K-MIND: Korean Multimodal INteraction Data for Dyadic Conversation Analysis

Jae Hee Yang[1], Yuha Shin[2], Saim Shin[1], Je Woo Kim[1], Jin Yea Jang[1,*]
[1]Korea Electronics Technology Institute Seongnam-si, Gyeonggi-do, South Korea
[2]MaumAI Seongnam-si, Gyeonggi-do, South Korea
{yangjaehee, sishin, jwkim, jinyea.jang}@keti.re.kr, yuhashin@maum.ai

Abstract

We present the Korean Multimodal INteraction Data (K-MIND), a large-scale corpus of dyadic Korean dialogue that is designed to capture the multimodal richness of social interaction. The dataset includes 292 participants and 200 sets (935 clips) spanning 115 hours and 30 minutes, all aligned across verbal, paraverbal, and nonverbal modalities such as transcripts, acoustic features, and visual signals. For these modalities, we propose a comprehensive annotation scheme that enables nuanced yet consistent labeling of complex communicative behaviors, balancing theoretical soundness with practical feasibility. We further report analysis results of the corpus, including label distributions, within- and cross-layer analyses. These analyses illuminate the key properties of dyadic K-MIND and demonstrate its utility for advancing research in human–computer interaction as well as in interdisciplinary domains. To ensure continuous refinement, the corpus and framework are being validated in complementary studies and have been extended to triadic interactions (K-MIND Triadic) that model group dynamics, which will be included in upcoming releases.

Keywords: multimodal interaction corpus, conversational AI, annotation scheme

1. Introduction

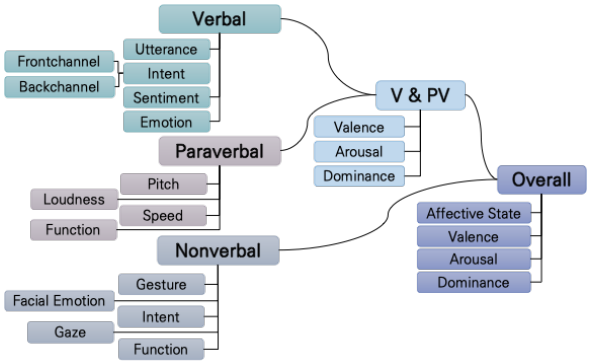

Figure 1: K-MIND Structure

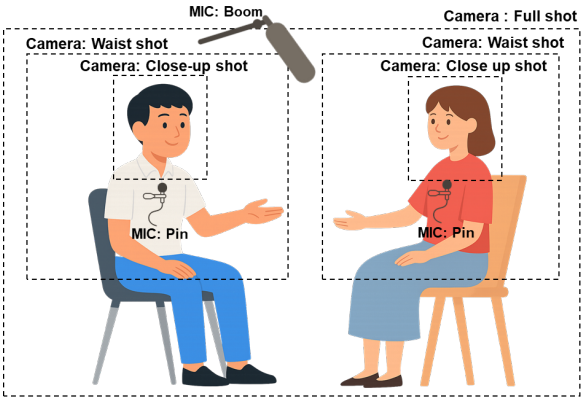

Figure 2: K-MIND Recording Environment and Acquisition Setup

Understanding the dynamics of human conversational interaction requires data that capture the full range of communicative behaviors. Over recent decades, numerous corpora have been developed with varying scales, modalities, and interactional settings (Serban et al., 2018). While early resources centered on audio or text (Godfrey et al., 1992; Canavan et al., 1997; Forsythand and Martell, 2007; Sordoni et al., 2015), the demand for multimodal corpora has grown (Serban et al., 2018), enabling analyses of not only what is said but also how it is expressed.

Among early multimodal efforts, the D64 Corpus (Oertel et al., 2013) and AMI Meeting Corpus (Kraaij et al., 2005; Renals et al., 2007) provided key insights into face-to-face interaction. Projects such as CCDb (Aubrey et al., 2013) and its 4D extension (Marshall et al., 2015) emphasized high-resolution video and motion tracking. MELD (Poria et al., 2019) has been widely used for multimodal emotion recognition, while SIMMC and SIMMC 2.0 (Moon et al., 2019; Kottur et al., 2021) target situated, task-oriented dialogues. A large-scale advance is represented by CANDOR (Reece et al., 2023), with unscripted video-mediated dialogues totaling over 850 hours. More recently, the Seamless Interaction corpus (Agrawal et al., 2025) scales face-to-face dyadic recordings to over 4,000 hours and enables interaction-conditioned modeling of audiovisual behavior. In parallel, large-scale multi-

* Corresponding author

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9105–9117
11–16 May 2026.

modal emotion recognition datasets such as SAMSEMO (Bujnowski et al., 2024) provide multilingual video-based resources for modeling affect at the level of individual scenes or segments.

Despite these advances, existing corpora still fall short of capturing the full complexity of human interaction—which unfolds through intertwined verbal, paraverbal (e.g., intonation, speech rate, stress), and nonverbal (e.g., gaze, gesture, posture) signals that jointly convey meaning and emotion in real time. Some resources achieve ecological validity at a small scale (e.g., CCDb), while others reach large scale through task-based, segment-centered, or behavior-generation-oriented designs (e.g., MELD, SIMMC, SAMSEMO, Seamless Interaction). However, few provide a hierarchically integrated representation that aligns these modalities within spontaneous, face-to-face conversational settings. Consequently, large-scale, naturalistic, and multimodally coordinated interaction data remain scarce, particularly in languages beyond English.

This scarcity is especially evident for Korean, where existing multimodal resources are limited in both scale and interactional coverage. The Korean Multimodal Discourse Corpus (Kim, 2015), for example, provides detailed gesture annotation but covers only six short dyadic cartoon-narration sessions (about 35 minutes), which constrains its representativeness for large-scale modeling of spontaneous interaction. The KEMDy20 corpus (Kim et al., 2020) offers a multimodal conversational dataset integrating speech, text, video, and physiological signals for emotion recognition, comprising approximately 7 hours of dyadic interaction (about 9,900 utterances from 80 participants). However, the interactions are elicited through emotion-inducing stimuli and are primarily designed for affective computing rather than for capturing the fine-grained dynamics of naturally unfolding discourse.

