Turn-taking features were selected to capture both the quantity and structure of verbal coordination between partners. The turn-level features characterize how much each speaker talks, how long their contributions are, and how quickly they respond to each other. The sequential features capture higher-order patterns in how speakers alternate, revealing whether the conversation follows a balanced exchange or is dominated by one partner. Together, these features provide a comprehensive description of the verbal coordination regime adopted by each dyad at each stage of learning.

The following 12 features are computed per speaker per block:

Turn-Level Features (9 features): (1) *Voice Activity Detection (VAD %)*, the percentage of block duration occupied by speech from that speaker, capturing overall verbal engagement with the task and partner; (2) *Speaking Ratio*, the proportion of total dyad speech time produced by each speaker, where 0.5 indicates perfect balance; (3) *Number of Turns*, the total count of speaking turns in the block; (4) *Turns per Minute*, turn count normalized by block duration, enabling comparison across blocks of different lengths; (5) *Average Turn Duration*, the mean duration of all turns in seconds, where short turns suggest concise, targeted communication and long turns suggest extended explanation;

(6) *Short Turn Ratio*, the proportion of turns shorter than 1 second, capturing back-channel utterances and brief acknowledgments; (7) *Long Turn Ratio*, the proportion of turns longer than 5 seconds, capturing extended instructional or explanatory segments; (8) *Overlap Ratio*, the proportion of each speaker's speech time that overlaps with the partner's speech; (9) *Average Gap Before Turn*, the mean inter-speaker silence preceding each turn onset, measuring response latency.

Sequential Structure (3 features): Beyond individual turn characteristics, the pattern of speaker alternation carries information about conversational structure. We model the speaker sequence as a first-order Markov chain and compute the 2×2 transition matrix $\mathbf{T}$, where $T_{ij} = P(\text{speaker}_j \mid \text{speaker}_i)$. From this we derive: (10) *Alternation Ratio*, the proportion of transitions that are speaker switches (off-diagonal entries), where high values indicate rapid back-and-forth exchange; (11) *Turn-Taking Entropy*, the conditional entropy of the transition matrix:

$$H = -\frac{1}{2} \sum_i \sum_j T_{ij} \log_2 T_{ij} \tag{4.6}$$

where $H = 1.0$ indicates maximally unpredictable transitions and $H = 0.0$ indicates completely deterministic turn structure; (12) *Floor Holding*, the self-transition probability $P(S \to S)$, capturing the tendency to retain the conversational floor across consecutive turns.

We additionally tested whether filtering segments below various duration thresholds (0.5 s, 1.0 s, 1.5 s, 2.0 s) improves statistical separation between conditions. Results showed that unfiltered data consistently produced equal or better interaction p-values across the majority of features, confirming that short utterances carry meaningful interactional signal in this cooperative gaming context and should be retained rather than discarded as noise.

4.5 Between-Condition Statistical Comparison

The statistical framework is applied to both the trajectory-derived and speech-derived features.

4.5.1 One-Way ANOVA

As a baseline, we apply one-way ANOVA on features averaged across all four blocks. This tests whether conditions differ in overall levels without considering temporal dynamics. For the turn-taking features, we test both a 3-group comparison (Expert vs. Novice$_{NE}$ vs. Novice$_{NN}$) and a 2-group comparison (NE dyads vs. NN dyads).

4.5.2 Mixed-Design ANOVA

To capture whether features evolve differently across conditions over time, we employ a mixed-design ANOVA with Group (NE vs. NN) as the between-subjects factor and Block (B_1–B_4) as the within-subjects factor. The critical test is the Group $\times$ Block interaction, which captures divergent temporal trajectories between conditions. A significant interaction indicates that the two groups change differently over blocks, which is the statistical signature of scaffolding.

Greenhouse-Geisser correction is applied when Mauchly's sphericity test is violated ($\varepsilon < 0.75$), adjusting the degrees of freedom to account for unequal variances of differences between block pairs.

4.5.3 Post-Hoc Tests and Robustness Checks

When the mixed ANOVA interaction is significant, Bonferroni-corrected paired t-tests identify which specific block transitions drive the effect. Six pairwise comparisons are conducted per group (B_1–B_2, B_1–B_3, B_1–B_4, B_2–B_3, B_2–B_4, B_3–B_4), with raw p-values multiplied by 6 to control the family-wise error rate.

As a non-parametric robustness check against normality violations, we compute per-dyad linear slopes ($B_1 \rightarrow B_4$) for each feature using ordinary least squares regression. These slopes capture the overall direction and rate of change for each dyad. Slopes are then compared between NE and NN conditions using the Mann-Whitney U test, which requires no distributional assumptions.

Between-group differences at each individual block are additionally assessed using Mann-Whitney U tests to identify at which block the conditions diverge.

Implementation Details

All analyses were implemented in Python 3.11 on a remote Ubuntu 22.04 workstation equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). Trajectory clustering used scikit-learn 1.8 and SciPy 1.17. Speaker diarization used PyTorch 2.10 (CUDA 12.8), pyannote-audio 4.0.4, and torchaudio 2.10 for neural segmentation and embedding extraction.

