Multi-Signal Tempo-Normalized Conversation Segmentation

Accounting for Actor Velocity, Message Length, and Discourse Coherence

Abstract

This document presents an enhanced framework for segmenting dialogue sessions that combines tempo-normalized temporal analysis with multi-signal boundary detection. Building on research in computational linguistics and discourse analysis, we extend pure temporal segmentation to incorporate lexical coherence, turn-taking patterns, and discourse markers. The framework accounts for actor-specific response velocities, message composition time, and cognitive processing demands, making it robust to heterogeneous actor characteristics in human-LLM and human-human conversations.

1 Introduction

Session boundary detection in dialogue systems is a fundamental challenge in computational linguistics. This document extends temporal segmentation frameworks to account for critical real-world phenomena:

1. **Actor-specific response velocities:** Different actors respond at systematically different speeds (e.g., LLMs respond near-instantaneously, while humans may take minutes).
2. **Message length effects:** Longer messages naturally require more time to compose, so raw temporal intervals should be normalized by message length to better capture conversation tempo.
3. **Multi-signal integration:** Research consistently shows that temporal features work best when combined with lexical coherence, topic continuity, and discourse markers (Kummamuru et al., 2009; Netten & van Someren, 2011).

The goal is to detect session boundaries based on changes in conversational tempo and discourse coherence, rather than absolute time alone, making the segmentation robust to heterogeneous actor characteristics.

2 Literature Foundation

2.1 Temporal Features in Dialogue Segmentation

Research in discourse analysis has established that temporal discontinuities serve as one component in multi-cue boundary detection systems. Netten & van Someren (2011) developed a Task-Adaptive Information Distribution system that segments dialogues using multiple cues including elapsed time between utterances, speaker switches, and lexical/syntactical markers. Their experiments demonstrated 90% accuracy in recognizing unit boundaries.

Kummamuru et al. (2009) proposed an unsupervised algorithm for segmenting conversational transcripts that combines lexical coherence, temporal proximity, and position information. This multi-feature approach proved more robust to noise than pure temporal or lexical methods alone.

2.2 Turn-Taking and Inter-Utterance Gaps

Research on turn-taking behavior provides empirical grounding for temporal thresholds. van Os et al. (2020) established that speakers follow a universal tendency to minimize pauses between turns while avoiding overlaps. They distinguish between "turn gaps" (positive temporal relations), "overlaps" (negative temporal relations), and characterize how deviations from expected timing affect perceived fluency.

Soma et al. (2022) established that humans can discern a pause in conversation at between 0.15 and 0.25 seconds, with the average being approximately 0.2 seconds. This provides a lower bound for meaningful silence detection.

2.3 Topic Continuity and Lexical Coherence

Rohde & Frank (2014) found that gaps between utterance onsets are shorter within topical sequences than at sequence boundaries. In their analysis of child-caregiver interactions, within-sequence gaps averaged 2.4–3.3 seconds versus 4.9–5.9 seconds at boundaries, supporting the integration of temporal and topical signals for boundary detection.

Morris-Adams (2016) discusses "marked topic changes" where there is no lexical or propositional link to preceding content. Such changes provide strong boundary signals independent of temporal gaps.

3 Extended Message Model

3.1 Message with Length and Complexity

Definition 1 (Extended Message). A message is a 5-tuple:

$$m = (t, a, c, \ell, \kappa)$$

where:

- $t \in \mathbb{N}$ is the timestamp
- $a \in A$ is the actor
- $c \in \Sigma^*$ is the message content
- $\ell \in \mathbb{R}^+$ is the message length (word count, character count, or token count)
- $\kappa \in \mathbb{R}^+$ is the message complexity score

3.2 Complexity Estimation

Following insights from cognitive load research, message complexity extends beyond length:

$$\kappa(c) = 1 + \gamma_1 \cdot \text{lexical\_diversity}(c) + \gamma_2 \cdot \text{syntactic\_depth}(c) + \gamma_3 \cdot \text{semantic\_novelty}(c)$$

where lexical diversity measures unique words / total words, syntactic depth measures average parse tree depth, and semantic novelty measures embedding distance from previous messages. Default: $\kappa = 1$ when complexity scoring is unavailable.