To address this gap, the present study introduces the Korean Multimodal INteraction Data (K-MIND)[1], a corpus capturing spontaneous dyadic interaction in real-time. Unlike task-based or scripted corpora, K-MIND prioritizes unscripted, naturalistic exchanges, recording extended conversations that reflect everyday communication. An overview of the dataset's hierarchical annotation structure is illustrated in Figure 1, which shows how verbal, paraverbal, and nonverbal signals are systematically organized and integrated into higher-level affective dimensions. The dataset encompasses a rich range of communicative cues—speech, prosody, gesture, facial expression, and interactional timing—across synchronized audio, video, and text. It further distinguishes itself through its scale, duration, and diversity of participant relationships, which are described in detail in Section 2.

As a large-scale and realistically grounded resource, K-MIND complements existing corpora and extends empirical foundations in an underrepresented language. This contribution supports cross-linguistic research and provides a foundation for developing culturally adaptive and socially aware AI systems. The remainder of this paper is organized as follows: Section 2 details the dataset design and collection, Section 3 outlines our multimodal annotation scheme, Section 4 reports data analysis results, Section 5 discusses implications, and Section 6 concludes the paper.

2. The Corpus

A total of 292 native Korean participants (165 females, 127 males; ages 10–79) were recruited. To ensure diversity, they were paired across relationship types (e.g., couples, friends, siblings, colleagues, grandparents–grandchildren), with some pairs real and others simulated. Each engaged in everyday social topics (e.g., relationships, stress, sports) to elicit varied emotions.

Conversations were recorded over a three-year period (2022–2024) using a five-camera setup consisting of two close-up shots (face), two waist shots (upper body), and one full shot (entire scene), along with a three-microphone configuration comprising one boom mic and two pin mics (Figure 2). Participants interacted either while seated or standing. Recordings were stored in MP4 format (1920×1080, 29.97/30 fps) and WAV format (48 kHz), capturing multimodal cues such as facial expressions, gestures, gaze, and vocal tone. The dataset comprises 200 sessions and 935 clips, totaling 115 hours and 35 minutes of recordings, with an average clip length of 7:02 (ranging from 2:01 to 17:04).

All sets were transcribed and temporally aligned with audio for verbal, paraverbal, and nonverbal annotation using a custom-built tool (Section A, Figure 9) in a three-stage process. Due to the scale of the dataset, the fine-grained temporal alignment of multimodal signals, and the complexity of the hierarchical annotation scheme, full double annotation was not feasible. Therefore, we designed a multistage expert review protocol to ensure consistency and validity of the annotations. In the first stage, verbal content was transcribed, and nonverbal behavior descriptions were documented using predefined dictionary entries specifying the agent, action, and object of each movement, thereby constraining annotator interpretation. In the second stage, annotation labels were assigned to each temporally aligned verbal, paraverbal, and nonverbal unit. In the third stage, two reviewers sequentially conducted error checking; disagreements were resolved through adjudicated consensus with the orig-

[1] https://github.com/AIRC-KETI/K-MIND

Intent (Abbrev)	Description and Example
(a) Request Subjective (Req.Subj)[a]	Speaker asks about listener's personal preferences, experiences, or opinions. *e.g., Do you like traveling?*
(b) Elaboration Subjective (Elab.Subj)	Speaker expresses own personal preferences, experiences, or opinions. *e.g., I like traveling.*
(c) Request Objective (Req.Obj)	Speaker seeks factual knowledge, explanations, or meanings. *e.g., Where is Gyeongbokgung Palace located?*
(d) Elaboration Objective (Elab.Obj)	Speaker provides factual knowledge, explanations, or meanings. *e.g., Gyeongbokgung Palace is located in Jongno-gu, Seoul.*
(e) Agreement (Agr)	Listener expresses consent or concurrence. *e.g., Totally agree.*
(f) Disagreement (Disagr)	Listener expresses opposition or disapproval. *e.g., If I were in your position, I wouldn't do it.*
(g) Evaluation	Listener conveys judgment or emotional response. *e.g., That is great!*
(h) Completion	Listener predicts and completes unfinished sentence. *e.g., A: It was so… B: Unexpected.*
(i) Clarification	Listener verifies or confirms information. *e.g., A: It broke. B: The window?*
(j) Suggestion	Speaker advises or proposes. *e.g., It's better not to touch to avoid infection.*
(k) Recall	Speaker evokes previously shared information or memories. *e.g., Haven't you said your mom majored in English literature?*
(l) Social Acts (Soc.Acts)	Speaker conveys apologies, gratitude, greetings, etc. *e.g., I'm sorry.*
(m) BC Continuer (BC Cont)	Backchannel encouraging continuation. *e.g., Mm-hmm./Uh-huh.*
(n) BC Understanding (BC Ustand)	Backchannel signaling comprehension. *e.g., Ah~*
(o) BC Negative Surprise (BC NegSurp)	Backchannel expressing shock or disapproval. *e.g., Oh no!/What?*
(p) BC Positive Surprise (BC PosSurp)	Backchannel expressing amazement or delight. *e.g., Wow!/Oh!*
(q) BC Request Confirmation (BC Req.Conf)	Backchannel seeking confirmation/clarification. *e.g., Really?*
(r) BC Affirmative (BC Affirm)	Backchannel conveying agreement. *e.g., Yeah./Right.*
(s) Encouragement	Utterances conveying support or uplifting message. *e.g., Everything will be fine./You can do it!*
(t) Reflection	Listener interprets speaker's cues and verbalizes inferred meaning. *e.g., A: (laughs) B: Judging by your smile, it seems something fun happened.*
(u) Unclassifiable	Cannot be categorized into the above types

Table 1: Categories of Verbal Intents

[a] Request Subjective was refined into two subtypes: Closed (yes/no, choice, confirmation) and Open (what, why, how, and others such as when, where, who).

