On the Structure of Address in Multi-Party Dialogue: From Discrete Labels to Continuous Levels

5th August 2026

SIGDIAL 2026 in Atlanta, USA

Taiga Mori, Koji Inoue, Divesh Lala, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan

Addressee Detection: The Standard Formulation

Prior work assumes address is DISCRETE: one label per utterance

- Human–human–computer: binary — system, or another human?
  Baba+ 2011; Shriberg+ 2012; Tsai+ 2015
- Multi-party human dialogue: pick ONE label — A / B / C, or “group”
  Jovanovic+ 2006; Ouchi & Tsuboi 2016; Gu+ 2021
- Predicted after the turn ends — post-hoc, utterance level
- Mainly an auxiliary task for turn-taking / response selection

But Is Address Really Discrete?

RQ: Is address graded, and does it relate to listener behavior?

- One utterance can address several listeners to different degrees
- Under a single-label scheme, disagreement is discarded as noise
- Address is modeled in isolation, apart from listener behavior
- Prediction happens only after the turn is complete

This Work: Three Contributions

Discrete label → continuous, per-listener address level

1. Evidence that address has graded structure
   → estimate it continuously, do not classify it

2. An operationalization: the address level
   → a latent value per listener, from categorical labels

3. It relates to turn-taking, gaze and backchannels
   → evidence for its validity

Dataset: Teidan Corpus

Teidan corpus: 36 Japanese triadic dialogues, 3,121 turns

- 12 groups × 3 topics
- Total 3 h 38 min
- 86.7 turns per dialogue
- Annotated: turns, backchannels, gaze

Addressee Annotation

3 annotators, 8 labels — agreement is only moderate (κ = 0.52)

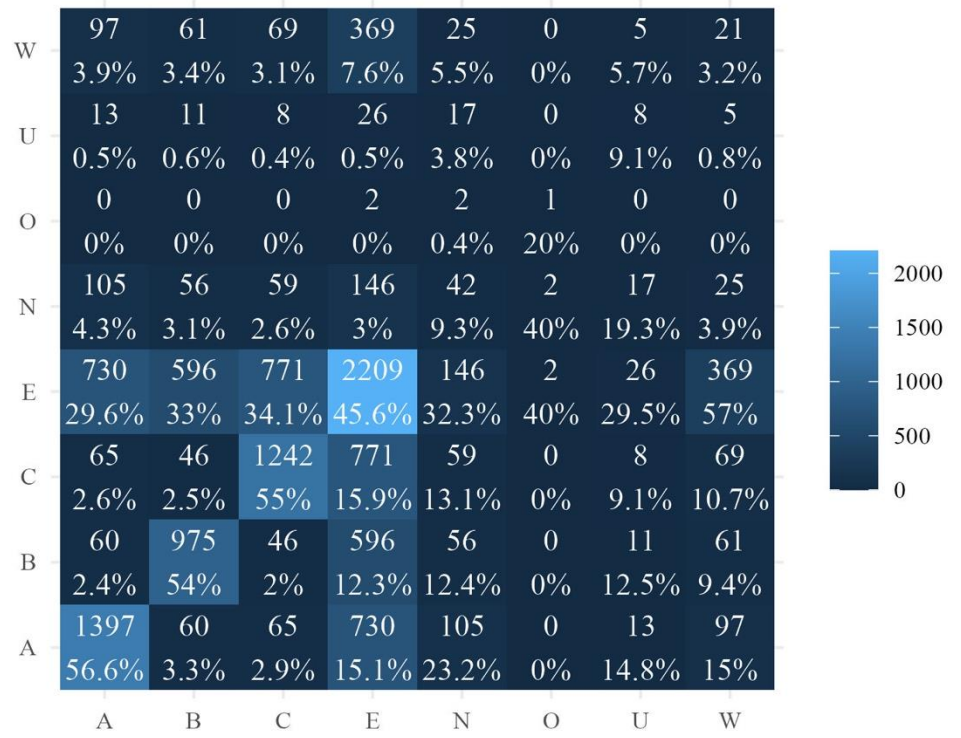
- Label set from Kadota+ 2024
- 3,078 turns, 3 annotators each
- Multimodal cues allowed
- No information after the turn

Label	Meaning	Count
A	participant A	1,932
B	participant B	1,390
C	participant C	1,751
E	everyone	3,529
W	whoever	334
N	none (monologue)	247
O	other (outside)	3
U	unknown	48