4 Actor-Specific Response Velocity

4.1 Motivation

In human-LLM conversations, response asymmetry is substantial:

- **Human → LLM:** Near-zero latency (LLM responds immediately)
- **LLM → Human:** Variable latency (human reads, thinks, types, may get distracted)

van Os et al. (2020) found that both "too eager" and "too reluctant" responses affect conversational dynamics differently, suggesting these transitions have qualitatively different properties.

4.2 Dual-Component Response Model

Definition 2 (Expected Response Time with Reading Component). For a human responding to an LLM message, the expected response time includes both processing and composition:

$$\mathbb{E}[\Delta t_i] = \beta_a^{\text{read}} \cdot \ell_{i-1} + \beta_a^{\text{write}} \cdot \ell_i \cdot \kappa_i$$

where β^{read} captures reading/processing time of the previous message (≈ 0.25 sec/word for humans), and β^{write} captures composition time (≈ 0.5 sec/word for humans). For LLM actors, both components approach zero ($\beta_{\text{LLM}}^{\text{read}} \approx 0$, $\beta_{\text{LLM}}^{\text{write}} \approx 0.05$ sec/word).

Novel Contribution

This dual-component model addresses a gap in the literature where processing time for received content is typically conflated with composition time.

4.3 Non-Linear Composition Scaling

Research on turn-taking (Bunning et al., 2011) shows that very short messages (backchannels like "ok", "yes") have different temporal properties than substantive turns. We model this with piecewise scaling:

$$\mathbb{E}[\text{comp}(m_i)] = \begin{cases} \beta_a^{\min} & \text{if } \ell_i < \ell_{\text{threshold}} \\ \beta_a \cdot \ell_i^{\gamma} \cdot \kappa_i & \text{otherwise} \end{cases}$$

where $\gamma < 1$ (typically 0.8–0.9) captures sublinear scaling due to typing momentum effects, and $\beta_a^{\min}$ represents minimum response time even for trivial messages (≈ 1.5 sec for humans).

5 Tempo-Normalized Inter-Exchange Intervals

5.1 Raw vs. Normalized IEI

Definition 3 (Raw Inter-Exchange Interval). The raw IEI between exchanges E_i and E_{i+1} is:

$$\Delta t_i^{\text{raw}} = t_{\text{start}}(E_{i+1}) - t_{\text{end}}(E_i)$$

Definition 4 (Normalized Inter-Exchange Interval). The tempo-normalized IEI accounts for expected reading and composition time:

$$\Delta t_i^{\text{norm}} = \Delta t_i^{\text{raw}} - \mathbb{E}[\text{read} + \text{comp}] = \Delta t_i^{\text{raw}} - \left(\beta_a^{\text{read}} \cdot \ell_{\text{prev}} + \beta_a^{\text{write}} \cdot \ell_{\text{next}}^{\gamma} \cdot \kappa_{\text{next}}\right)$$

5.2 Tempo Ratio

Definition 5 (Tempo Ratio).

$$\rho_i = \frac{\Delta t_i^{\text{raw}}}{\mathbb{E}[\text{read} + \text{comp}] + \epsilon}$$

Interpretation:

- $\rho_i \approx 1$: Natural conversational flow
- $\rho_i \gg 1$: Potential session boundary
- $\rho_i < 1$: Faster-than-expected response (eager response or pre-composed message)

6 Multi-Signal Boundary Detection

Following the literature consensus that temporal features work best in combination with other signals (Kummamuru et al., 2009; Netten & van Someren, 2011), we propose a multi-signal detection framework.

6.1 Lexical Coherence Signal

Definition 6 (Lexical Coherence Score). The coherence between adjacent exchanges:

$$\text{coh}(E_i, E_{i+1}) = \text{cosine\_similarity}(\text{embed}(E_i), \text{embed}(E_{i+1}))$$

A lexical break is detected when $\text{coh}(E_i, E_{i+1}) < \theta_{\text{sim}}$ (typically 0.3–0.5). This captures topic changes with no lexical or propositional link to preceding content (Morris-Adams, 2016).