Function	Description and Example
Amplification	Emphasizing content words via speed, stress, or pitch *e.g., I went to the market YESTERDAY. (emphasizing time)*
Expressivity	Conveying emotions or attitudes through vocal modulation *e.g., I wanted to … try it… (lowering volume and slowing down to express disappointment)*
Grammar	Marking grammatical functions with intonation *e.g., You're coming? (raising intonation to indicate a question)*
Others	None of the above

Table 2: Functions of Paraverbal Behavior[a]

[a] Utterances could carry multiple labels.

inal annotator.

This annotation process was conducted over a period of two years and five months (2023–2025) by a total of 21 contributors. In total, 19 individuals participated in transcription and labeling as annotators and 5 individuals served as reviewers for validation and quality control, with 3 contributors participating in both roles. To ensure consistency and accuracy, regular feedback meetings were held at least once a week during the first and the second year and at least once a month during the third year. In addition to the full review of the dataset, random spot-checking was performed to identify and correct potential annotation errors.

Function	Description and Example
Representational	Depicts objects, actions, or concepts *e.g., arm swinging to depict 'running'*
Highlighting	Emphasizes amount, strength, or degree *e.g., spreading arms while saying 'a lot'*
Pointing	Indicate referents *e.g., pointing to oneself while saying 'I'*
Focusing	Marks distinguishing points in discourse, often rhythmic or repetitive *e.g., tapping fingers on the table to emphasize a key point*
Responsive	Shows affective response or attentiveness *e.g., nodding to show agreement*
Attitudinal	Expresses attitudes or emotional states *e.g., yawning to show boredom*
Conventional	Conveys culturally established meaning *e.g., waving for greeting*

Table 3: Funtions of Nonverbal Behavior

Label	Description and Example
Happiness	Positive state toward a subject/object *e.g., happiness, excitement, enthusiasm, delight*
Sadness	Mild sadness or disappointment *e.g., sadness, regret, sorrow, melancholy, disappointment*
Anger	Resistance or confrontational mindset *e.g., anger, displeasure, annoyance, irritation*
Surprise	Spontaneous reaction to the unexpected *e.g., surprise, astonishment, amazement, wonder*
Afraid	Unease from perceived threat or danger *e.g., concern, worry, anxiety, distress, restlessness*
Fear	Intense fear response to threat *e.g., fear, dread, terror, fright*
Disgust	Strong aversion to something offensive *e.g., disgust, aversion, antipathy*
Contempt	Disrespect or disdain toward someone/something *e.g., contempt, mockery*
Shame	Embarrassment over flaws/mistakes *e.g., humiliation, guilt, disgrace*
Hope	Hope or anticipation for the future *e.g., hope, desire, dream, eagerness*
Interest	Interest in ongoing topics *e.g., interest, amusement, attention*
Boredom	Lack of interest in ongoing topics *e.g., boredom, indifference, weariness*
Thinking	Hesitation or indecision *e.g., hesitation, indecision, uncertainty*
Humour	Humor through wordplay or comedy *e.g., joke, wordplay*
Neutral	Emotionally flat state not fitting other categories

Table 4: Categories of Overall Affective States

3. Annotation Levels for Multimodal Interaction Data

The multimodal annotation was organized into three independent layers—verbal, paraverbal, and nonverbal—with dedicated annotator groups for each layer. All features were manually annotated according to modality-specific guidelines, as described in the following subsections, except for gaze and prosodic features, which were automatically extracted.

3.1. Verbal Level

Verbal communication here refers to verbal behaviors in spoken interaction, represented through grammatical and semantic structures. While drawing on prior verbal studies(Kim, 2002; Clark and Haviland, 1975; Ito et al., 2020; Lee, 2016), our own discourse analysis of communicative patterns observed in the K-MIND corpus allowed us to identify seven core features of human communication: people (i) exchange both subjective and objective information (see Table 1 (a–d)); (ii) respond empathically to their partner's utterances (see Table 1 (e–g)); (iii) actively engage by predicting, completing, or clarifying what is said (see Table 1 (h–j, s, f)); (iv) offer advice or suggestions when appropriate (see Table 1 (l)); (v) retrieve content by recalling shared experiences or history (see Table 1 (k)); (vi) perform social speech acts such as apologizing, expressing gratitude, and greeting (see Table 1 (l)); and (vii) provide timely backchannels that signal understanding and engagement (see Table 1 (m–r)).

3.2. Paraverbal Level

The paraverbal level includes prosodic properties that convey emotions, attitudes, and emphasis beyond literal meaning. In K-MIND, these features—specifically loudness (intensity), pitch (F0), and speech rate were automatically extracted using openSMILE(Eyben et al., 2010, 2013), and annotators classified segments deviating from neutrality into four functions—Amplification, Expressivity, Grammar[2], and Others—based on modified categories from prior studies (Eggins and Slade, 2004; Martin and White, 2003; Chun, 2002) (Table 2).

3.3. Nonverbal Level

The nonverbal level focuses on facial expressions and movements of the hand, head, and upper body.

[2]Grammar was identified through automatic extraction of intonation patterns marking grammatical functions.

Agent	Action	Theme
Hand (one/both)	clapping, covering, counting, folding, folding arms, grabbing, hitting, lifting, pointing, scratching/rubbing, spreading, touching, unfolding, waving, shaping[a]	arm, back, belly, chest, ear, face, forehead, hand, head, jaw, mouth, nail, neck, nose, shoulders, waist
Head	circling, head up, leaning back, nodding, shaking, turning left or right, turtle neck	oneself, someone, others
Upper Body (Torso)	leaning, moving, sitting back, sitting straightly	NA
Shoulder	circling, shaking, shrugging	NA
Face (Mouth)	biting lips, biting nails, fully smiling, laughing, opening, pouting, pressing lips, puffing, smirking	NA
Face (Eye(s))	closing, opening wide, narrowing, winking, raising, rolling	NA
Face (Eyebrows)	raising, frowning	NA

Table 5: Tagging Categories for Nonverbal Behavior Annotation

[a] Unlike other labels, the label 'shaping' requires a natural language description.