Chapter 5

Results and Discussion

Summary

This chapter presents the results of the multimodal analysis. We report learning curves, unsupervised strategy profiles, longitudinal strategy evolution, dyadic strategy synchrony, turn-taking dynamics, and prosodic features. Each results section includes interpretation. A general discussion synthesizes findings across modalities and considers implications.

5.1 Learning Curves

Both conditions show significant learning concentrated in the first block transition: trial completion time drops sharply from Block 1 to Block 2 (Wilcoxon signed-rank, $p = 0.002$ for both NE and NN), with no significant improvement thereafter ($B_2 \rightarrow B_3$ and $B_3 \rightarrow B_4$, all $p > 0.13$). This indicates a power-law learning pattern where the task is largely acquired within the first 40 trials.

NE dyads consistently perform better than NN at every block, though the groups do not differ significantly at baseline (Block 1: NE = 30.6 s, NN = 38.4

s, $p = 0.212$). A significant between-group gap emerges at Block 2 (NE = 18.0 s vs. NN = 24.3 s, $p = 0.031$) and persists through Block 4 (NE = 16.1 s vs. NN = 20.8 s, $p = 0.076$). Overall learning rates are nearly identical (NE: -47.5%, NN: -45.9%), suggesting that expert presence does not accelerate the rate of learning. Rather, NE dyads consolidate to a lower performance asymptote, indicating that the expert enables faster convergence to efficient task execution after the initial learning phase.

Figure 5.1 shows the mean trial completion time per block for both conditions.

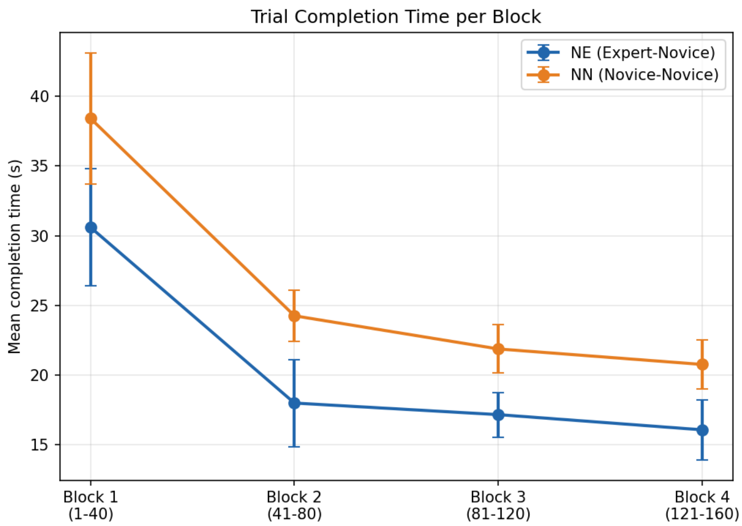

Figure 5.1: Mean trial completion time per block for NE and NN dyads. Error bars indicate standard error. Both groups improve sharply from Block 1 to Block 2, then plateau at different asymptotes.

Within-dyad variability (mean standard deviation of trial times per block) also decreases across blocks for both conditions (NE: 26.9 s in Block 1 to 11.1 s in Block 4; NN: 34.0 s to 14.5 s), indicating that both groups become more consistent over time. However, NN dyads remain more variable than NE at every block, suggesting less consolidated strategies.

Figure 5.2 shows the within-dyad variability per block for both conditions.

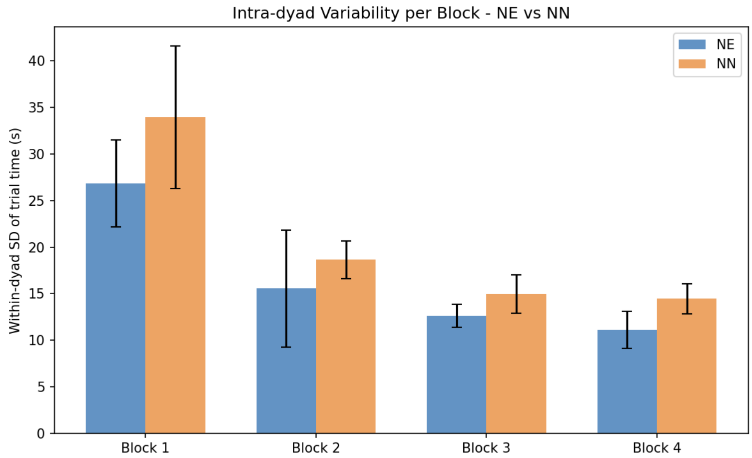

Figure 5.2: Within-dyad performance variability across blocks. Each bar represents the standard deviation of trial completion times computed within each dyad, then averaged across the 10 dyads per condition. Error bars indicate standard error of these per-dyad standard deviations across dyads.

