Multimodal Classification of Co-speech Gesture Pragmatic Function in Storytelling

Jinqian Zhang
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
harushinken@jaist.ac.jp

Sixia Li
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
lisixia@jaist.ac.jp

Candy Olivia Mawalim
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
candylim@jaist.ac.jp

Shogo Okada*
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
okada-s@jaist.ac.jp

Abstract

Gestures, as essential co-speech behaviors in human communication, carry rich pragmatic functions. Accurately recognizing these functions could enhance an agent's ability to understand communicative behavior. In spontaneous storytelling scenarios—unlike lab-controlled settings—gesture functions exhibit high variability and are strongly influenced by individual speaker differences, making it difficult for unimodal systems to reliably capture their pragmatic intent. To address this challenge, we collected and annotated naturally occurring co-speech gestures in narrative dialogues, assigning each gesture one of six pragmatic function labels. We further propose a multimodal sequential classification model that encodes skeletal motion, acoustic prosody, and facial dynamics using separate bidirectional LSTM networks. These modality-specific encodings are fused via cross-modal attention and a gated mechanism to capture temporal dependencies and complementary information across modalities. Experimental results demonstrate that our tri-modal system achieves 62.5% accuracy and a weighted F1 score of 0.62 on the six-way classification task, outperforming uni-modal and bi-modal baselines by 3–11%. Ablation analysis reveals that skeletal features provide the most discriminative power for the majority of gesture functions, acoustic features are critical for specific categories, and facial features—though weak in isolation—substantially enhance overall performance when integrated via attention.

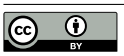
CCS Concepts

• **Human-centered computing** → **Gestural input**; *Human computer interaction (HCI)*; • **Computing methodologies** → *Supervised learning by classification*; Neural networks.

Keywords

Multimodal interaction, Gesture recognition, Storytelling, Machine learning, Human–agent interaction

*Corresponding author



HAI '25, Yokohama, Japan

ACM ISBN 979-8-4007-2178-6/25/11
https://doi.org/10.1145/3765766.3765771

ACM Reference Format:
Jinqian Zhang, Sixia Li, Candy Olivia Mawalim, and Shogo Okada. 2025. Multimodal Classification of Co-speech Gesture Pragmatic Function in Storytelling. In *13th International Conference on Human-Agent Interaction (HAI '25), November 10–13, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3765766.3765771

1 Introduction

Beyond spoken language, nonverbal signals play a crucial role in human communication [15]. These signals include gestures, facial expressions, eye gaze, and most notably, co-speech gestures that accompany speech. Unlike command gestures used in user interfaces, co-speech gestures serve pragmatic functions: they convey semantic content, organize discourse, and regulate interaction [28, 45].

This research addresses two core questions: (1) how could we automatically recognize and classify the pragmatic functions of co-speech gestures in natural storytelling dialogues, and (2) how could we quantify the contributions of different modalities to this classification, in order to empower socially aware agents and enhance the naturalness of interaction?

Automatic classification of co-speech gesture functions in natural settings faces two major challenges. First, the mapping between gestures and functions is highly variable: even the same speaker may produce different gestures for the same content, making it difficult to infer intent based solely on skeletal trajectories or hand shapes. Second, different speakers have varying gesture frequencies, styles, and durations, especially for spontaneous gestures produced intentionally or unintentionally while speaking. This requires the model to have a higher degree of generalizability. [18]. Although gesture generation and interaction have been actively explored in the domain of virtual agents [45], conventional computational gesture studies primarily focus on form-based recognition or command mapping [9, 36], often relying on lab-induced, pre-defined gesture tasks [36]. While Okada et al. [31] have used multimodal features for modeling gesture functions, systematized, fine-grained classification in natural interactions remains underexplored.

In real-world human–agent interactions, fine-grained function recognition is highly valuable. For example, an educational robot that detects "emphasis" gestures could highlight key text, while a customer service agent identifying "hesitation" gestures could adjust its response timing to improve user experience. The ability to

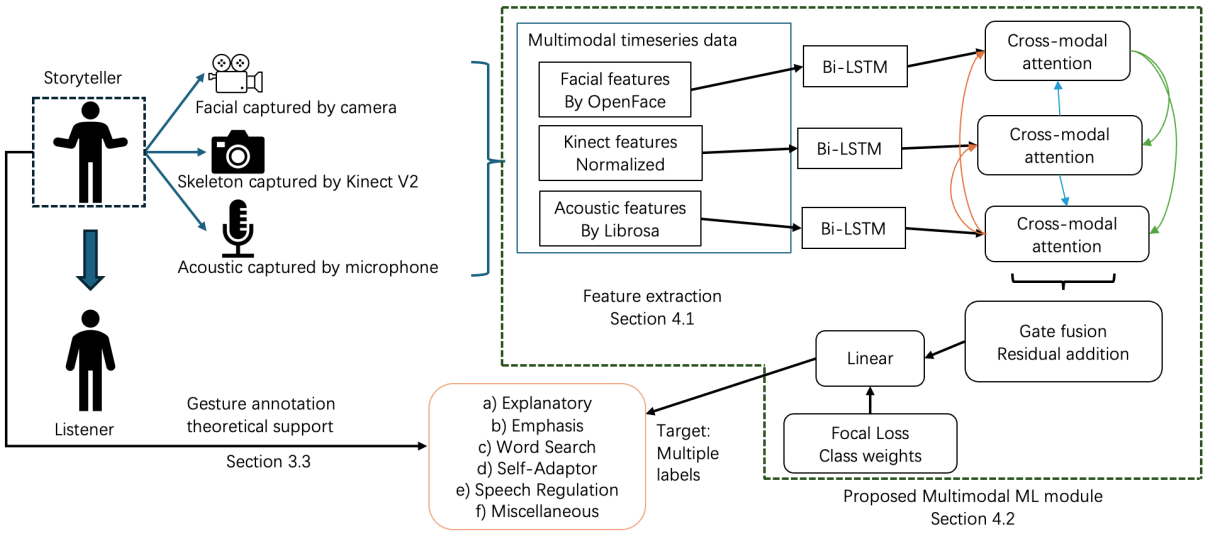


Figure 1: Overview of the proposed framework: We collected spontaneously produced gestures from speakers during storytelling and annotated them based on theories. Facial, skeletal, and acoustic features from participants are used as model inputs, with multi-class pragmatic gesture function labels as outputs.

accurately recognize a wider range of pragmatic gesture functions is essential for building socially intelligent agents capable of natural interaction [40].