Annotators Disagree — Systematically

Individual vs. group is not a clean line: labels leak into E

- A / B / C agree with themselves > 50%
- 2nd partner is always E (~30%)
- E with itself 45.6%; A/B/C ~15%
- W co-occurs with E, not with itself



Two Representations of Addressivity

The label gives 0 or 1; the level gives anything in between

Address LABEL — 0 or 1

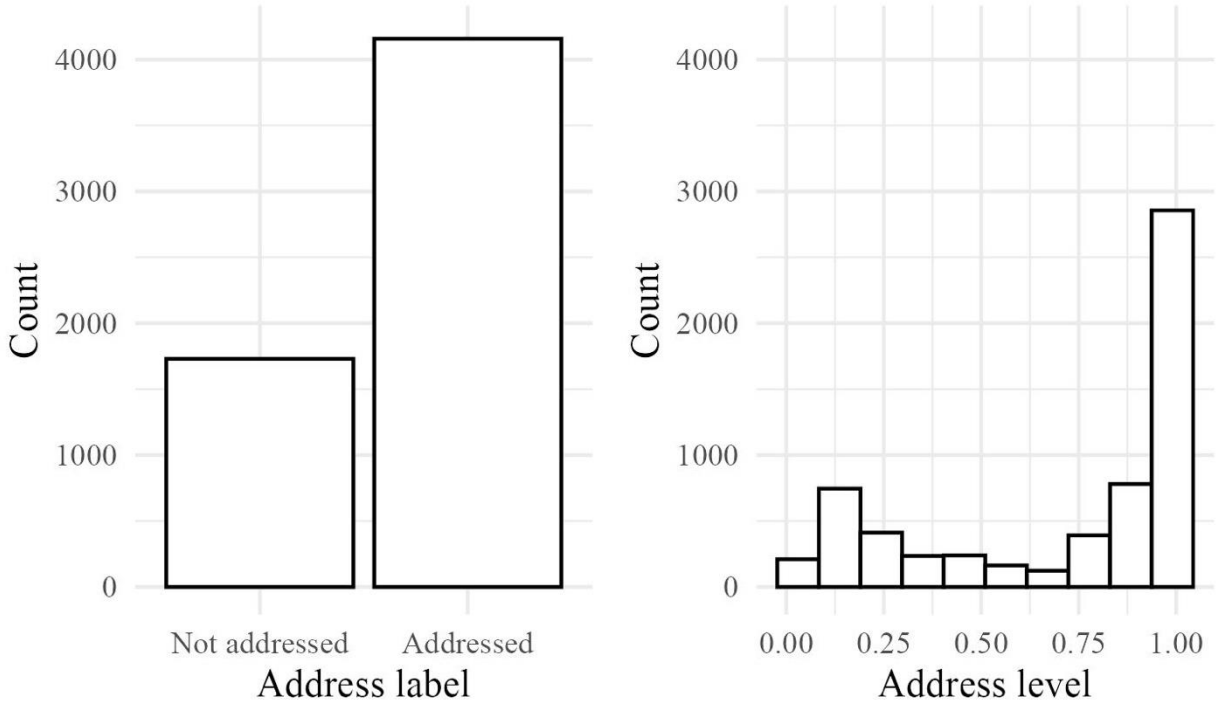
- Merge W into E; drop U
- Majority vote per turn
- Listener or E → addressed; otherwise → not addressed

Address LEVEL — 0 to 1

- Latent $\theta \in (0, 1)$ per listener
- Ideal points: L1 (1,0), L2 (0,1), E (1,1), N (0,0)
- $Pr = \text{softmax}(\alpha - \lambda \|\theta - v\|^2)$

Distributions of Label and Level

Partially addressed states exist — not only 0 or 1



- Label: addressed ≈ 2× not addressed
- Level: mass at 0 and at 1 …
- … but the middle is not empty