Only clearly manifested gestures were annotated, using the [agent:action:(theme)] format (Kim, 2015; Son, 2012), where 'agent' denotes the moving body part, 'action' the gesture, and 'theme' the affected part or object (optional). Annotators marked onset/offset and described the primary stroke[3], drawing from a predefined dictionary or natural language when necessary (Table 5). Gestures were classified into seven categories—Representational, Highlighting, Pointing, Focusing, Responsive, Attitudinal, and Conventional (Table 3[4]) (Tan et al., 2010). Gaze information was extracted from video using an automated gaze estimation method (Ryan et al., 2025). Based on the estimated gaze likelihood toward the interlocutor, gaze was labeled as on-person when the participant was estimated to be looking at the interlocutor, and off-person otherwise.

3.4. Emotion and Overall Affective State

To describe emotions, we adopted the six basic categories—Happiness, Sadness, Fear, Surprise, Anger, Disgust (Ekman, 1992), and —and extended them to nine by adding Neutral, Contempt, and Other (Öhman, 2020). Emotion was annotated from transcripts at the verbal level, excluded for paraverbal cues, and confined to facial expressions for the nonverbal level, as other body movements were less reliable indicators. Unlike prior studies that analyzed channels separately (Öhman, 2020), K-MIND introduces an additional layer—Overall Affective State—to capture holistic impressions of speakers' emotions or attitudes, acknowledging that verbal, nonverbal, and paraverbal cues cannot be cleanly weighted[5]. Beyond emotion and overall affective state, two further dimensions were annotated: Sentiment, targeting evaluative verbal content (e.g., preference, approval, disapproval), with neutral otherwise; and VAD—Valence (positive–negative), Arousal (activation), and Dominance (sense of control) (Mehrabian and Russell, 1974; Russell, 1980; Russell and Barrett, 1999). Together, these layers support analyses of how different perspectives, alone or combined, capture multiple aspects of affect in multimodal interaction.

4. Descriptive Statistics and Interactional Patterns in K-MIND

This section analyzes annotation distributions to highlight key properties of the dataset.

[3] Rather than distinguishing gesture phases such as preparation, stroke, hold, and retraction (Kendon, 1967), we focused on the communicative functions conveyed by each gesture.

[4] Annotators reported some difficulty in distinguishing meaningless movements from the focusing function; however, no systematic confusion was observed for the other categories due to distinct operational criteria in the guidelines. Highlighting, in particular, was anchored to adverbial modifiers and therefore did not overlap with other functions.

[5] Unlike previous studies reporting confusion or increased difficulty in distinguishing negative emotions (Laukka et al., 2005; Busso et al., 2008; Singh et al., 2023), confusion between negative emotions such as fear and anger was not prominent in our data, likely due to the everyday conversational topics used in our corpus, which rarely elicited fear, and the multimodal integration of verbal, paraverbal, and nonverbal cues. Annotators favored a finer-grained affective inventory, while the overall affective state was intentionally designed as an impression-based label, making its subjectivity a theoretically motivated property rather than a source of inconsistency.