According to McNeill [28], co-speech gestures are particularly frequent in storytelling contexts. We collected a storytelling dialogue corpus and annotated each gesture with one of six pragmatic function labels: Explanatory, Emphasis, Word Search, Self-Adaptor, Speech Regulation, and Miscellaneous. As gesture functions are not solely expressed through hand movements but are also deeply influenced by prosody, facial expressions, and head movements, uni-modal approaches often fail to capture their full meaning [12]. Recent advances in multimodal machine learning have significantly improved the analysis of fused audio, visual, and physiological signals [2, 41]. To this end, we propose a tri-modal classification framework based on Bi-LSTM encoders, cross-modal attention, and gated fusion to comprehensively predict six pragmatic gesture functions. Figure 1 illustrates our overall experimental pipeline and the general architecture of our model.

Our main contributions are as follows.

(1) We propose a new six-way taxonomy of gesture pragmatic functions based on existing theories and annotate a newly collected storytelling dataset accordingly;
(2) We design and evaluate a multimodal classification model integrating skeleton motion, speech prosody, and facial information;
(3) We conduct systematic ablation studies to quantify the contributions of each modality and the attention mechanism, offering insights for multimodal human-agent interaction.

In the future, interactive agents capable of understanding and generating gestures with appropriate pragmatic functions would significantly improve the fluency and credibility of human–machine dialogues. This work lays a solid foundation toward that goal [30].

2 Related Work

In the fields of social science and linguistics, a vast body of research has examined the role of gestures in human communication. McNeill [28] proposed a method for analyzing discourse gestures and investigated the relationship between spoken utterances and hand movements. He argued that gestures serve pragmatic functions and tend to appear more frequently in certain communicative contexts -particularly in "storytelling tasks "where a speaker narrates a story to a naive listener. Our research also adopts this experimental paradigm to elicit gesture data. Kendon [17] further introduced a four-phase model of gesture (preparation, stroke, hold, and retraction), identifying the stroke phase as the most information-rich segment. In our research, we similarly segment gestures and focus our analysis on the stroke phase.

Co-speech gestures serve multiple pragmatic functions, including conveying semantic content, structuring discourse, and regulating interaction. McNeill's classical taxonomy categorized gestures into iconic, deictic, metaphoric, and beat types, each associated with distinct functions [28]. Recent studies have further expanded this framework by emphasizing the tight coupling among gesture, speech prosody, discourse context, and speaker intent [12, 39].

Traditional computational gesture recognition has primarily focused on analyzing formal features such as hand shape and trajectory. However, recent research has increasingly shifted toward a function-oriented perspective. Kopp et al.[20] proposed a framework for modeling gesture functions in human–computer interaction and demonstrated that function-aware models outperform simple motion-matching approaches. Kelly and Tran[16] examined the role of co-speech gestures in conveying emotion and pragmatic intent, underscoring the importance of precise functional annotation. The GeSTICS corpus [14] offers fine-grained, function-level

annotations for gestures, reflecting a growing academic interest in functional modeling.

Accurately recognizing the pragmatic functions of gestures is critical for developing socially intelligent interactive agents. For example, Voss et al.[44] proposed an augmented co-speech gesture generation system that embeds pragmatic meaning into synthesized gestures, leading to more natural interactions. Robots capable of identifying turn-taking and emphasis gestures demonstrate smoother interactions and higher user satisfaction[40]. These findings highlight the importance of incorporating both multimodal and functional perspectives in gesture analysis to enable more effective human–machine interaction.

Given the inherently multimodal nature of gestures, analyzing their pragmatic functions requires integrating speech and facial signals. Kucherenko et al.[22] showed that combining gesture with prosodic and visual features significantly improves function prediction accuracy. Pelachaud et al.[37] similarly argued that socially interactive agents require integrated models of gesture, voice, and facial dynamics to accurately infer user intent and emotion. Imamura et al.[13] investigated the synergetic functional spectrum of head movements and facial expressions in conversation; Otsuchi et al.[35] analyzed the timing of impression formation based on head movement and linguistic features. Additional studies explored modeling the functions of head motion, such as predicting subjective impressions from head movements [34], recognizing functional head gestures using inflation–deflation networks [42], and capturing head interaction signals in multiparty conversation via deep transfer learning [29]. These results demonstrate that head dynamics also provides rich pragmatic information.

Unlike this extensive line of research on modeling head motion functions, our work focuses on classifying the pragmatic functions of co-speech hand gestures using a multimodal modeling approach. Previous work by Okada et al.[31] explored inferring lexical content from spontaneous gestures, and[32] investigated whether gestures convey information based on multimodal features of both the speaker and the listener. More recent multimodal research has confirmed that combining skeletal motion, facial expressions, and acoustic features enhances the detection of communicative cues such as hesitation, engagement, and emphasis [33].

In contrast to studies that define a fixed inventory or staged structure of gestures, we focus on naturally occurring co-speech gestures in narrative contexts. Our goal is to automatically classify the pragmatic functions of spontaneous gestures using multimodal information. To address the variability in motion–function mapping, we incorporate additional modalities such as facial expression and gaze to extract shared features. To mitigate speaker-specific variability, we employ normalization techniques (e.g., body-size normalization) and constrain the task to "content-specific explanation" to elicit more consistent gesture patterns. Successful classification of gesture pragmatic functions not only aids intelligent agents in understanding the speaker's cognitive state and communicative behavior but also lays the foundation for generating gestures with appropriate pragmatic functions in future systems.

3 Dataset and Annotation

3.1 Data Collection

We constructed a multimodal storytelling corpus based on McNeill's narrative-recount paradigm [28]. A total of 37 native Japanese speakers (6 males and 31 females) participated in a video explanation task. Each participant first watched two approximately 10-minute animated shorts, "Tweety–Canary Row" and "Tweet–S.O.S.", and was allowed to take brief notes. They then recounted the story content to a listener who had not seen the videos. The listener was allowed to ask questions at any time, which encouraged the natural production of co-speech gestures by the storyteller. We recorded 72 sessions in total, each lasting approximately 10 minutes, and simultaneously collected data using a Kinect V2 sensor (skeleton + depth), a directional microphone (audio), and a high-resolution video camera (facial/head). The participants were between 20 and 40 years old. Figure 2 illustrates the overall data collection procedure.