6.2 Discourse Marker Signal

Explicit session markers provide strong boundary signals independent of tempo:

```
marker_patterns = {
    greetings: ['hi', 'hello', 'hey', 'hi again'],
    new_topic: ['new question', 'different topic', 'unrelated but'],
    closings: ['thanks', 'bye', 'that's all']
}
```

6.3 Rhythm Variance Signal

Following turn-taking research (Sacks, Schegloff, & Jefferson, 1978, as cited in van Os et al., 2020), sudden changes in conversational rhythm can indicate session boundaries:

$$\text{rhythm\_var}_i = \text{std}(\{\rho_{i-w}, \ldots, \rho_{i-1}, \rho_i\})$$

High rhythm variance indicates the conversation has entered a different "mode."

6.4 Combined Boundary Function

Definition 7 (Multi-Signal Boundary Function).

$$\begin{aligned} B_{\text{multi}}(E_i, E_{i+1}, H) = {} & \text{marker\_break}(E_{i+1}) && \text{[Strong signal]} \\ \vee {} & (\text{tempo\_break}(\Delta t_i^{\text{norm}}) \wedge \text{lexical\_break}(\text{coh}_i)) && \text{[Validated tempo]} \\ \vee {} & (\text{tempo\_break}(\Delta t_i^{\text{norm}}) \wedge \Delta t_i^{\text{norm}} > \tau_{\text{absolute}}) && \text{[Absolute threshold]} \\ \vee {} & (\text{lexical\_break}(\text{coh}_i) \wedge \text{rhythm\_break}(\text{rhythm\_var}_i)) && \text{[Topic + rhythm]} \end{aligned}$$

Novel Contribution

This multi-signal integration prevents false boundaries when users pause mid-conversation but continue the same topic, while catching topic shifts with modest temporal gaps.

7 Adaptive Threshold Computation

7.1 IQR-Based Adaptive Thresholds

Using normalized IEIs in history $H = \{\Delta t_0^{\text{norm}}, \ldots, \Delta t_{i-1}^{\text{norm}}\}$:

$$\tau_{\text{adaptive}} = Q_3(H) + k \cdot \text{IQR}(H) \quad (1)$$

$$\tau = \max(\tau_{\text{min}}, \tau_{\text{adaptive}}) \quad (2)$$

Recommended: $\tau_{\text{min}} = 60$ seconds, $k = 1.5$, window size = 20 exchanges.

7.2 Context-Dependent Baselines

Rohde & Frank (2014) found distinct temporal patterns within vs. between topical sequences. We maintain separate baselines:

β_a^{within} — baseline for within-session responses (tighter)

β_a^{initial} — baseline for session-initiating responses (looser)

8 Enhanced Detection Algorithm

Algorithm 1 Multi-Signal Tempo-Normalized Boundary Detection

```
Require: Exchange $E_{i+1}$, previous exchange $E_i$, history $H$, semantic model $M$
Require: Parameters: $\tau_{\min}$, $k$, $\theta_{\text{sim}}$, $\theta_{\text{rhythm}}$, $\lambda$
Compute $\Delta t_i^{\text{raw}} = t_{\text{start}}(E_{i+1}) - t_{\text{end}}(E_i)$
Compute expected time with dual components:
    $\mathbb{E}[\text{response}] = \beta_a^{\text{read}} \cdot \ell_{\text{prev}} + \beta_a^{\text{write}} \cdot \max(\ell_{\text{next}}, \ell_{\min})^{\gamma} \cdot \kappa_{\text{next}}$
Compute $\Delta t_i^{\text{norm}} = \max(0, \Delta t_i^{\text{raw}} - \mathbb{E}[\text{response}])$
Compute $\rho_i = \Delta t_i^{\text{raw}} / (\mathbb{E}[\text{response}] + \epsilon)$
Lexical signal:
    $\text{coh} = \text{semantic\_similarity}(E_i, E_{i+1})$
    $\text{lexical\_break} = (\text{coh} < \theta_{\text{sim}})$
Marker signal:
    $\text{marker\_break} = \text{contains\_session\_marker}(\text{first\_message}(E_{i+1}))$
Rhythm signal:
    $\text{rhythm\_var} = \text{std}(\{\rho_{i-w}, \ldots, \rho_i\})$
    $\text{rhythm\_break} = (\text{rhythm\_var} > \theta_{\text{rhythm}})$
Tempo signals:
    Add $\Delta t_i^{\text{norm}}$ to history $H$ (keep last 20)
    $\tau = \max(\tau_{\min}, Q_3(H) + k \cdot \text{IQR}(H))$
    $\text{tempo\_break} = (\Delta t_i^{\text{norm}} \geq \tau) \lor (\rho_i \geq \rho_{\min})$
Combined decision:
if marker_break then
    return BOUNDARY {Strong explicit signal}
end if
if tempo_break $\land$ (lexical_break $\lor$ $\Delta t_i^{\text{norm}} > \tau_{\text{absolute}}$) then
    return BOUNDARY {Validated tempo break}
end if
if lexical_break $\land$ rhythm_break then
    return BOUNDARY {Topic shift with rhythm change}
end if
return NO BOUNDARY
Update actor baselines using EMA: $\beta_a \leftarrow \lambda \cdot \text{observed\_rate} + (1 - \lambda) \cdot \beta_a$
```