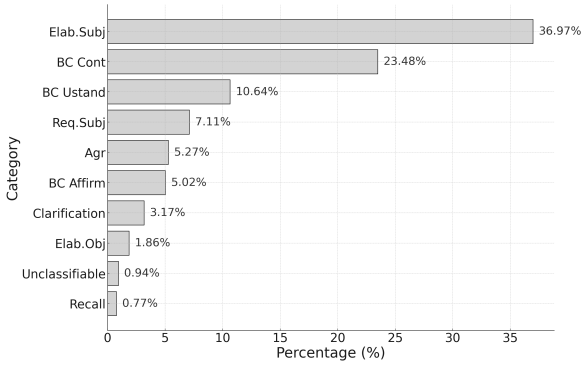

Figure 3: Top 10 Verbal Intents

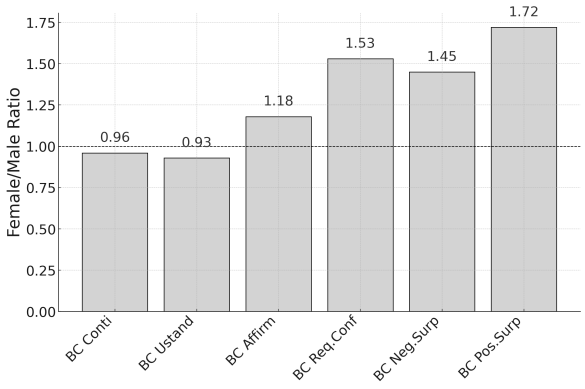

Figure 4: Relative Frequency of Backchannel Types by Gender

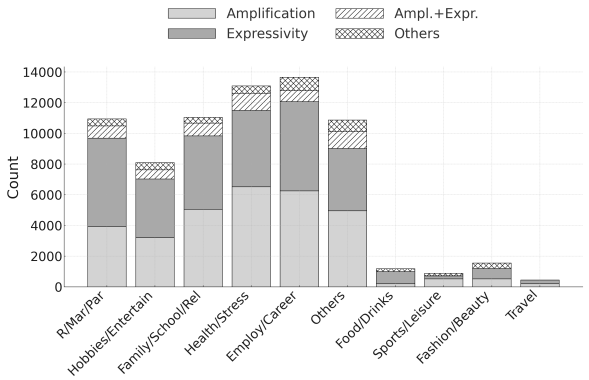

Figure 5: Paraverbal Function Distribution by Topic

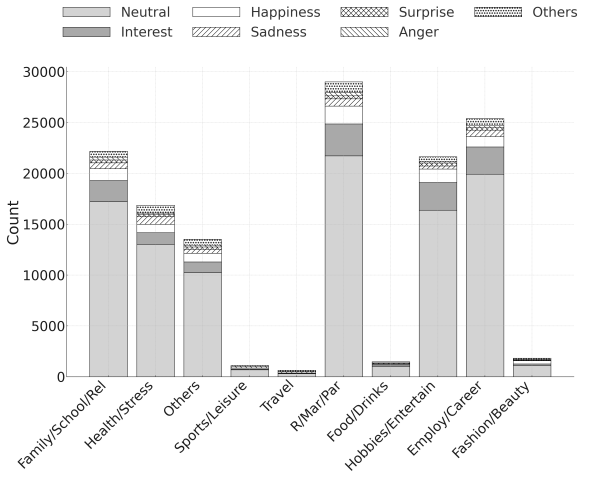

Figure 6: Overall Affective State Distribution by Topic

4.1. Verbal Behaviors

In total, 133,595 verbal labels were annotated, with the top 10 intents collectively accounting for 95.23% of all labels (Figure 3).The top three verbal intents, Elaboration Subjective, BC Continuer, and BC Understanding, together accounted for 71.09% of all utterances, indicating that most interactions centered on personal sharing supported by listener feedback. Within the top ten, empathic backchannels (BC Continuer, BC Understanding, BC Affirmative) comprised 39.14%, underscoring the role of attentiveness and approval in sustaining dialogue. Request Subjective (7.11%) and Clarification (3.17%) further highlight how listeners contributed to the interaction by asking questions and verifying information.

To assess the single-channel feasibility of verbal analysis, we analyzed six verbal backchannel subtypes (per 1,000 utterances)(Figure 4). Gender (Female/Male) ratios revealed clear demographic variation: BC Continuer (0.96) and BC Understanding (0.93) were close to balance, showing little gender difference in basic conversational flow. By contrast, BC Affirmative (1.18), BC Request Confirmation (1.53), and both BC Negative Surprise (1.45) and BC Positive Surprise (1.72) were considerably higer for women, indicating that women employed these affective backchannels more frequently.

4.2. Paraverbal Behaviors

As for paraverbal functions, Expressivity (44.55%, n=33,733) and Amplification (42.57%, n=32,235) predominated, followed by Amplification & Expressivity (7.84%, n=5,940) and Others (5.04%, n=3,814), indicating that paraverbal cues were mainly realized through emphasis and emotional expression.

When analyzing the distribution of paraverbal functions by topic (Figure 5), clear domain-specific patterns emerged. Expressivity was most frequent in topics such as Romance/Marriage/Parenting (52.6%, n=5,767), Hobbies/Entertainment (47.3%, n=3,832), Food/Drinks (67.3%, n=798), Fashion/Beauty (43.9%, n=683), and Travel (50.1%, n=245), whereas Amplification dominated in Health/Self-management/Stress (49.8%, n=6,524), Employment/Career/Workplace (45.8%, n=6,262), Family/School/Interpersonal Relations (45.8%, n=5,057), Sports/Leisure (58.8%), and Others (45.6%). Amplification & Expressivity appeared in

most topics at a supplementary level (0.2–10.0%). Overall, the distribution highlights topic-sensitive tendencies: affective domains favored Expressivity, whereas informational or pragmatic domains favored Amplification, with hybrid realizations consistently remaining secondary across topics.

4.3. Nonverbal Behaviors

In total, 233,492 gestures were annotated, with the 20 most frequent agent–action combinations (Table 6) accounting for 84.76%. Head movements were most prominent: nodding alone represented over one-third of all gestures and predominantly served responsive functions. Additional head gestures—tilting, shaking, lifting, and thrusting forward—raised their share to nearly 45%. Hand gestures were more diverse, including pointing, waving, raising, spreading, hitting, unfolding, and shaping, and were mainly used for rhythmic marking, deictic functions, or emphasis. Facial gestures such as smiling, fully smiling, eyebrow raising/narrowing, and eye-widening were less frequent but essential for affective expression. Overall, the distribution highlights the centrality of head movements in regulating interaction, the multifunctionality of hands for focusing and deixis, and the specialization of facial gestures in emotional communication.

The top three nonverbal function categories—Attitudinal (31.26%, n=72,980), Responsive (31.24%, n=72936), and Focusing (21.72%, n=50,710)—accounted for about 84% of all behaviors, suggesting that gestures primarily conveyed stance, feedback, or discourse marking. Representational (7.50%, n=17,505) and Pointing (7.16%, n=16,718) served secondary but notable roles, whereas Highlighting (1.00%, n=2,337) and Conventional (0.13%, n=306) were rare.

4.4. Integrative Analysis of Verbal, Paraverbal, and Nonverbal Behaviors

4.4.1. Overall Affective State

This study introduces the annotation of overall affective state (n=133,596), acknowledging that humans process verbal, paraverbal, and nonverbal cues holistically and instantaneously without fixed weights. About three-quarters of the data were Neutral (76.03%), while positive affect—Interest and Happiness—accounted for 15.47%. The non-neutral remainder was 8.50%, comprising negative affect (6.00%; Sadness, Anger, Afraid, Disgust, Fear, Shame, Contempt, Boredom) and other states (2.51%). This indicates that K-MIND mainly captures everyday conversations marked by neutral and positive affect, rather than strong or extreme emotional expressions.

Table 6: Top 20 Agent–Action Gesture Combinations

Rank	Agent+Action	Frequency (%)
1	Head+Nodding	36.22
2	Eyebrows+Raising	5.54
3	One Hand+Pointing	4.88
4	Two Hands+Waving	4.70
5	One Hand+Waving	3.85
6	Mouth+Smiling	3.47
7	Two Hands+Spreading	2.74
8	Head+Tilting	2.71
9	Head+Shaking	2.28
10	Head+Lifting	2.17
11	Two Hands+Pointing	2.10
12	Mouth+Fully Smiling	2.10
13	One Hand+Raising	1.76
14	Eyebrows+Narrowing	1.74
15	One Hand+Hitting	1.54
16	Eyes+Widening	1.50
17	Head+Thrusting Forward	1.44
18	One Hand+Spreading	1.41
19	Two Hands+Shaping	1.32
20	Two Hands+Unfolding	1.28

When analyzing Overall Affective State by topic (Figure 6), Neutral dominated most categories. In Romance/Marriage/Parenting, however, Interest (10.9%) and Happiness (6.0%) co-occurred with a wider range of negative affects such as Sadness and Anger. Positive affects were more prominent in Hobbies/Entertainment (Interest 12.6%, Happiness 6.1%) and Fashion/Beauty (Interest 8.3%, Happiness 17.4%), whereas negative affects were relatively higher in Family/School/Interpersonal Relations (Sadness 2.6%, Anger 1.3%) and Health/Self-management/Stress (Sadness 4.5%, Anger 0.8%), reflecting conflict and stress. Anger was notably more frequent in Sports/Leisure (16.5%) and Travel (19.8%) than in other topics[6].