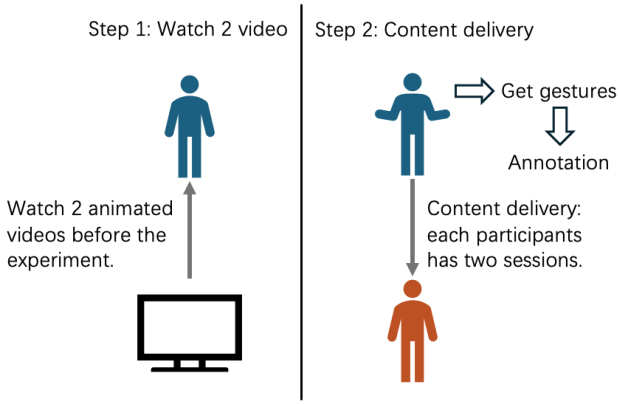


Figure 2: Storytelling task for data collection

3.2 Gesture Unit Segmentation

We adopt Kendon's gesture model[17] for segmentation. In Kendon's model, the neutral "rest position" surrounds each Gesture Unit (G-unit), defined as the movement interval between successive rests. A G-unit could further be decomposed into up to five Gesture Phases:

1. Preparation: hand moves from rest toward the gesture start.
2. Pre-stroke hold: brief hold signaling stroke onset.
3. Stroke: the core movement conveying meaning.
4. Post-stroke hold: brief hold at stroke end.
5. Retraction: hand returns to rest.

Not all phases appear in every gesture; some phases may be omitted or re-ordered, and "Hold" phases could prolong shape for emphasis[18]. We automatically detect rest-to-rest intervals via velocity thresholds on skeletal keypoints and then manually refine stroke boundaries and holds.

3.3 Functional Tagset and Theoretical Support

Building on [38], this research assigns functional labels to gestures occurring in communicative contexts. According to prior literature, gesture functions could be broadly categorized into two main types:

- **Content-Expressive Functions:** Gestures that convey semantic content or supplement the spoken message—for example, mimicking a pitching or batting motion when describing a baseball scene.
- **Interactional-Regulatory Functions:** Gestures that regulate the flow of interaction, such as managing turn-taking or providing listener feedback—for example, touching one's head during hesitation or extending a palm to fill a pause.

Based on this theoretical framework, we define six fine-grained pragmatic gesture-function labels tailored to storytelling contexts. Each label is grounded in gesture pragmatics literature and supported by corresponding research, Figure 3 shows examples of pragmatic functions of co-speech gestures.

(1) **Explanatory**
Gestures that depict scene elements, actions, or relationships to support listener understanding. Co-speech gestures could enhance comprehension by representing and highlighting important content in speech [4, 7, 11].
(2) **Emphasis**
Emphasis gestures consist of rhythmic punctuations aligned with speech prosody to reinforce semantic prominence; such functions are described in Kendon's foundational work [17] and elaborated by McClave [26].
(3) **Word Search**
Hesitation or deliberation gestures appearing during lexical retrieval (e.g., hand-to-chin poking). These "hesitation gestures" have been characterized by Kita et al. [19] and recognized as signals of planning difficulty in speech production.
(4) **Self-Adaptor**
Self-touch actions (e.g., scratching head, rubbing face) that regulate emotional or cognitive states without conveying semantic content. Kita et al. [19] distinguish these from communicative gestures, and Lausberg's NEUROGES framework treats self-touch as a regulator category [23].
(5) **Speech Regulation**
Gestures that manage turn-taking or fill pauses (e.g., open-palm "yield" gesture). Kendon [17] discusses their role in discourse coordination.
(6) **Miscellaneous**
Other infrequent or context-specific movements not covered by the above categories.

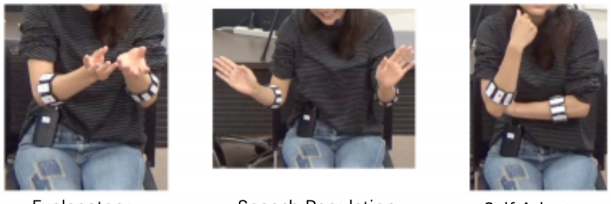


Figure 3: The example of gesture pragmatic function

Each Stroke (the core meaningful movement phase) of a Gesture Unit is assigned exactly one primary label. Overlapping beat strokes may receive concurrent Emphasis tags when they serve a reinforcing function.

Table 1: Distribution of gesture by functional label (total: 14,897).

Functional Label	Count
Explanatory	7,475
Speech Regulation	3,737
Self-Adaptor	1,443
Word Search	1,282
Emphasis	534
Miscellaneous	426

3.4 Annotation Procedure

We ensured high-quality functional annotations through a three-stage protocol. First, in the training & pilot phase, two gesture-literate annotators familiarized themselves with our six-label taxonomy and the Kendon/Kita phase model [17, 19], then independently labeled a 10 % pilot subset; any disagreements were used to refine our annotation guidelines. Next, during annotation & cross-check, the remaining data were split between the annotators, who each labeled half of the stroke segments before swapping their annotations and resolving ambiguities through mutual review. Finally, in the adjudication & spot-check phase, we jointly resolved the fewer than 5 % of segments still in dispute and then randomly spot-checked another 10 % of all labels to verify consistency against the finalized guidelines. We ultimately annotated a total of 14,897 gesture pragmatic function labels across the entire dataset, Table 1 summarizes the total number of annotated gesture segments per functional category.

4 Methodology

Our framework ingests three synchronized streams—Kinect skeletal motion, acoustic prosody, and facial/head dynamics—and processes each into fixed-length feature tensors. All modalities are aligned to Kinect's 8-digit millisecond timestamps; for acoustic and OpenFace sequences lacking native millisecond marks, we infer equivalent stamps by matching keyframes to the nearest Kinect timepoint.

4.1 Feature Extraction

In our framework, to enable the model to accurately capture the pragmatic functions of co-speech gestures, we perform dedicated preprocessing and feature extraction for three key modalities: skeletal motion, speech prosody, and facial dynamics. Table 2 provides an overview of the extraction procedures, tools used, and final output dimensions for each modality. The following sections detail the specific processing methods for each feature stream.