Statistical Models

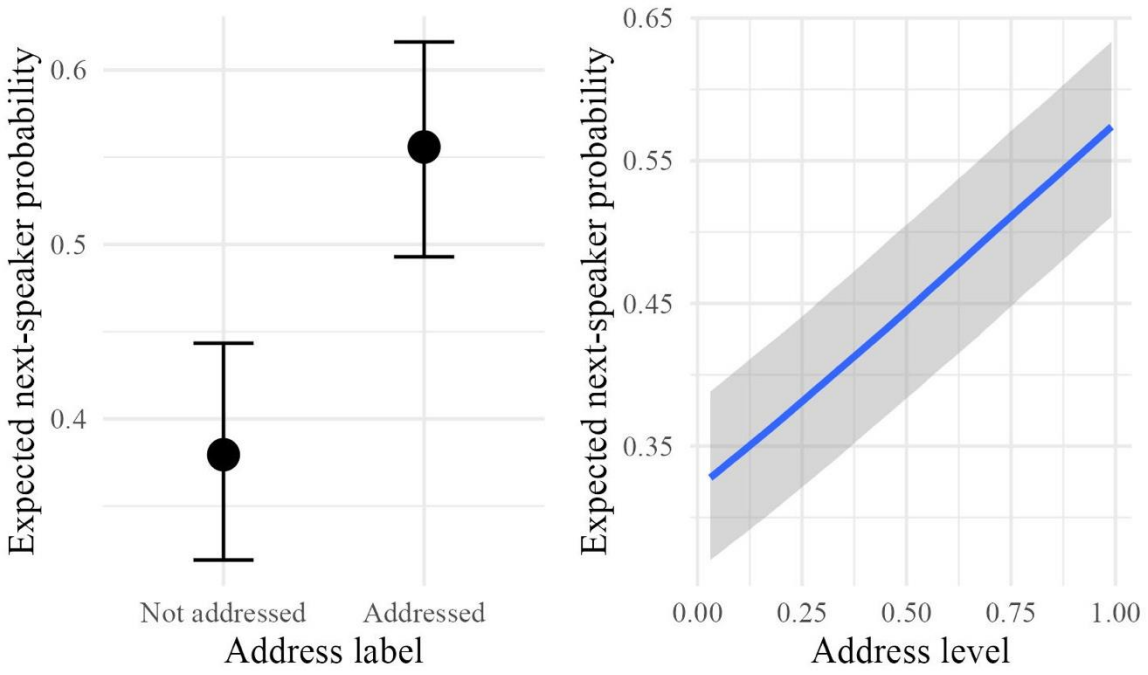
Label vs. level: same data, compared by predictive fit (ELPD)

- Bayesian GLMM (brms); fixed effect: address label OR level
- Random intercepts: turn, session, speaker, listener
- Compared by leave-one-out cross-validation

Outcome (per listener, per turn)	Model	Observations
Became the next speaker	Bernoulli (logit)	5,820
Gaze at the current speaker	Beta (proportion)	5,744
Number of backchannels	Poisson + log(duration)	4,920

Result 1: Next Speaker

Higher addressivity → next speaker; the level fits better



left: label / right: level

Address label

β = 0.72 [0.59, 0.84]