9 Hierarchical Session Structure

Following discourse analysis research (Artigas Miralles et al., 2019), conversations often have hierarchical structure. Sessions may contain multiple topic episodes:

$$\text{Session} \supset \text{Episodes} \supset \text{Exchanges} \supset \text{Messages}$$

Definition 8 (Hierarchical Segmentation). Session boundaries are detected using tempo + multi-signal methods (high threshold), while episode boundaries use lexical coherence methods (lower threshold). This distinguishes true session breaks from within-session topic shifts.

10 Worked Example

Consider a human-LLM conversation with: $\beta^{\text{read}}_{\text{human}} = 0.25$ sec/word, $\beta^{\text{write}}_{\text{human}} = 0.5$ sec/word, $\beta^{\text{read}}_{\text{LLM}} \approx 0$, $\beta^{\text{write}}_{\text{LLM}} = 0.05$ sec/word, $\gamma = 0.9$.

Ex	Actor	ℓ_{prev}	ℓ_{curr}	Δt^{raw}	$\mathbb{E}[\text{resp}]$	Δt^{norm}	ρ	coh
E_1	Human	—	10	—	—	—	—	—
E_2	LLM	10	50	2	2.5	0	0.8	0.72
E_3	Human	50	8	15	16.6	0	0.9	0.68
E_4	LLM	8	40	1	2.0	0	0.5	0.75
E_5	Human	40	12	8	14.4	0	0.6	0.71
E_6	LLM	12	35	3	1.75	1.25	1.7	0.65
red!10 E_7	Human	35	5	420	12.2	407.8	34.4	0.31

Table 1: Example human-LLM conversation with computed metrics. Row E_7 (highlighted) shows a detected boundary.

Analysis: E_7 shows: (1) tempo_break = True ($\Delta t^{\text{norm}} = 407.8 \gg \tau$), (2) lexical_break = True (coh = $0.31 < 0.5$), (3) $\rho = 34.4 \gg 1$. The multi-signal approach confirms boundary with high confidence. Note that without normalization, E_3's raw gap of 15 seconds might falsely trigger a boundary, but after accounting for reading (50 words × 0.25) and writing (8 words × 0.5), the normalized IEI is 0.

11 Evaluation Metrics

Following established metrics from computational linguistics (Rohde & Frank, 2014):

- **WindowDiff:** Compares the number of boundaries posited within a sliding window to the true count, addressing issues with false positive/negative weighting.
- **P_k metric:** Probability that two random points are correctly classified as same/different segment.
- **Boundary precision/recall:** Standard metrics against human-labeled session boundaries.
- **Within-session tempo variance:** Should be lower than between-session variance (validates tempo model).
- **Topic coherence:** Sessions should maintain semantic consistency (validates multi-signal integration).

12 Recommended Configuration

13 Novel Contributions Beyond Prior Literature

This framework makes several contributions that extend beyond the existing literature:

1. **Tempo ratio (ρ) as primary signal:** While prior work uses absolute temporal thresholds or simple normalization, the tempo ratio provides a dimensionless measure of deviation from expected conversational rhythm that is actor-agnostic and adaptive.