4.1.1 Skeletal Motion Features. We obtain 3D joint coordinates from Kinect V2. Kinect provides joint position data in a right-handed Cartesian coordinate system centered at the sensor's optical origin, where the forward direction is the Z_+ axis, the rightward direction is the X_+ axis, and the upward direction is the Y_+ axis. Therefore, the 3D positions of joints precisely reflect the spatial

Table 2: Overview of modality-specific feature sets

Modality	Preprocessing	Features Extracted	Output Shape
Skeletal motion	Kinect V2; select 9 joints; center on shoulders; Z-score std.	3D coordinates of 9 joints (X, Y, Z)	$(T_{\text{Kinect}}, 27)$
Acoustic	16 kHz; 33 ms hop, 50 ms window via Librosa	RMS energy; F_0 (YIN); 13 MFCCs	$(T_{\text{audio}}, 15)$
Facial & head	OpenFace 2.0 extraction	68 landmarks; 17 AUs; 3D head pose; gaze; HOG	$(T_{\text{Facial}}, 713)$

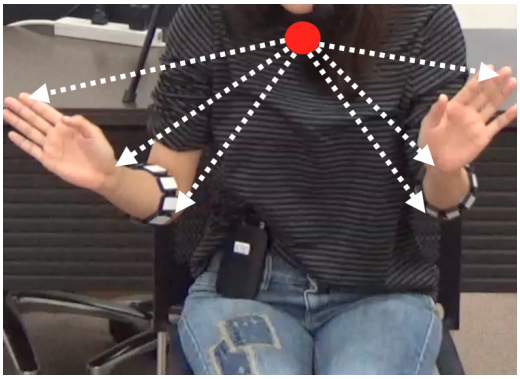


Figure 4: Overview example of processing skeletal motion features

distribution of the human body in the real world, rather than merely representing a projection onto the image plane.

To mitigate coordinate discrepancies caused by sensor placement and individual differences, we first perform joint selection and centering. Specifically, we retain nine keypoints on each side (Head, Shoulder, Elbow, Wrist, Hand) and use the midpoint of the shoulders as the origin to establish a person-centric reference frame.

Next, we apply Z-score normalization independently to each spatial axis (X, Y, Z) by subtracting the session-wide mean and dividing by the standard deviation. This standardization stabilizes the distribution of time-series features and facilitates downstream learning [8]. Figure 4 illustrates an overview of our preprocessing pipeline. The final output is a tensor of shape $(T_{\text{Kinect}}, 27)$, where T_{Kinect} denotes the total number of Kinect frames, and the 27 channels correspond to the 3D coordinates of 9 selected joints.

4.1.2 Acoustic Features. We process speech signals at a 16 kHz sampling rate using the Librosa library [27]. Audio is framed with a 33 ms hop size (approximately 30 fps) and a 50 ms window. For each frame, we extract three types of low-level features: root-mean-square (RMS) energy, which reflects signal intensity variations; the fundamental frequency (F_0), estimated via the YIN algorithm within the 80–500 Hz range to capture the speaker's vocal-fold vibration rate [6]; and 13 Mel-Frequency Cepstral Coefficients ($\text{MFCC}_1, \ldots, \text{MFCC}_{13}$), which encode the spectral envelope of the audio [5]. As a result, each utterance is represented as a tensor of shape $(T_{\text{audio}}, 15)$, where T_{audio} is the number of frames and the 15 channels correspond to [energy, F_0, $\text{MFCC}_1, \ldots, \text{MFCC}_{13}$].

4.1.3 Facial Features. Facial and head dynamics are extracted using OpenFace 2.0 [3], which provides over 713 features per frame. These include 68 two-dimensional facial landmark coordinates, intensities and binary presence indicators for 17 facial Action Units, three-dimensional head pose (pitch, yaw, roll), eye gaze direction vectors, and HOG-based appearance descriptors alongside additional visual features. After concatenation, each video sequence is represented as a tensor of shape $(T_{\text{Facial}}, 713)$. Such high-dimensional facial behavior features have been demonstrated to enhance affective and gesture recognition performance [25].

4.2 Model Architecture

Our core classifier is a multimodal sequential model that jointly captures intra-modal temporal dynamics and inter-modal interactions. It comprises the following components:

(1) **Bidirectional LSTM Encoders**: Three independent bidirectional LSTM layers encode the acoustic, facial, and skeletal feature streams, capturing both forward and backward temporal dependencies [10]. For each modality $m \in \{\text{ac}, \text{of}, \text{ki}\}$, we apply a Bi-LSTM:

$$\mathbf{H}^m = \text{BiLSTM}_m(\mathbf{X}^m) \in \mathbb{R}^{T \times 2h},$$

where $\mathbf{X}^m \in \mathbb{R}^{T \times d_m}$ is the input feature sequence, T the sequence length, d_m input dim, and $h = 256$ the hidden size per direction [10].

(2) **Cross-Modal Attention Mechanism**: To allow each modality to incorporate complementary information from the others, we implement a cross-modal attention module. Specifically, we treat the LSTM output of one modality as the query, and the concatenated outputs of the other two modalities as the keys and values for computing scaled dot-product attention. This mechanism enables, for example, acoustic signals to dynamically attend to relevant facial and body movements [21, 43]. Let $\mathbf{H}^m \in \mathbb{R}^{B \times T \times 2h}$ be the bidirectional-LSTM output for modality m, and let $\bar{m}$ denote the other two modalities, whose outputs we concatenate as $\mathbf{H}^{\bar{m}} = [\mathbf{H}^i; \mathbf{H}^j] \in \mathbb{R}^{B \times T \times 2h}$. We first project into query, key, and value spaces:

$$\mathbf{Q}^m = \mathbf{H}^m \mathbf{W}^Q, \quad \mathbf{K}^{\bar{m}} = \mathbf{H}^{\bar{m}} \mathbf{W}^K, \quad \mathbf{V}^{\bar{m}} = \mathbf{H}^{\bar{m}} \mathbf{W}^V,$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{2h \times h}$. This yields $\mathbf{Q}^m, \mathbf{K}^{\bar{m}}, \mathbf{V}^{\bar{m}} \in \mathbb{R}^{B \times T \times h}$.

We compute attention scores with optional masking (to ignore padding):

$$\text{scores}_{b,t,t'} = \begin{cases} -\infty, & \text{if mask}_{b,t'} = 0, \\ \frac{\mathbf{Q}^m_{b,t} \mathbf{K}^{\bar{m}}_{b,t'}}{\sqrt{h}}, & \text{otherwise}, \end{cases}$$

then normalize:

$$\mathbf{A}^m = \text{softmax}(\text{scores}) \in \mathbb{R}^{B \times T \times T},$$

$$\text{attn}^m = \mathbf{A}^m \mathbf{V}^{\bar{m}} \in \mathbb{R}^{B \times T \times h}.$$

Finally, we project back to $2h$ dimensions and pool over time:

$$\text{Attn}^m = \text{attn}^m \, \mathbf{W}^{\text{out}}, \quad \mathbf{W}^{\text{out}} \in \mathbb{R}^{h \times 2h},$$

$$\mathbf{z}^m = \frac{1}{T} \sum_{t=1}^{T} \text{Attn}^m_{:,t,:} \in \mathbb{R}^{B \times 2h}.$$

This produces a fixed-size summary $\mathbf{z}^m$ for each modality, which is then fed into the gated fusion layer.