4.4.2. Cross-Layer Analyses: Multimodal Integration

K-MIND's structure enables both independent and aligned analyses across channels. When verbal intents were paired with prosodic emphasis (Figure 7), paraverbal functions showed clear intent-specific patterns. Speaker-driven contributions—Elaboration Subjective, Elaboration Objective, and Recall—relied on Amplification highlighting the role of prosody in clarifying information and

[6]Normalized test results (per 1,000 utterances) also revealed clear patterns by gender and age. Women expressed Surprise (2.59×), Fear (1.94×), Happiness (1.43×), and Sadness (1.37×) more frequently than men, but used Humor less (0.32×). Age-related tendencies showed that teenagers displayed more Happiness, young adults more Interest, and older groups more Neutral.

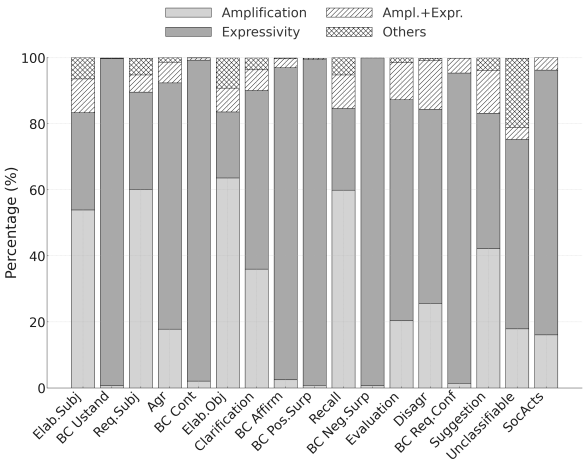

Figure 7: Paraverbal Function Distribution by Verbal Intent

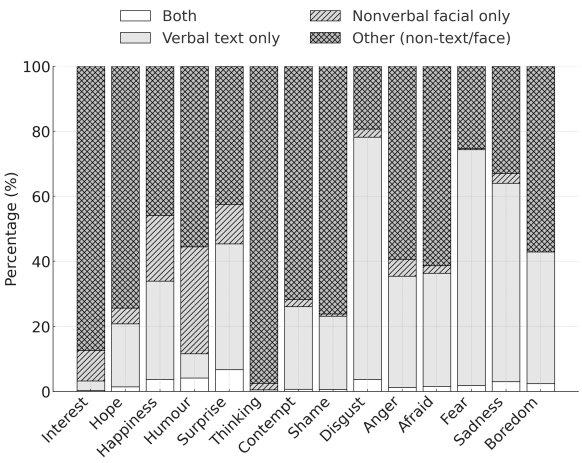

Figure 8: Distribution of Overall Affective State-Verbal Text Emotion-Nonverbal Facial Emotion Combinations

asserting stance. By contrast, listener-oriented intents—backchannel subtypes including BC Continuer, BC Understanding, BC Negative/Positive Surprise, BC Request Confirmation, and BC Affirmative—were marked almost exclusively by Expressivity. Mixed profiles appeared in Clarification, Suggestion, Disagreement, and Agreement. Overall, prosodic cues align systematically with interactional roles: speakers use amplification to deliver and highlight content, while listeners employ expressivity to signal attunement and sustain coordination.

Figure 8 shows the distribution of overall affective state–verbal text emotion–nonverbal facial emotion combinations, normalized per state (Neutral excluded). Emotional cues are broadly distributed across channels, often non-overlapping and incongruent. Thinking is most strongly associated with Other (97.5%), followed by Interest (87.4%), Shame (76.3%), and Hope (74.4%). Disgust (74.5%) and Fear (72.5%) are predominantly verbal, whereas Humour shows a relatively high proportion of facial-only signals (32.9%). Cross-channel co-occurrence (Both) remains low overall, peaking at Surprise (6.7%). These findings indicate that while Other (non-text/face) channels dominate affective signaling, verbal, paraverbal, and nonverbal modalities function in complementary ways across different affective states.

5. Discussion

Regarding verbal intents, speakers primarily shared personal thoughts, opinions, and experiences, reflecting the everyday and socially oriented nature of the topics. Listeners, in turn, responded empathically, sustained the conversational flow through questions, and verified information. These interactional dynamics—shaped by subjective sharing and active listening—align with prior findings that everyday talk serves primarily affective and relational rather than factual functions (Eggins and Slade, 2004; Coupland, 2014).

Building on these findings, the verbal behaviors captured in this dataset open up opportunities for both verbal and applied studies. A finer categorization of Request Subjective utterances (Closed vs. Open) may reveal which question types most effectively sustain conversational flow. In addition, analyzing less-examined phenomena such as Reflection, Recall, and Completion can highlight subtle patterns of language use that are crucial for modeling human-like conversational AI.

Moving to paraverbal behaviors, this layer provides a useful basis for examining how prosodic cues vary across topics. Deviations from neutral prosody not only signal emotional involvement but also highlight semantically salient content. At a broader level, the contrast between Amplification in affective domains (e.g., personal relationships) and Expressivity in task-oriented domains (e.g., work, career) reveals how cultural values and social norms shape conversational styles. Such contrasts warrant further analysis of generational, gender, and cultural variation. From a psychological perspective, contexts characterized by Expressivity illuminate when speakers foreground emotion over information. From an applied perspective, topic-sensitive paraverbal patterns can guide conversational AI, and integrating paraverbal with nonverbal cues enables deeper insight into multimodal coordination and turn-taking.