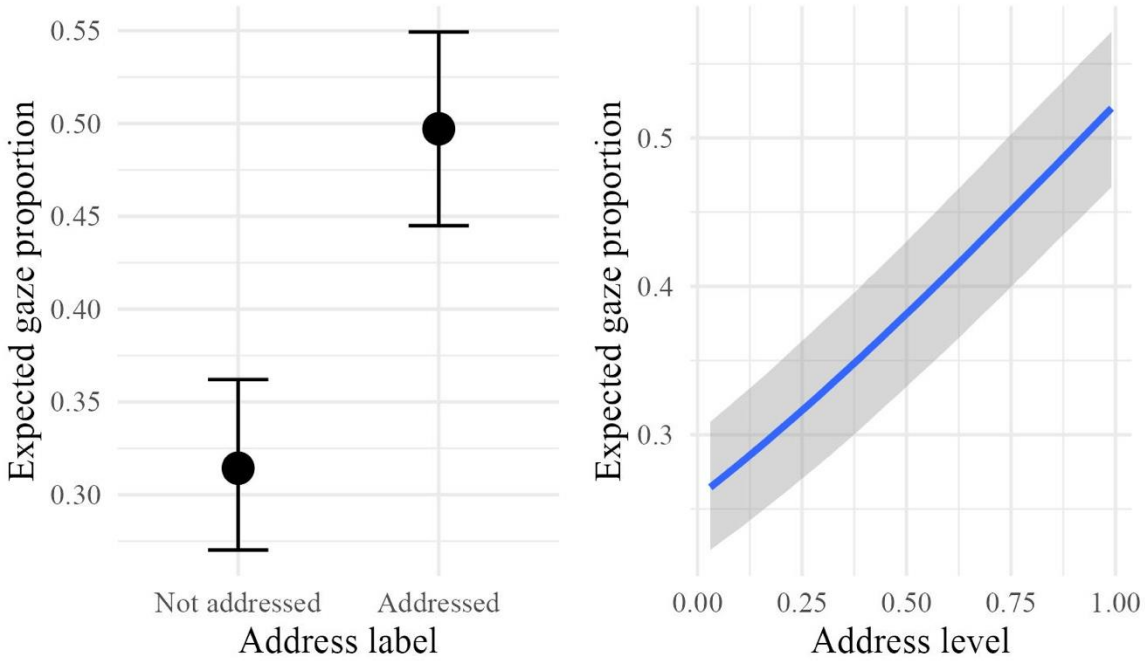
Address level

β = 1.06 [0.89, 1.23]

ELPD: +14.4 (SE 4.9)

Result 2: Listener Gaze

Higher addressivity → longer gaze; the largest gain of the three



left: label / right: level

Address label

β = 0.77 [0.69, 0.85]

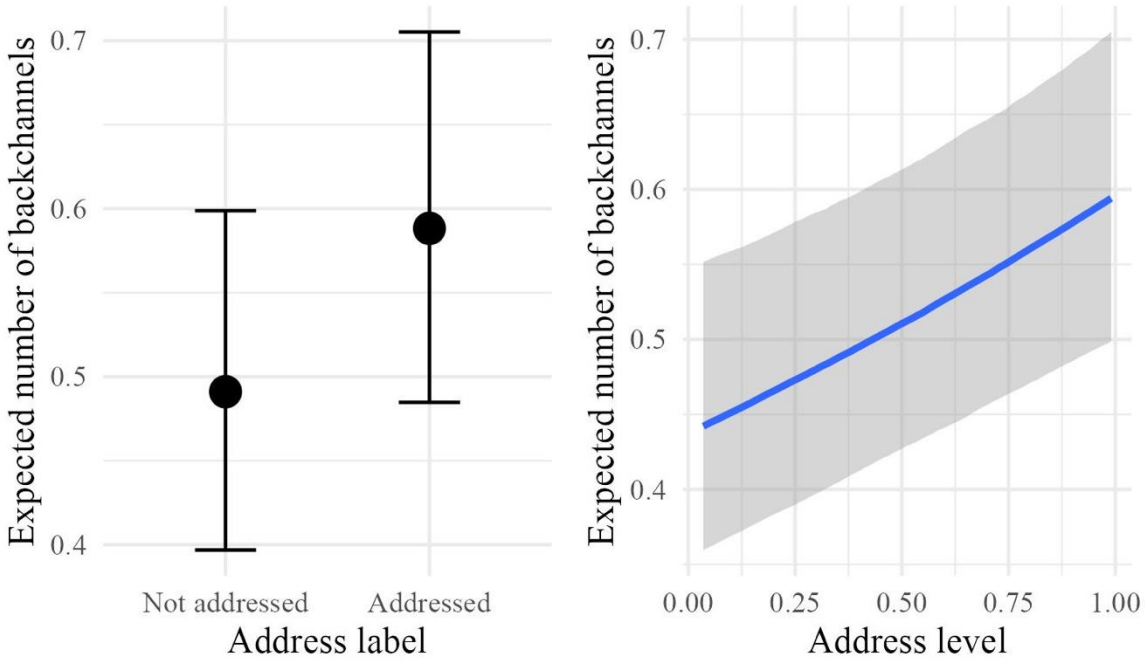
Address level

β = 1.15 [1.04, 1.26]

ELPD: +39.2 (SE 7.9)

Result 3: Backchannels

Higher addressivity → more backchannels, but other factors too



left: label / right: level

Address label

β = 0.18 [0.08, 0.28]

Address level

β = 0.31 [0.17, 0.45]

ELPD: +3.5 (SE 1.6) — smallest

Qualitative Example: Double Address

One answer serves two questions: primary C, weaker B

01 C: What were we to do in Tokyo?
gazing at B

02 B: Was there something to do?
gazing at A

03 A: Shopping.
gazing at C

- 03 answers B in sequence, but C in content
- 2 annotators: C, C, E
- So C is primary, B is weaker

Conclusions & Future Work

Estimate address continuously and online, with listener behavior

- Per-listener levels keep what majority votes discard
- Scalable: independent of the number of participants
- Future: online joint modeling of gaze, backchannels, turn-taking
- Limits: Japanese triads; 3 annotations per turn

Thank you!

Contact: mori@sap.ist.i.kyoto-u.ac.jp