(3) **Focal Loss for Class Imbalance**: Our gesture-function dataset exhibits significant class imbalance (e.g., "Explanatory" far more frequent than "Emphasis" or "Miscellaneous"), which could cause standard cross-entropy to be dominated by the majority class. To mitigate this, we employ the Focal Loss proposed by Lin *et al.* [24]:

$$\text{FL}(p_t) \;=\; -\alpha_t \, (1 - p_t)^{\gamma} \; \log(p_t)$$

where p_t is the model's estimated probability for the true class, $\gamma > 0$ is a focusing parameter that down-weights well-classified examples, and $\alpha_t \in (0, 1)$ balances the importance of each class. By emphasizing hard, minority-class examples, Focal Loss helps the model learn features for under-represented gesture functions more effectively [24, 46].

(4) **Gated Fusion**: The three attention-weighted feature vectors are concatenated and passed through a sigmoid-activated gating layer, producing a weight vector for each modality. We split the weight vector into three parts, normalize each, and compute a weighted sum to obtain a fused representation that adaptively highlights the most informative modality for each sample [1]. We concatenate $\mathbf{z} = [\mathbf{z}^{\text{ac}}; \mathbf{z}^{\text{of}}; \mathbf{z}^{\text{ki}}] \in \mathbb{R}^{3h}$ and compute

$$\mathbf{g} = \sigma(\mathbf{W}_g \mathbf{z} + \mathbf{b}_g) \in (0, 1)^{3h}.$$

Split $\mathbf{g} = [\mathbf{g}^{\text{ac}}; \mathbf{g}^{\text{of}}; \mathbf{g}^{\text{ki}}]$ with each $\mathbf{g}^m \in \mathbb{R}^h$, then normalize

$$\tilde{\mathbf{g}}^m = \frac{\mathbf{g}^m}{\sum_{n \in \{\text{ac,of,ki}\}} \mathbf{g}^n + \epsilon}.$$

The fused feature is given by

$$\mathbf{z}^{\text{fuse}} = \sum_m \tilde{\mathbf{g}}^m \odot \mathbf{z}^m,$$

where $\odot$ denotes element-wise multiplication. This mechanism allows the model to adaptively adjust the contribution of each modality based on the input sample, highlighting the most informative modality.

(5) **Classification Layer**: The fused feature $\mathbf{z}^{\text{fuse}} \in \mathbb{R}^{2h}$ is fed into a single linear layer and softmax to predict the six gesture classes:

$$\hat{\mathbf{p}} = \text{softmax}\big(\mathbf{W}_c \, \mathbf{z}^{\text{fuse}} + \mathbf{b}_c\big) \quad \text{where } \mathbf{W}_c \in \mathbb{R}^{6 \times 2h}, \; \mathbf{b}_c \in \mathbb{R}^6.$$

The model's final prediction is the class k with the highest probability $\hat{p}_k$.

5 Experiments

5.1 Experimental Setup

Gesture pragmatic-function classification is cast as a six-way task (Explanatory, Emphasis, Word Search, Self-Adaptor, Speech Regulation, Miscellaneous). We used five stratified folds (7:1.5:1.5 split, seed=42) throughout.

All variants share hyperparameters: batch size 64, Bi-LSTM hidden 256, learning rate 1e−3, up to 80 epochs with early stopping (patience 40). Models are trained with Adam to minimize weighted focal loss ($\gamma = 2$). We select the best checkpoint by validation accuracy and report accuracy, precision, recall, and F1 on each test fold.

5.2 Ablation Research Design

To quantify the contribution of each modality and interaction mechanism, we implement the following variants, all trained under the same splits and hyperparameters:

(1) **Uni-Modal Baselines:**
- A: Acoustic only
- F: Facial only
- K: Kinect only

Each uses a single bidirectional LSTM plus global pooling for classification.

(2) **Bi-Modal Variants:**
- A + F: Acoustic + Facial
- A + K: Acoustic + Kinect
- F + K: Facial + Kinect

Each fuses two modalities via a single cross-modal attention module.

(3) **Tri-Modal Model (ALL):** Combines all three modalities with cross-modal attention for each stream, followed by a gated fusion layer.

(4) **Tri-Modal Without Attention:** To verify the effectiveness of the cross-modal attention mechanism and to measure how dependent various gesture functions are on inter-modal attention, we remove all cross-modal attention modules from the tri-modal model. Instead, we pass the acoustic, facial, and skeletal streams independently through Bi-LSTMs, apply average-pooling to each output, and simply concatenate the three pooled vectors before feeding them into the final classification layer.

Comparing these variants on validation and test sets allows us to assess the effectiveness of multimodal integration, attention mechanisms, and gated fusion in predicting gesture pragmatic functions.

6 Results and Discussion

6.1 Overview

In this section, we first present the overall performance of all model variants on the six-way gesture pragmatic-function classification task. We then analyze per-category performance to understand how each modality and their combinations contribute to recognizing specific gesture functions.

6.2 Overall Performance

Table 3 reports the mean accuracy and weighted F1 over five folds for all model variants. As expected, the full tri-modal system (A+F+K) achieves the best performance (62.47 % accuracy, 0.62 F1). When the cross-modal attention mechanism is removed (Tri-modal No-Attention), accuracy drops to 57.04 % and weighted F1 to 0.59

Table 3: Five-fold average performance (mean ± std.**) of all model variants.**

Model	Accuracy (%)	Weighted-F1
Tri-modal (A+F+K)	**62.47 ± 0.84**	**0.62 ± 0.01**
Tri-modal No-Attention	57.04 ± 0.85	0.59 ± 0.01
Bi-modal (A+K)	59.19 ± 1.57	0.59 ± 0.01
Bi-modal (F+K)	57.86 ± 1.66	0.57 ± 0.01
Bi-modal (A+F)	51.44 ± 1.63	0.46 ± 0.01
Uni-modal (K)	52.36 ± 0.88	0.54 ± 0.01
Uni-modal (A)	49.32 ± 1.81	0.50 ± 0.02
Uni-modal (F)	48.19 ± 1.58	0.45 ± 0.01

(±0.01), indicating a clear degradation. Among bi-modal pairs, Acoustic+Kinect (A+K) outperforms both Facial+Kinect (F+K) and Acoustic+Facial (A+F), showing that skeletal motion combined with prosodic cues is particularly powerful, while facial features add less when paired with body data. The large drop for A+F (51.44 % acc, 0.46 F1) demonstrates that face and audio alone fail to capture many gesture functions. In the uni-modal setting, Kinect alone (52.36 % acc, 0.54 F1) is strongest, followed by acoustic (49.32 % acc, 0.50 F1), and facial (48.19 % acc, 0.45 F1) being weakest. Overall, body motion is the single most informative channel, prosody contributes notably to certain classes, and facial cues—though useful in attention-based fusion—are insufficient in isolation.