Parameter	Value	Notes
$\beta_{\text{human}}^{\text{read}}$	0.25 sec/word	Reading speed ∼250 wpm
$\beta_{\text{human}}^{\text{write}}$	0.5 sec/word	Typing ∼40 wpm + thinking
$\beta_{\text{LLM}}^{\text{write}}$	0.05 sec/word	Near-instantaneous
γ (length scaling)	0.9	Sublinear composition
ℓ_{min}	1 word	Minimum effective length
τ_{min}	60 seconds	Minimum gap threshold
k (IQR multiplier)	1.5	Adaptive threshold
θ_{sim}	0.4	Lexical coherence threshold
θ_{rhythm}	2.0	Rhythm variance threshold
λ (EMA rate)	0.1	Slow baseline adaptation
Window size	20 exchanges	History for IQR computation

Table 2: Recommended parameter configuration for human-LLM conversations.

2. **Dual-component response model:** Separating reading/processing time from composition time addresses the asymmetric nature of human-LLM interactions not captured in prior dialogue segmentation work.

3. **Sublinear length scaling:** The $\gamma < 1$ exponent captures typing momentum effects observed empirically but not formalized in prior segmentation frameworks.

4. **Explicit integration of multi-signal validation:** While prior work acknowledges that temporal and lexical features are complementary, this framework provides explicit combination rules that prevent false positives from any single signal.

5. **Adaptive per-actor baselines:** Online estimation of actor-specific response characteristics using exponential moving averages, allowing the system to adapt to individual users.

6. **Hierarchical session/episode distinction:** Using different threshold levels for session boundaries versus topic episodes within sessions.

14 Conclusion

By incorporating actor-specific response velocities, normalizing by message length, and integrating lexical coherence and discourse markers, we transform raw temporal segmentation into a robust multi-signal tempo-based segmentation framework. This approach:

- Handles asymmetric response times in human-LLM interactions
- Accounts for both reading and composition time based on message length
- Focuses on conversational rhythm rather than absolute time
- Validates temporal signals against lexical coherence to reduce false positives
- Adapts to individual actor characteristics over time
- Provides hierarchical session/episode structure
- Uses established evaluation metrics from computational linguistics

Key insight: Session boundaries are detected when the tempo changes significantly AND discourse coherence supports the break, not just when absolute time gaps appear.

References

Artigas Miralles, L., Vilaregut Puigdesens, A., Feixas Viaplana, G., Mateu Martínez, C., Seikkula, J., & Vall Castelló, B. (2019). Dialogue and dominance in couple therapy for depression: Exploring therapists' responses in creating collaborative moments. *Family Process*, 59(3), 1080–1093.

Bunning, K., Smith, C., Kennedy, P., & Greenham, C. (2011). Examination of the communication interface between students with severe to profound and multiple intellectual disability and educational staff during structured teaching sessions. *Journal of Intellectual Disability Research*, 57(1), 39–52.

Kummamuru, K., Padmanabhan, D., Roy, S., & Venkata Subramaniam, L. (2009). Unsupervised segmentation of conversational transcripts. *Statistical Analysis and Data Mining*, 2(4), 231–245.

Linders, G. M. & Louwerse, M. M. (2023). Surface and contextual linguistic cues in dialog act classification: A cognitive science view. *Cognitive Science*, 47(10).

Morris-Adams, M. (2016). Negotiating topic changes: Native and non-native speakers of English in conversation. *International Journal of Applied Linguistics*, 26(3), 366–383.

Netten, N. & van Someren, M. (2011). Improving communication in crisis management by evaluating the relevance of messages. *Journal of Contingencies and Crisis Management*, 19(2), 75–85.

Rohde, H. & Frank, M. C. (2014). Markers of topical discourse in child-directed speech. *Cognitive Science*, 38(8), 1634–1661.

Rossiter, E. L. (2021). Measuring agenda setting in interactive political communication. *American Journal of Political Science*, 66(2), 337–351.

Soma, C. S., Wampold, B. E., Flemotomos, N., Peri, R., Narayanan, S., Atkins, D. C., & Imel, Z. E. (2022). The silent treatment? Changes in patient emotional expression after silence. *Counselling and Psychotherapy Research*, 23(2), 378–388.

van Os, M., de Jong, N. H., & Bosker, H. R. (2020). Fluency in dialogue: Turn-taking behavior shapes perceived fluency in native and nonnative speech. *Language Learning*, 70(4), 1183–1217.