5.2 Unsupervised Strategy Profiles

The HDBSCAN clustering pipeline reveals three distinct strategy clusters, consistent across all four blocks. We interpret these as progressive levels of task proficiency based on their feature profiles and validated execution times.

Cluster 0 (Unskilled) characterizes poor performance: low path efficiency (PE $\approx$ 0.73–0.78), large lateral deviations (MaxDev $\approx$ 0.28), frequent lateral oscillations ($\mu \approx 0.73$), erratic forward progress ($R^2 \approx 0.71$), and slow execution ($\mu \approx$ 18–24 s). This profile indicates feedback-dependent, exploratory behavior in which the player does not yet know which key combinations produce which directions and relies on trial-and-error correction.

Cluster 2 (Proficient) emerges from Block 2 onward as a transitional strategy: high but sub-perfect path efficiency (PE $\approx$ 0.96), fast execution ($\mu \approx$ 3.3 s), but with small residual direction changes and occasional lateral corrections. This suggests a learned but not yet fully automated control scheme: the player knows the correct direction but still requires minor online adjustments.

Cluster 1 (Masterful) represents a fully automated feedforward plan: near-perfect path efficiency (PE $\approx$ 1.03), minimal deviation (MaxDev $\approx$ 0.05), zero corrections, zero lateral oscillations, and rapid completion ($\mu \approx$ 3.1–3.5 s). The player launches directly toward the target with no corrective inputs, indicating complete internalization of the control mapping.

The expertise hierarchy across these clusters is statistically confirmed by task execution times. The one-sided Mann-Whitney U test confirms a monotonic decrease in execution time across levels: trials utilizing the Unskilled strategy are significantly slower than those using the Proficient strategy ($p < 0.001$), and Proficient trials are significantly slower than Masterful trials ($p = 0.022$).

5.3 Longitudinal Strategy Evolution

In Block 1, both conditions have the majority of their trials categorized within the Unskilled cluster (NE: 90.7%, NN: 97.1%), indicating that even pre-trained experts are initially disrupted by the novel cooperative coupling mechanism. The expert's prior knowledge of control mechanics does not immediately translate into efficient cooperative performance when the proximity constraint is introduced for the first time.

A notable shift occurs in Block 2: membership in the Unskilled cluster drops (NE: 51.5%, NN: 65.3%) as both Proficient and Masterful strategies emerge. NE dyads show stronger adoption of efficient strategies (Proficient + Masterful: 48.5% for NE vs. 34.7% for NN).

By Block 4, the divergence consolidates: NE dyads maintain 45.7% efficient strategies (14.1% Masterful + 31.6% Proficient), while NN dyads regress toward poor performance (72.4% Unskilled). This regression in NN dyads may reflect fatigue, loss of motivation, or the absence of a structured coordination partner to maintain efficient strategies under increasing fatigue.

This pattern suggests that expert scaffolding accelerates the transition through the intermediate Proficient stage, though it does not guarantee full automation (Masterful). The expert's presence appears to stabilize efficient strategies once acquired, preventing the regression observed in NN dyads.

Figure 5.3 displays the adoption frequency of each strategy across blocks in both conditions.

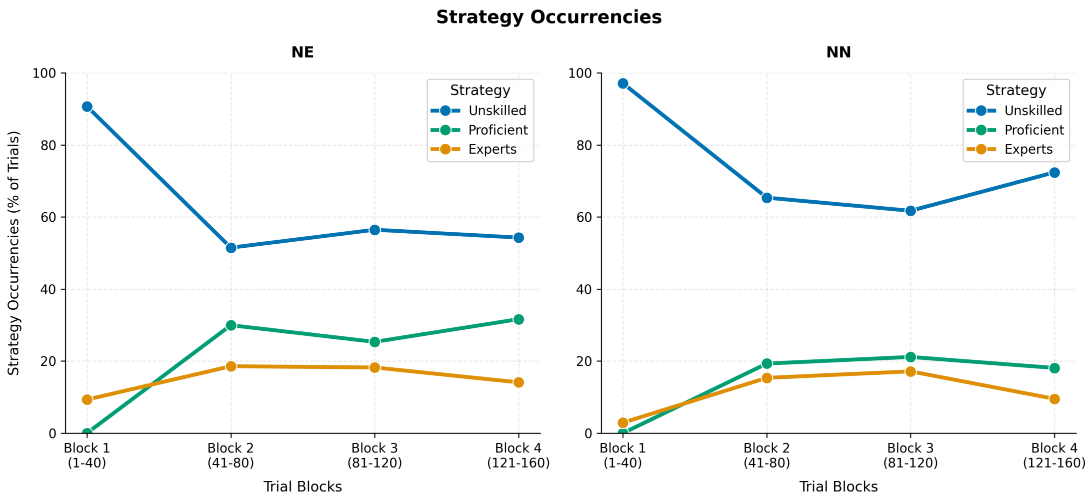

Figure 5.3: Adoption frequency of strategies in each experimental condition across trial blocks. NE dyads transition to efficient strategies faster and maintain them, while NN dyads regress toward Unskilled by Block 4.