The performance gap between the full tri-modal model and the No-Attention variant shows that cross-modal attention allows acoustic, facial, and skeletal features to be dynamically aligned at the temporal level. For example, hesitation in speech may be accompanied by subtle facial micro-expressions or preparatory skeletal movements. Without cross-modal attention, each modality's high-level representation never "sees" or aligns with information from the other two streams, preventing the model from exploiting complementary signals.

Table 4: Tri-modal model classification results by class (mean ± std.**).**

Class	Precision	Recall	F1-score
Explanatory	0.77 (± 0.02)	0.79 (± 0.04)	0.78 (± 0.01)
Word Search	0.42 (± 0.04)	0.49 (± 0.06)	0.45 (± 0.02)
Speech Regulation	0.55 (± 0.06)	0.46 (± 0.07)	0.49 (± 0.03)
Self-Adaptor	0.51 (± 0.05)	0.57 (± 0.06)	0.53 (± 0.02)
Emphasis	0.18 (± 0.06)	0.20 (± 0.10)	0.19 (± 0.08)
Miscellaneous	0.35 (± 0.05)	0.31 (± 0.04)	0.33 (± 0.04)

Table 4 presents the five-fold average precision, recall and F1-score for each gesture function using the full tri-modal model (A + F + K). Several observations emerge:

- **Explanatory**: Highest F1 of 0.78(±0.01) reflects that explanatory gestures are abundant and exhibit consistent prosodic and body-motion patterns.
- **Word Search**: F1 = 0.45(±0.02) indicates that hesitation cues in audio, supplemented by slight facial and skeletal signals, are captured moderately well.
- **Speech Regulation**: Recall (0.46±0.07) slightly lags precision (0.55±0.06), suggesting the model is conservative in predicting turn-taking gestures but does so accurately when it does.
- **Self-Adaptor**: Balanced performance (F1 = 0.53 ±0.02) shows body-centric actions like scratching are reliably distinguished, with facial subtleties and prosody providing minor disambiguation.
- **Emphasis**: Lowest F1 of 0.19 (±0.08) reveals that emphasis gestures are rare and exhibit high variability in modality signals, calling for more data or specialized features.
- **Miscellaneous**: F1 = 0.33 (±0.04) underlines that unstructured, diverse gestures remain the hardest to classify, however, these may contain new gesture pragmatic functions with underlying commonalities, suggesting that future work could benefit from refining and expanding the label categories.

Overall, our model demonstrates strong capability on well-represented classes, while performance on infrequent or highly variable gestures highlights directions for future data collection and feature refinement.

6.3 Per-Category Performance

Table 5 reports the five-fold average F1-scores (mean ± std) for each gesture function and each model variant. This allows us to see how modalities contribute differently across the six classes.

Explanatory Gestures. Explanatory gestures constitute the largest and most stable class in our dataset. All modalities distinguish this class well, with the tri-modal model achieving the highest weighted F1 score of 0.78 ± 0.01. Notably, combining skeletal data with any other modality—especially acoustic—yields nearly equivalent performance (e.g., bi-modal A+K F1 = 0.75 ± 0.01), indicating that body-motion features are most critical for recognizing explanatory gestures, while speakers also exhibit characteristic prosodic patterns during explanation.

Word Search Gestures. Word search gestures typically coincide with speaker hesitation and pitch fluctuations, making the acoustic modality the primary contributor. The acoustic-only model achieves the best F1 of 0.46 ± 0.05, with the tri-modal model close behind (0.45 ± 0.02). In contrast, combinations lacking acoustic information (e.g., Facial+Kinect) fail to capture these features. Interestingly, no bi-modal combination of A+K improves upon the acoustic-only baseline, yet the tri-modal model does, underscoring the importance of three-way synergy for this class.

Speech Regulation Gestures. Speech regulation gestures convey both prosodic and bodily signals. The tri-modal model achieves an F1 of 0.49 ± 0.03, and the bi-modal Acoustic+Kinect model attains 0.48 ± 0.02. Facial information alone or paired with acoustic performs poorly, indicating that these gestures are primarily driven by body motion and speech rhythm. Nevertheless, the attention mechanism in the tri-modal model leverages some facial cues to provide a modest boost.

Table 5: Five-fold Weighted F1 (mean ± std.**) by class and model.**

Model	Explanatory	Word Search	Speech Reg.	Self-Adaptor	Emphasis	Misc.
Tri-modal (A+F+K)	**0.78 ± 0.01**	0.45 ± 0.02	**0.49 ± 0.03**	0.53 ± 0.02	**0.19 ± 0.08**	0.33 ± 0.04
Tri-modal No-Attention	0.75 ± 0.02	**0.49 ± 0.02**	0.44 ± 0.07	0.47 ± 0.01	0.18 ± 0.03	**0.37 ± 0.04**
Bi-modal (A+K)	0.75 ± 0.01	0.34 ± 0.06	0.48 ± 0.02	**0.54 ± 0.05**	0.14 ± 0.05	0.28 ± 0.06
Bi-modal (F+K)	0.74 ± 0.01	0.19 ± 0.02	0.47 ± 0.03	0.54 ± 0.04	0.15 ± 0.05	0.29 ± 0.04
Bi-modal (A+F)	0.69 ± 0.01	0.00	0.44 ± 0.06	0.02 ± 0.04	0.07 ± 0.10	0.00
Uni-modal (K)	0.72 ± 0.02	0.12 ± 0.07	0.45 ± 0.02	0.49 ± 0.02	0.12 ± 0.04	0.24 ± 0.03
Uni-modal (A)	0.70 ± 0.02	0.46 ± 0.05	0.32 ± 0.06	0.12 ± 0.07	0.17 ± 0.02	0.23 ± 0.02
Uni-modal (F)	0.69 ± 0.02	0.03 ± 0.06	0.38 ± 0.07	0.01 ± 0.01	0.14 ± 0.03	0.00

Self-Adaptor Gestures. Self-adaptor gestures (e.g., scratching, touching the face) are almost entirely defined by body movements and bear no direct relation to explanatory content, making them difficult to characterize with prosodic or facial signals. A Kinect-only model already achieves 0.49 ± 0.02; fusing with facial or acoustic data improves this slightly to 0.54 ± 0.05 and 0.54 ± 0.04, respectively. The tri-modal model (0.53 ± 0.02) underperforms bi-modal combinations, perhaps due to increased complexity.