With respect to nonverbal behaviors, interaction-related gestures—such as Attitudinal, Responsive, and Focusing—were dominant, while Representational and Pointing gestures were less frequent. This suggests that gestures primarily served coor-

dination rather than lexical concretization. Head and hand movements were the main expressive channels, whereas facial gestures, though less frequent, specialized in conveying emotional nuances. These patterns invite further research into how gestures shape turn-taking (Kendon, 1967; Stivers et al., 2009; Levinson and Torreira, 2015) and how the use of specific body parts varies across cultures, genders, and topics. Despite their smaller proportion, facial gestures act as crucial affective signals, allowing distinctions such as joyful vs. social smiles (Ekman and Friesen, 2015; Niedenthal et al., 2010) and revealing context-dependent meanings. The differentiated use of gestural channels by speakers (e.g., hands for amplification, head for stance) and listeners (e.g., nods, facial expressions for empathy) provides valuable cues for designing multimodal dialogue systems capable of role-based gestural responses.

Concerning the overall affective state, the dataset primarily captures stable, everyday interactions—reflecting ordinary rather than conflict-driven conversations. Nevertheless, topic-specific affective spectra reveal distinct emotional patterns, offering a foundation for studying how empathy and positivity are organized in neutral and supportive exchanges, and for contrasting them with dialogues involving conflict or extreme emotion.

Finally, cross-layer analyses demonstrate systematic alignment between verbal and paraverbal cues: speaker-driven intents were typically associated with Amplification, whereas listener-oriented backchannels were marked by Expressivity. This suggests that prosody reinforces interactional roles by distinguishing content delivery from affective attunement. Importantly, the integration of Overall Affective State, Verbal Emotion, and Nonverbal Emotion across time-overlapping segments shows that multiple modalities operate in complementary ways. Together, these findings underscore that real-life language use relies on multimodal interplay rather than single-channel expression.

These insights have direct implications for conversational AI. Multimodal agents should model interactional roles, enabling speaker-oriented amplification and listener-oriented backchanneling as distinct behaviors. Because affective meaning is distributed across modalities without full temporal overlap, unimodal emotion recognition is inherently limited; multimodal integration is therefore essential for detecting implicit and relational signals in everyday interaction.

Beyond these observations, the K-MIND corpus provides a foundation for advancing multimodal and socially grounded research. Rather than a static dataset, it serves as a platform for exploring how language, emotion, and embodiment interact across communicative layers. Future work may extend these findings by examining cross-linguistic and cross-cultural variation, longitudinal shifts in communication styles, and multimodal modeling for adaptive conversational AI. Through its scale, diversity, and fine-grained alignment, K-MIND contributes to both the theoretical understanding of human interaction and the development of affect-aware, culturally adaptive dialogue systems.

6. Conclusions

The Korean Multimodal INteraction Data (K-MIND) provides a large-scale, naturalistic corpus that captures the multilayered nature of human interaction across verbal, paraverbal, and nonverbal dimensions. By offering synchronized multimodal signals with systematic annotation, it establishes a solid foundation for empirical research on communication and affective dynamics in Korean. Beyond its linguistic value, K-MIND supports interdisciplinary exploration in multimodal human–computer interaction, social cognition, and AI-driven dialogue modeling. Future extensions—such as the planned Triadic K-MIND for multiparty interaction—will enable investigation of more complex dynamics including turn management, role negotiation, topic transitions, and cross-cultural variation, thereby broadening both the scope and the societal relevance of this resource.

Limitations

While K-MIND provides a large-scale, multimodal view of Korean interaction, it has certain limitations. The recordings are confined to dyadic settings and may not fully capture broader social dynamics. Camera presence might also have influenced participants' behavior, leading to slight deviations from their usual conversational style.

Ethics Statement

All participants provided informed consent and were briefed on the study's purpose and data use. Personally identifiable information was anonymized, and video data are shared only for approved academic research. The collection and release of K-MIND comply with ethical standards and the Korean Personal Information Protection Act (PIPA).

Acknowledgements

This work was supported by an IITP grant (No. RS-2022-II220608) funded by the Ministry of Science and ICT (MSIT), Republic of Korea, and by the Signature Project of the Basic Research Program (401C5B16) of the Korea Electronics Technology Institute.

7. Bibliographical References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Zhiyuan Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhoefer. 2025. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset.

Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. 2013. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282.

Pawel Bujnowski, AmirAli Zadeh, Kwan-Yee Lee, Louis-Philippe Morency, and Taehwan Kim. 2024. Samsemo: A multilingual multimodal dataset for emotion recognition. In *Proceedings of the Language Resources and Evaluation Conference (LREC-COLING 2024)*, Torino, Italy. European Language Resources Association (ELRA).

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. Callhome american english speech. *Linguistic Data Consortium*.

Dorothy M. Chun. 2002. *Discourse intonation in L2: from theory and research to practice*. J. Benjamins.

H.H. Clark and S.E. Haviland. 1975. *Comprehension and the Given-new Contract*. Social sciences working paper. School of Social Sciences, University of California, Irvine.

Justine Coupland. 2014. *Small talk*. Routledge.

S. Eggins and D. Slade. 2004. *Analysing Casual Conversation*. Equinox Textbooks and Surveys in Linguistics. University of Toronto Press.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman and Wallace V. Friesen. 2015. Felt, false, and miserable smiles. *Nonverbal Behavior*, 6:238–252.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*, pages 835–838. ACM.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*, pages 1459–1462. ACM.

Eric N Forsythand and Craig H Martell. 2007. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Koichiro Ito, Masaki Murata, Tomohiro Ohno, and Shigeki Matsubara. 2020. Relation between degree of empathy for narrative speech and type of responsive utterance in attentive listening. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 696–701.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Haeyeon Kim. 2002. Collaborative turn completion in korean conversation. *Language Research*, pages 1,281–1,316.

Hansaem Kim. 2015. A research on non-verbal behavior of korean: Focused on gestures accompanied by interjection [in korean]. *Journal of Bangyo Language and Literature*, 65:5–28.

Jinsung Kim et al. 2020. Kemdy20: Korean multimodal emotion dialog dataset.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Petri Laukka, Patrik N. Juslin, and Roberto Bresin. 2005. A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, 19(5):633–653.