5.4 Dyadic Strategy Synchrony

The within-condition permutation test (100,000 iterations) reveals a significant difference in strategy joint adoption between the two conditions.

NE dyads exhibit an observed trial-level joint adoption rate of 5.31%, which is highly significant above the shuffled chance baseline of $3.55\% \pm 0.40\%$ ($p < 0.001$).

This statistically confirms the presence of active, real-time strategy coordination between the novice and the pre-trained expert: they are not merely converging individually on the dominant strategy but are actively aligning their strategies trial by trial.

Conversely, NN dyads exhibit a joint adoption rate of 4.25%, which is statistically indistinguishable from the chance baseline of $3.62\% \pm 0.40\%$ ($p = 0.072$). Although both novices individually improve over blocks (as shown in the learning curves), their trial-by-trial strategy choices remain statistically independent. They converge to similar strategies over time but without genuine interpersonal coupling.

This finding demonstrates that expert presence enables a qualitatively different form of coordination: not just individual improvement toward the same endpoint, but active dyadic synchronization of strategies in real time.

Figure 5.4 shows the results of the within-condition permutation test.

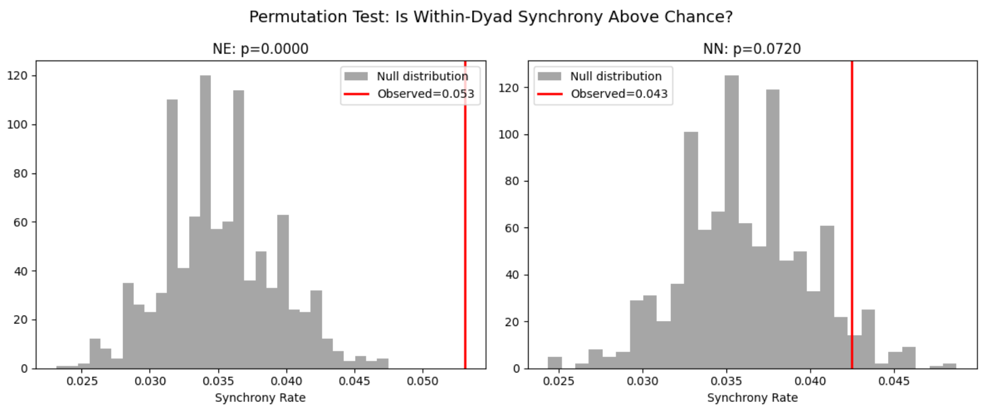

Figure 5.4: Permutation test results. Grey histograms show the null distribution of synchrony rates from 100,000 shuffled partner assignments. Red lines indicate the observed synchrony. Left: NE dyads synchronize well above chance ($p < 0.001$). Right: NN dyads do not significantly exceed chance ($p = 0.072$).

Figure 5.5 shows the overall dyadic strategy synchrony across blocks for both conditions.

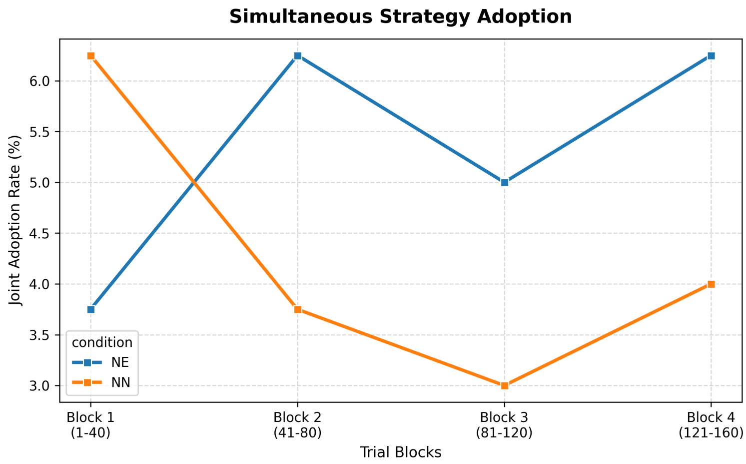

Figure 5.5: Overall dyadic strategy synchrony over blocks. NE dyads increase synchrony from Block 1 onward, while NN dyads decline from their initial peak. The lines cross at Block 2, after which NE consistently exceeds NN.

Decomposing synchrony by cluster (Figure 5.6) reveals the nature of this divergence.