Emphasis Gestures. Emphasis gestures rely on subtle coordination of movement and prosody but suffer from data sparsity, making stable feature learning challenging. All models perform poorly (best acoustic-only F1 = 0.17 ± 0.02), reflecting that speakers may increase vocal intensity when emphasizing content. The tri-modal model provides only a marginal gain (0.19 ± 0.08), suggesting that additional data or specialized features are needed to capture emphasis intent.

Miscellaneous Gestures. The miscellaneous class encompasses diverse, unstructured movements, leading to uniformly low performance across all models (tri-modal F1 = 0.33 ± 0.04). Modalities lacking skeletal or acoustic information (e.g., Facial+Acoustic) completely fail to recognize these gestures. Future work might subdivide this class or design finer-grained features.

Without Cross-modal Attention. The model without cross-modal attention shows a marked decline in classification performance for "Explanatory," "Speech Regulation," "Self-Adaptor," and "Emphasis" compared to the full model, indicating that these pragmatic functions of gestures rely heavily on intermodal alignment and dynamic fusion. In contrast, the "Word Search" category—which depends more on a single modality (acoustic)—improves slightly in the no-attention model (0.49 ± 0.02). One possible explanation is that, without the attention module, the model is less disturbed by other modalities and thus captures hesitation cues in the audio more purely. The ablation results for the attention mechanism further demonstrate that temporal alignment and dynamic interaction between modalities are crucial for capturing gesture pragmatic functions in realistic conversational settings. They also indirectly suggest that, in natural communication, most gesture functions emerge from the interplay of multiple modalities.

Summary. Overall, skeletal (Kinect) data is essential for most gesture functions; acoustic cues aid in detecting hesitation and emphasis; facial features contribute minimally in isolation but could be leveraged via attention in the tri-modal model. Our results demonstrate that tri-modal fusion significantly improves pragmatic-function classification in realistic explanatory scenarios, yet rare classes ("Miscellaneous," "Emphasis") remain challenging.

6.4 Limitations

From the perspective of limitations, the current dataset suffers from class imbalance, which makes it challenging to model gesture functions for low-frequency categories. Although the model's overall performance on the 6-class classification task is not exceptionally high, we consider it a preliminarily acceptable result given the number of classes and the degree of data imbalance. Additionally, our study participants were predominantly female, which may limit the generalizability of our findings. The precise impact of this gender imbalance remains an open question, as existing research on the topic lacks conclusive evidence.

7 Conclusion

This research proposes a multimodal framework that extracts acoustic, facial, and skeletal features from a storytelling scenario to automatically classify the pragmatic functions of co-speech gestures in real-world communicative contexts. Experimental results show that the full tri-modal system achieves the highest overall accuracy and weighted F1 score, significantly outperforming uni-modal and bi-modal models. We found that skeletal motion is most crucial for the majority of gesture functions, acoustic features play a leading role in identifying hesitation and emphasis gestures, and facial features—while weak in isolation—serve as effective auxiliary signals within the attention mechanism. Furthermore, the performance improvement brought by the cross-modal attention mechanism indicates that the pragmatic functions of gestures in natural communication rely on the interaction between multiple modalities. Future work would focus on further expanding and diversifying the dataset to mitigate the issue of label imbalance, as well as incorporating richer dialogue context and stronger temporal modeling mechanisms to enhance the model's ability to understand the pragmatic functions of co-speech gestures.

Acknowledgments

This work was partially supported by JSPS KAKENHI (22H00536, 23H03506), JST Moonshot R&D program (JPMJMS2031), and JST, CRONOS (JPMJCS24K7).

References

[1] Juan Arévalo, Thamar Solorio, Mario Montes-y Gómez, and Francisco A. González. 2017. Gated Multimodal Units for Information Fusion. arXiv preprint arXiv:1702.01992.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443. doi:10.1109/TPAMI.2018.2798607

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[4] Eliza L. Congdon, Miriam A. Novack, Neon Brooks, Naureen Hemani-Lopez, Lucy O'Keefe, and Susan Goldin-Meadow. 2017. Better together: Simultaneous presentation of speech and gesture in math instruction supports generalization and retention. *Learning and Instruction* 50 (2017), 65–74. doi:10.1016/j.learninstruc.2017.03.005

[5] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 4 (1980), 357–366.

[6] Alain de Cheveigné and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111, 4 (2002), 1917–1930.

[7] Jan P. de Ruiter. 2000. The production of gesture and speech. In *Language and Gesture*, David McNeill (Ed.). Cambridge University Press, Cambridge, UK, 284–311.

[8] Michelle B. Del Rosario, Stephen J. Redmond, and Nigel H. Lovell. 2015. Tracking the evolution of smartphone sensing for monitoring human movement. *Sensors* 15, 8 (2015), 18901–18933.

[9] Sergio Escalera, Vassilis Athitsos, and Isabelle Guyon. 2017. Challenges in multimodal gesture recognition. In *Gesture Recognition*, Isabelle Guyon, Vassilis Athitsos, and Sergio Escalera (Eds.). Springer, Cham, Switzerland, 1–60. Lecture Notes in Computer Science, vol. 10275.

[10] Alex Graves and Alex Graves. 2012. Long Short-Term Memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 37–45.

[11] Angela Grimminger, Katharina J. Rohlfing, and Prisca Stenneken. 2010. Children's lexical skills and task demands affect gestural behavior in mothers of late-talking children and children with typical language development. *Gesture* 10, 2–3 (2010), 251–278. doi:10.1075/gest.10.2-3.07gri

[12] Judith Holler and Stephen C. Levinson. 2019. Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences* 23, 8 (2019), 639–652. doi:10.1016/j.tics.2019.04.009

[13] Mai Imamura, Ayane Tashiro, Shiro Kumano, and Kazuhiro Otsuka. 2023. Analyzing Synergetic Functional Spectrum from Head Movements and Facial Expressions in Conversations. In *Proceedings of ICMI*. 42–50.