Hyewon Lee. 2016. A study on empathic listener response strategies for promoting communication: focusing on korean language learners.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, Volume 6 - 2015.

Andrew David Marshall, Paul L Rosin, Jason Vandeventer, and Andrew Aubrey. 2015. 4d cardiff conversation database (4d ccdb): A 4d database of natural, dyadic conversations. *Auditory-Visual Speech Processing,{AVSP} 2015*, pages 157–162.

James R Martin and Peter R White. 2003. *The language of evaluation*, volume 2. Springer.

Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology.* the MIT Press.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Paula M. Niedenthal, Martial Mermillod, Marcus Maringer, and Ursula Hess. 2010. The simulation of smiles (sims) model: Embodied simulation and the meaning of facial expression. *Behavioral and Brain Sciences*, 33(6):417–433.

Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2013. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1):19–28.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *DHN postproceedings*, pages 134–144.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Steve Renals, Thomas Hain, and Hervé Bourlard. 2007. Recognition and understanding of meetings the ami and amida projects. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 238–247. IEEE.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell and Lisa Feldman Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. 2025. Gaze-lle: Gaze target estimation via large-scale learned encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28874–28884.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue and Discourse*, 9(1):1–49. Publisher Copyright: © 2018 Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, Joelle Pineau.

Pranaydeep Singh, Luna De Bruyne, Orphée De Clercq, and Els Lefever. 2023. Misery loves complexity: Exploring linguistic complexity in the context of emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12871–12880. Association for Computational Linguistics.

Hyunjung Son. 2012. Construction of a Korean multimodal corpus. *Language Facts and Perspectives*, 30:55–79.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Ning Tan, Gaëlle Ferré, Marion Tellier, Edlira Cela, Mary-Annick Morel, Jean-Claude Martin, and Philippe Blache. 2010. Multi-level annotations of nonverbal behaviors in french spontaneous conversation. In *International Conference for Language Resources and Evaluation*, pages 74–79.

A. Annotation Tool

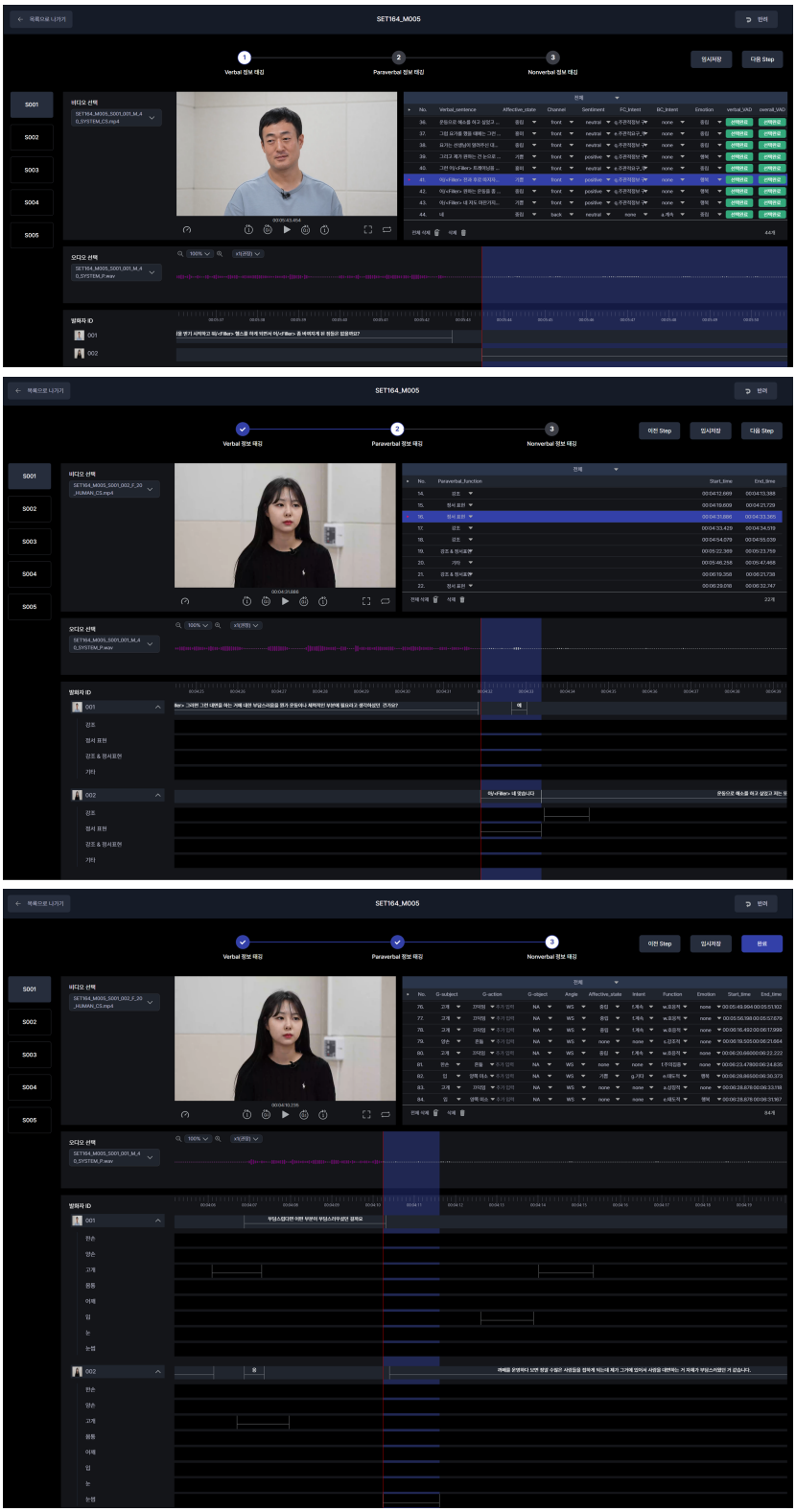

Figure 9: K-MIND annotation tool interface: (top) verbal, (middle) paraverbal, and (bottom) nonverbal.