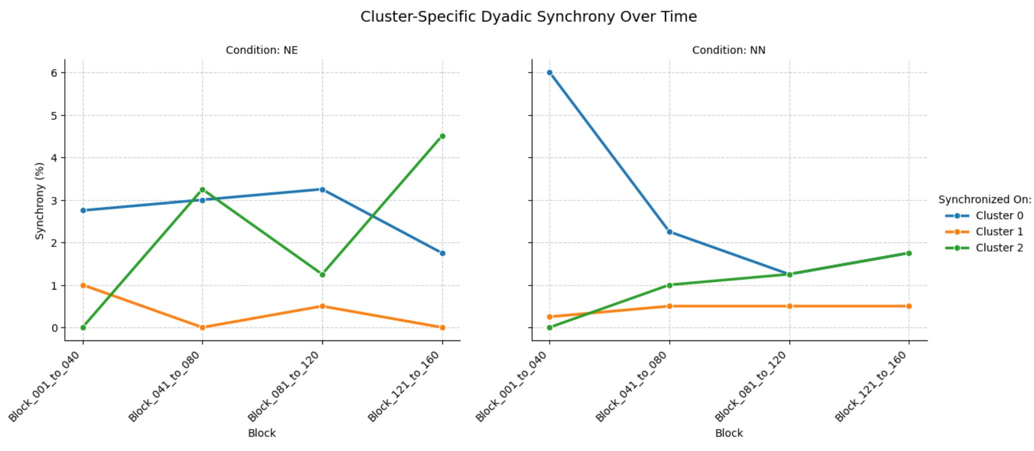

Figure 5.6: Cluster-specific dyadic synchrony over blocks. Left: NE dyads shift from synchronizing on Cluster 0 (Unskilled) to Cluster 2 (Proficient) by Block 4. Right: NN dyads lose their initial trivial Cluster 0 alignment without developing synchrony on efficient strategies.

5.5 Turn-Taking Dynamics Across Conditions

One-way ANOVA on turn-taking features aggregated across all four blocks yields no significant group differences for any of the 12 features tested (all $p > 0.22$). This motivates the temporal analysis below, as overall averages may obscure divergent patterns that develop over blocks.

The mixed ANOVA (Group × Block) reveals two significant interactions.

VAD % diverges across blocks ($p = 0.048$): NE dyads progressively reduce voice activity from Block 1 (39.8%) to Block 4 (23.0%), a monotonic decline confirmed by post-hoc tests ($B_1 \to B_3$: $p = 0.030$). This pattern reflects increasing communication efficiency as the dyad develops shared task understanding and requires less verbal coordination. NN dyads show no such monotonic trend (B_1: 26.1%, B_3: 27.3%, B_4: 18.8%), with speech density fluctuating erratically across blocks.

The Speaking Ratio trend reveals who drives the VAD reduction ($p = 0.147$): in NE dyads, the expert's share of total speech drops from 0.52 in Block 1 to 0.37 in Block 4, while the novice's share rises correspondingly. This is consistent with a scaffolding dynamic in which the expert initially provides intensive verbal guidance and progressively withdraws as the novice internalizes the task. NN dyads show no directional shift in speaking balance.

Notably, NE dyads start with higher VAD% at Block 1 (NE: 39.8% vs. NN: 26.1%, Mann-Whitney $p = 0.089$), suggesting that the expert invests heavily in verbal guidance from the outset before progressively withdrawing.

Average Turn Duration exhibits the strongest interaction ($p = 0.027$) with opposing trajectories: NE turns shorten across blocks (from 1.67 s in B_1 to 1.54 s in B_3), while NN turns lengthen (from 1.66 s in B_1 to 2.02 s in B_3). The between-group divergence peaks at Block 3, where NN turns are significantly longer than NE turns (NE: 1.54 s vs. NN: 2.02 s, Mann-Whitney $p = 0.045$). Longer turns in NN dyads may reflect less efficient verbal coordination, where speakers require more words to convey the same coordination intent that NE dyads achieve with shorter, more targeted utterances.

Figure 5.7 shows the temporal evolution of VAD% and average turn duration across blocks.

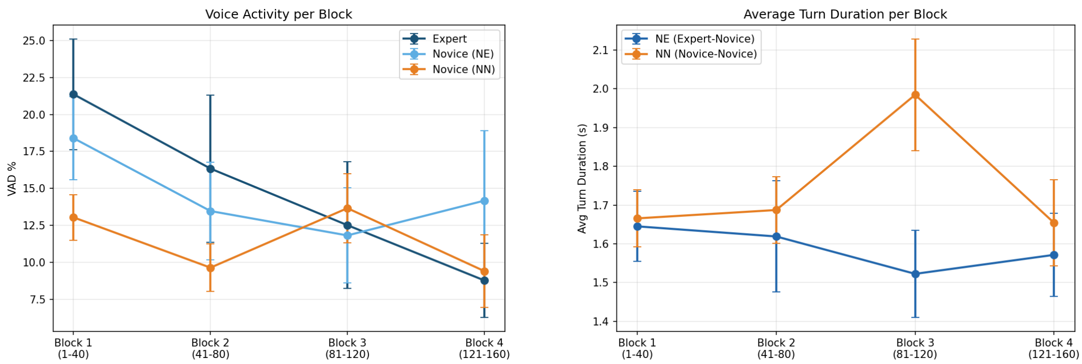

Figure 5.7: Left: voice activity per block for Expert, Novice (NE), and Novice (NN). Right: average turn duration per block for NE and NN dyads. Error bars indicate standard error.