[14] Gedion Y. Kebe, Mehmet D. Birlikci, Antoine Boudin, Ryo Ishii, Jean-Marc Girard, and Louis-Philippe Morency. 2024. GeSTICS: A Multimodal Corpus for Studying Gesture Synthesis in Two-party Interactions with Contextualized Speech. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. ACM, 1–10. doi:10.1145/3652988.3673917

[15] Spencer D. Kelly, Dale J. Barr, R. Brent Church, and Kathleen Lynch. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language* 40, 4 (1999), 577–592. doi:10.1006/jmla.1998.2624

[16] Spencer D. Kelly and Quynh A. Ngo Tran. 2023. Exploring the Emotional Functions of Co-Speech Hand Gesture in Language and Communication. *Topics in Cognitive Science* (2023). doi:10.1111/tops.12657

[17] Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, UK.

[18] Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes* 24, 2 (2009), 145–167. doi:10.1080/01690960802586188

[19] Sotaro Kita, Ingebor van Gijn, and Harry van der Hulst. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Gesture and Sign Language in Human-Computer Interaction*, Ipke Wachsmuth and Martin Fröhlich (Eds.). Springer, Berlin, Germany, 23–35. doi:10.1007/BFb0052986

[20] Stefan Kopp. 2017. Computational Gesture Research: Studying the Functions of Gesture in Human–Agent Interaction. In *Why Gesture?* John Benjamins Publishing Company, Amsterdam, The Netherlands, 267–284.

[21] Dushyant N. Krishna and Anupama Patil. 2020. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. In *Interspeech*. 4243–4247.

[22] Taisia Kucherenko, Richard Nagy, Michael Neff, Henrik Kjellström, and Gustav E. Henter. 2021. Multimodal Analysis of the Predictability of Hand-Gesture Properties. *arXiv preprint* arXiv:2108.05762 (2021). https://arxiv.org/abs/2108.05762

[23] Hedda Lausberg. 2013. The NEUROGES system for detailed gesture analysis. In *Gesture and Speech in Interaction*. Springer, Cham, Switzerland, 45–68.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2999–3007.

[25] Gavin Littlewort, Jennifer Whitehill, Tian Wu, Ivan Fasel, Michael Frank, Javier Movellan, and Marian Bartlett. 2011. The computer expression recognition toolbox (CERT). In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 298–305.

[26] Evelyn Z. McClave. 2001. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 33, 5 (2001), 855–878. doi:10.1016/S0378-2166(00)00048-3

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference (SciPy 2015)*. 18–24.

[28] David McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago, IL.

[29] Takashi Mori and Kazuhiro Otsuka. 2021. Deep Transfer Learning for Recognizing Functional Interactions via Head Movements in Multiparty Conversations. In *Proceedings of ICMI*. 370–378.

[30] Simba Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *Computer Graphics Forum* 42, 7 (2023), 569–596. doi:10.1111/cgf.14776

[31] Shogo Okada, Masashi Bono, Kenji Takanashi, Yasuo Sumi, and Kazuhiro Nitta. 2013. Context-based Conversational Hand Gesture Classification in Narrative Interaction. In *Proceedings of the ACM International Conference on Multimodal Interaction*. ACM, 303–310. doi:10.1145/2522848.2522886

[32] Shogo Okada and Kazuhiro Otsuka. 2017. Recognizing Words From Gestures: Discovering Gesture Descriptors Associated With Spoken Utterances. In *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*. IEEE, 302–308. doi:10.1109/FG.2017.22

[33] Keiko Onishi, Hiroshi Tanaka, and Shogo Nakamura. 2024. Multimodal Voice Activity Projection for Turn-Taking and Effects on Speaker Adaptation. *IEICE Transactions on Information and Systems* (2024). doi:10.1587/transinf.2024HCP0002

[34] Shumpei Otsuchi, Yoko Ishii, Momoko Nakatani, and Kazuhiro Otsuka. 2021. Prediction of Interlocutors' Subjective Impressions Based on Functional Head-Movement Features in Group Meetings. In *Proceedings of ICMI*. 352–360.

[35] Shumpei Otsuchi, Koya Ito, Yoko Ishii, Ryo Ishii, Shinichirou Eitoku, and Kazuhiro Otsuka. 2023. Identifying Interlocutors' Behaviors and its Timings Involved with Impression Formation from Head-Movement Features and Linguistic Features. In *Proceedings of ICMI*. 336–344.

[36] M. A. Ozdemir, D. H. Kisa, O. Guren, et al. 2022. Hand gesture classification using time–frequency images and transfer learning based on CNN. *Biomedical Signal Processing and Control* 77 (2022), 103787.

[37] Catherine Pelachaud, Carlos Busso, and Dirk Heylen. 2021. Multimodal Behavior Modeling for Socially Interactive Agents. In *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics, Volume 1: Methods, Behavior, Cognition*. Springer, Cham, Switzerland, 259–310.

[38] Hironori Saitō and Sōtarō Kita. 2002. *Gesture, Action, and Meaning*. Taishukan Shoten, Tokyo, Japan.

[39] Christopher Saund and Stacy Marsella. 2021. Gesture Generation. In *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics, Volume 1: Methods, Behavior, Cognition*. Springer, Cham, Switzerland, 213–258.

[40] Samuel Saunderson and Goldie Nejat. 2019. How Robots Influence Humans: A Survey of Nonverbal Communication in Social Human–Robot Interaction. *International Journal of Social Robotics* 11, 4 (2019), 575–608. doi:10.1007/s12369-019-00523-0

[41] Zheng Sun and Yutao Liu. 2021. Multimodal learning for affect recognition: Recent advances and challenges. *IEEE Multimedia* 28, 4 (2021), 57–68. doi:10.1109/MMUL.2021.3104688

[42] Kazuki Takeda and Kazuhiro Otsuka. 2021. Inflation-Deflation Networks for Recognizing Head-Movement Functions in Face-to-Face Conversations. In *Proceedings of ICMI*. 361–369.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).

[44] Holger Voß and Stefan Kopp. 2023. Augmented Co-Speech Gesture Generation: Including Form and Meaning Features to Guide Learning-Based Gesture Synthesis. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. ACM, 1–8. doi:10.1145/3570945.3607337

[45] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014), 209–232. doi:10.1016/j.specom.2013.10.001

[46] Suhang Wang, Lin He, and Leman Akoglu. 2020. Learning from Imbalanced Data: Review of Methods and Applications. *ACM Transactions on Knowledge Discovery from Data* 14, 5 (2020), 1–41.