5.6 Prosodic Features Across Conditions

The prosodic analysis using the 88 eGeMAPSv02 features revealed no significant differences between conditions for any feature, either in static comparisons (one-way ANOVA, all $p > 0.6$) or in temporal interactions (mixed ANOVA Group $\times$ Block, all $p > 0.3$). Effect sizes were uniformly small ($\eta^2 < 0.03$).

This result is expected and informative: prosodic features (pitch, formants, loudness, voice quality) primarily reflect individual vocal anatomy and habitual speaking style rather than expertise level or cooperative dynamics. An expert does not speak with a fundamentally different voice than a novice. The distinguishing signal between conditions lies in interaction patterns (how speakers coordinate) rather than in individual voice characteristics (how speakers sound). This negative finding confirms that the turn-taking effects reported above reflect genuine changes in communication structure rather than confounds from voice quality differences between groups.

5.7 Back-Channel Filtering Results

Systematic testing of five duration thresholds (no filter, ≥ 0.5 s, ≥ 1.0 s, ≥ 1.5 s, ≥ 2.0 s) revealed that unfiltered data produces the best or equal-best interaction p-values for the majority of features. The two significant interactions (VAD% and Average Turn Duration) remain significant across four of five thresholds, confirming their robustness to methodological choices regarding back-channel inclusion.

The stability of results across thresholds indicates that back-channel utterances in this cooperative gaming context carry meaningful interactional information and should not be excluded. Unlike formal meeting recordings where back-channels may add noise to dominance measures, the short utterances in our data reflect active verbal engagement with the task and the partner.

5.8 General Discussion

The trajectory and turn-taking analyses converge on the same conclusion: expert presence restructures the temporal dynamics of dyadic coordination across modalities. At the trajectory level, NE dyads transition to efficient strategies faster and achieve trial-level joint adoption significantly above chance, while NN dyads never exceed chance-level coordination. In speech, NE dyads progressively reduce verbal output with the expert yielding conversational control, while NN dyads fluctuate without directional structure. Notably, preliminary static comparisons on aggregated turn-taking features yield no significant group differences.

The scaffolding effect is exclusively temporal, emerging only through Group × Block interaction analysis.

The progressive transfer of verbal control we observe (expert speaking ratio dropping from 0.52 to 0.37 while the novice's share rises correspondingly) aligns with the theoretical framework of scaffolding as originally described by Wood, Bruner and Ross [10]: the more competent partner provides structured support calibrated to the learner's current ability, then fades that support as competence develops. What our data add to this classical model is that the same fading pattern operates simultaneously in task execution (strategy joint adoption) and in verbal interaction (turn-taking dynamics), suggesting that scaffolding is not a single-channel process but a multimodal one in which action and speech are coupled components of the same coordination regime.

Our longitudinal design also extends recent findings by Haraped et al. [7], who reported that expertise did not create overall coordination asymmetry within a single 5-minute interaction, with asymmetries emerging only when data were disaggregated by conversational turns. Our results are consistent with theirs at the static level, but reveal that expertise effects emerge when coordination is tracked over repeated practice blocks rather than within a single session. This suggests that scaffolding requires time to manifest: it is a developmental process that unfolds over tens of minutes and multiple task repetitions, not a property visible in a brief snapshot of interaction.

The regression of NN dyads toward Unskilled strategies in Block 4 (72.4%), despite having achieved partial proficiency in earlier blocks, warrants specific interpretation. Without expert stabilization, efficient strategies appear fragile: novice-novice dyads can discover effective coordination patterns through exploration but cannot sustain them under the accumulating demands of a long session. The expert may therefore serve not only as an initial scaffold that accelerates strategy discovery, but as an ongoing stabilizer that prevents regression. This dual role (accelerator + stabilizer) may explain why NE dyads maintain 45.7% efficient strategies in Block 4 while NN dyads collapse back toward exploratory behavior.

From a systems perspective, the pipeline developed in this thesis constitutes the analytical backbone that a real-time adaptive collaborative system would require: automated speaker identity resolution from raw audio, temporal tracking of verbal coordination patterns, and unsupervised detection of task strategy states without prior labeling. These components operate on the same multimodal streams (audio, trajectory data) that would be available to an intelligent collaborative interface during live interaction, making the transition from offline analysis to online monitoring a matter of engineering optimization rather than fundamental redesign. The multimodal coordination signatures we identify, specifically declining speech density, shortening turn duration, and above-chance strategy joint adoption, could serve as real-time indicators of scaffolding progress. A system detecting the absence of these patterns (as observed in NN dyads) could intervene with adaptive support, mimicking the expert's natural withdrawal of guidance as coordination stabilizes.

Chapter 6

Conclusions

6.1 Summary of Findings

This thesis presented DyadicTickTacking, a multimodal analysis framework combining unsupervised trajectory clustering with automated speaker diarization and turn-taking analysis to investigate how expertise asymmetry shapes cooperative coordination over time. We compared 10 Novice-Expert and 10 Novice-Novice dyads performing a cooperative task across 160 trials organized in four blocks, analyzing both cursor trajectory strategies and verbal interaction patterns.

Our findings address the three research questions posed in the Introduction.

Regarding the first question, whether coupling with an expert improves task performance, the learning curve analysis revealed that both conditions learn at equivalent rates (NE: -47.5%, NN: -45.9%), with learning concentrated in the first block transition ($p = 0.002$ for both). However, NE dyads consolidate to a significantly lower performance asymptote, with the between-group gap emerging at Block 2 ($p = 0.031$) and persisting through Block 4 ($p = 0.076$). Expert presence does not accelerate the rate of learning but enables faster convergence to efficient task execution.

Regarding the second question, whether participants adopt distinct strategies and how expert presence shapes their evolution, HDBSCAN clustering revealed three emergent strategy profiles (Unskilled, Proficient, Masterful) validated by monotonically decreasing execution times ($p < 0.001$ and $p = 0.022$). NE dyads transition to efficient strategies faster (48.5% efficient by Block 2 vs. 34.7% for NN) and maintain them, while NN dyads regress toward Unskilled performance by Block 4 (72.4%). The permutation test (100,000 iterations) confirmed that NE dyads achieve trial-by-trial strategy joint adoption significantly above chance ($p < 0.001$), indicating genuine real-time interpersonal coordination. NN dyads do not exceed chance-level synchrony ($p = 0.072$), meaning they converge individually without active dyadic coupling.

Regarding the third question, whether turn-taking behaviors co-evolve with task learning differently across conditions, mixed-design ANOVA revealed two significant Group × Block interactions. VAD% diverges across blocks ($p = 0.048$): NE dyads monotonically reduce speech density from 39.8% to 23.0%, with the expert's share of total speech dropping from 0.52 to 0.37 while the novice's share rises correspondingly. Average Turn Duration shows opposing trajectories ($p = 0.027$): NE turns shorten while NN turns lengthen, diverging significantly at Block 3 ($p = 0.045$). NN dyads show no directional verbal structure across blocks. Prosodic features revealed no significant differences between conditions, confirming that the turn-taking effects reflect genuine changes in communication structure rather than voice quality confounds.

A key methodological finding is that the scaffolding effect is exclusively temporal: one-way ANOVA on aggregated features yields no significant group differences, confirming that static comparisons are insufficient. The effect emerges only when coordination dynamics are tracked longitudinally through interaction analysis. This convergent multimodal evidence demonstrates that expert presence restructures how coordination develops over time rather than simply producing better individual performance.

6.2 Limitations

Several limitations should be acknowledged.

The sample size of 10 dyads per condition limits statistical power. With larger samples, the trending effects observed for several features (Turns per Minute, Speaking Ratio, Floor Holding) might reach significance, and the borderline results (e.g., $p = 0.048$ for VAD%) would carry more confidence.

The absence of ground-truth diarization annotations means that speaker assignment accuracy relies on convergence across 10 parameter configurations and manual verification of a held-out 5-minute clip rather than formal Diarization Error Rate evaluation against human annotations.

The majority of dyads (86%) consisted of previously acquainted individuals. While this reflects naturalistic conditions, familiar pairs may coordinate more easily regardless of expertise condition, potentially attenuating the difference between NE and NN groups.

A subset of dyads communicated in a second language (3 NE, 2 NN), which may have constrained verbal output relative to native-language pairs independently of the expertise manipulation.

Sessions exceeded 50 minutes for most dyads. Fatigue in later blocks may partially account for reduced speech density in both conditions. However, the monotonic decline in NE versus the erratic pattern in NN suggests that fatigue alone cannot explain the divergent temporal dynamics observed.

6.3 Future Directions

Several directions for future work emerge from this thesis.

The video recordings collected during the experiment contain gaze direction, facial expressions, and co-speech gesture data that were not analyzed in the present work. Integrating these additional modalities could provide a richer characterization of the scaffolding process, revealing whether the expert leads visual attention in the same progressive-withdrawal pattern observed in speech.

The original experimental design included competitive conditions in which the proximity coupling penalizes closeness rather than distance. Analyzing these conditions would reveal whether the coordination signatures identified here are specific to cooperation or reflect more general effects of expertise asymmetry on interaction structure.

The three emergent strategy clusters and the permutation-based synchrony metric developed here could be applied to other cooperative paradigms beyond DyadicTickTacking, testing whether similar expertise-dependent coordination patterns emerge in different task domains with different control interfaces.

From an engineering perspective, the multimodal coordination signatures we identified (declining speech density, shortening turn duration, above-chance strategy joint adoption) could serve as real-time indicators for adaptive collaborative systems. An intelligent interface monitoring these signals during a cooperative task could detect when scaffolding is progressing naturally and when intervention is needed, mimicking the expert's behavior of providing structured guidance and then progressively withdrawing. The DyadicTickTacking paradigm itself offers a controlled testbed for prototyping and evaluating such adaptive systems before deployment in more complex collaborative settings.

Finally, replication with larger and more controlled samples, explicitly balancing familiarity, language proficiency, and gender composition across conditions, would strengthen the generalizability of the scaffolding patterns observed here and enable more sophisticated statistical modeling such as linear mixed-effects models that can account for individual differences within conditions.

References

[1] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 70–76, 2006. 1, 6, 7

[2] L. M. Fawcett and A. F. Garton, "The effect of peer collaboration on children's problem-solving ability," *British Journal of Educational Psychology*, vol. 75, no. 2, pp. 157–169, 2005. 1, 4, 14

[3] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978. 2, 7, 14

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009. 3, 4, 13

[5] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010. 3, 8, 36

[6] C. Raman, N. R. Prabhu, and H. Hung, "Perceived conversation quality in spontaneous interactions," *IEEE Transactions on Affective Computing*, 2022. 3, 9, 13, 36

[7] L. Haraped, S. E. Huber, W. F. Bischof, and A. Kingstone, "Looking, pointing, and talking together: How dyads of differential expertise coordinate

attention during conversation," *PLoS ONE*, vol. 19, no. 12, p. e0315728, 2024. 3, 4, 7, 8, 13, 50

[8] R. Vink, M. L. Wijnants, A. H. N. Cillessen, and A. M. T. Bosman, "Co-operative learning and interpersonal synchrony," *Nonlinear Dynamics, Psychology, and Life Sciences*, vol. 21, no. 2, pp. 189–215, 2017. 4, 13

[9] J. Ayache, A. Connor, S. Marks, D. J. Kuss, D. Rhodes, A. Sumich, and N. Heym, "Exploring the "dark matter" of social interaction: Systematic review of a decade of research in spontaneous interpersonal coordination," *Frontiers in Psychology*, vol. 12, p. 718237, 2021. 4, 12, 13

[10] D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem solving," *Journal of Child Psychology and Psychiatry*, vol. 17, no. 2, pp. 89–100, 1976. 4, 7, 14, 50

[11] P. E. Keller, G. Novembre, and M. J. Hove, "Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1658, p. 20130394, 2014. 7

[12] L. Noy, E. Dekel, and U. Alon, "The mirror game as a paradigm for studying the dynamics of two people improvising motion together," *Proceedings of the National Academy of Sciences*, vol. 108, no. 52, pp. 20947–20952, 2011. 7

[13] J. van de Pol, M. Volman, and J. Beishuizen, "Scaffolding in teacher–student interaction: A decade of research," *Educational Psychology Review*, vol. 22, no. 3, pp. 271–296, 2010. 7, 14

[14] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974. 8

[15] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009. 9

[16] P. Tzirakis, A. Baird, J. Brooks, E. Parada-Cabaleiro, L. Stappen, S. Rao, T. Lebryk, J. P. Cłapa, and J. Madsen, "The 2026 acii dyadic conversations (daikon) workshop & challenge," *arXiv preprint arXiv:2605.02672*, 2026. 9

[17] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012. 10, 31

[18] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *24th INTERSPEECH Conference*, pp. 1983–1987, 2023. 10

[19] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. 10, 33

[20] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 160–172, Springer, 2013. 11, 29

[21] D. F. Crouse, "On implementing 2d rectangular assignment algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1679–1696, 2016. 11, 30

[22] R. Rein, C. Button, K. Davids, and J. Summers, "Cluster analysis of movement patterns in multiarticular actions: A tutorial," *Motor Control*, vol. 14, no. 2, pp. 211–239, 2010. 12

[23] S. Mulroy, J. A. Gronley, W. Weiss, C. Newsam, and J. Perry, "Use of cluster analysis for gait pattern classification of patients in the early and late recovery phases following stroke," *Gait & Posture*, vol. 18, no. 1, pp. 114–125, 2003. 12

[24] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," *ACM Computing Surveys*, vol. 50, no. 2, pp. 1–34, 2017. 12

[25] K. Onishi, H. Tanaka, and S. Nakamura, "Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation," in *Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23)*, pp. 386–388, ACM, 2023. 13

[26] C. Vesper, R. P. R. D. van der Wel, G. Knoblich, and N. Sebanz, "Making oneself predictable: reduced temporal variability facilitates joint action coordination," *Experimental Brain Research*, vol. 211, no. 3, pp. 517–530, 2011. 13

[27] I. S. Plank, J. C. Koehler, A. M. Nelson, N. Koutsouleris, and C. M. Falter-Wagner, "Automated extraction of speech and turn-taking parameters in autism allows for diagnostic classification using a multivariable prediction model," *Frontiers in Psychiatry*, vol. 14, p. 1257569, 2023. 14

[28] D. Rocchesso, A. Bellino, and A. Perez, "Ticktacking – drawing trajectories with two buttons and rhythm," in *Proceedings of the Sound and Music Computing Conferences*, vol. 2023-June, pp. 63–71, 2023. 16

[29] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *24th INTERSPEECH Conference*, pp. 1983–1987, 2023.