Learning Co-Speech Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation

Esam Ghaleb
University of Amsterdam
e.ghaleb@uva.nl

Bulat Khaertdinov
Maastricht University

Wim Pouw
Radboud University

Marlou Rasenberg
Meertens Institute

Judith Holler & Aslı Özyürek
Radboud University & MPI for
Psycholinguistics

Raquel Fernández
University of Amsterdam
raquel.fernandez@uva.nl

ABSTRACT

In face-to-face dialogues, the form-meaning relationship of co-speech gestures varies depending on contextual factors such as what the gestures refer to and the individual characteristics of speakers. These factors make co-speech gesture representation learning challenging. How can we learn meaningful gestures representations considering gestures' variability and relationship with speech? This paper tackles this challenge by employing self-supervised contrastive learning techniques to learn gesture representations from skeletal and speech information. We propose an approach that includes both unimodal and multimodal pre-training to ground gesture representations in co-occurring speech. For training, we utilize a face-to-face dialogue dataset rich with representational iconic gestures. We conduct thorough intrinsic evaluations of the learned representations through comparison with human-annotated pairwise gesture similarity. Moreover, we perform a diagnostic probing analysis to assess the possibility of recovering interpretable gesture features from the learned representations. Our results show a significant positive correlation with human-annotated gesture similarity and reveal that the similarity between the learned representations is consistent with well-motivated patterns related to the dynamics of dialogue interaction. Moreover, our findings demonstrate that several features concerning the form of gestures can be recovered from the latent representations. Overall, this study shows that multimodal contrastive learning is a promising approach for learning gesture representations, which opens the door to using such representations in larger-scale gesture analysis studies.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → **Computer vision tasks**; **Computer vision representations**;

KEYWORDS

Gesture analysis; face-to-face dialogue; representation learning; intrinsic evaluation; diagnostic probing.

1 INTRODUCTION

Co-speech hand gestures are intentionally used along with speech to convey meaning [43]. For instance, representational iconic gestures depict objects, events or actions through various representational techniques such as enacting, tracing, and hand-shaping [31]. Gesture analysis is an active research area in fields such as Human-Computer Interaction (HCI) [39], Sign Language Recognition (SLR) [29, 32], and human behavior analysis [17, 34], where sensory data collected through wearable sensors [22] or, more commonly, through passive sensors like RGB or depth cameras are widely used for studying gestures [50, 59, 60].

In face-to-face interaction, the form-meaning relationship of co-speech gestures is influenced by various situational and contextual factors, including what a gesture refers to and the characteristics of individual speakers. Although multiple current studies aim to model and represent gestures, there are prominent areas with room for improvement, particularly concerning gesture representation learning in conversations [18, 19, 41, 61, 62, 64]. *First*, most studies train deep learning architectures from scratch on specific downstream tasks, including gesture segmentation [18, 19, 61] or generation [41, 62, 64]. Thus, the employed objectives are focused on the task-specific discriminative and generative power of the models rather than on their ability to effectively encode general meaningful properties of gestures and relationships between them. The research literature has already pointed out the lack of models to represent gestures [41, 62, 64]. For example, some authors use autoencoders to extract and compare latent representations to evaluate synthetic gestures [41, 62, 64]. *Second*, much of the research has also focused on emblems (which are conventionalized and can stand alone without speech) [6, 33, 65] or pure beat gestures (rhythmic gestures that lack semantic content), particularly within monologue-based datasets as surveyed in [45]. Yet, in real-world interactions, gestures are semantically, pragmatically, and temporally related to speech, especially in face-to-face conversations. For instance, along with speech, representational gestures during face-to-face interaction facilitate the identification of referents and the establishment of shared understanding [25].

In this study, we address these limitations by focusing on learning representations of co-speech gestures in face-to-face dialogue.

These representations should exhibit desirable intrinsic properties, i.e., properties that align with expert intuitions. We propose to learn such representations using self-supervised learning approaches. In particular, building upon a pre-trained sign-language recognition model [29], we exploit unimodal (skeletal) and multimodal (speech-skeletal) contrastive learning to encode co-occurring gestural and speech patterns without using any manually annotated features. We opt for contrastive learning objectives because they allow us to simultaneously refine visual gesture representations and ground them in robust speech features learned by a foundation speech model. Furthermore, such a framework opens up possibilities for integrating other sources of information, e.g., semantic information through models processing text, in future work.

For training models with the proposed objectives, we utilize a dataset of naturalistic face-to-face *referential communication* rich with representational iconic gestures collected by Rasenberg *et al.* [52]. In this dataset, speakers engage in a conversation to identify objects that they are not familiar with. Gesture strokes (i.e., the most meaning-bearing component of a gesture [43]) are manually segmented by experts, and we take these segments as our starting point for representation learning. Moreover, a subset of pairs of gestures by two different speakers were coded by the dataset authors as similar or dissimilar with respect to five form features: handedness, position, shape, rotation, and movement. We use this annotation to evaluate the intrinsic quality of the gesture representations learned by our models. To summarise, we make the following contributions:

- We propose a self-supervised contrastive learning approach to simultaneously learn co-speech gesture representations through skeletal information while grounding them in spoken language.
- We present extensive intrinsic evaluations of the learned representations by investigating how well they capture pairwise gesture similarity. Our results show a significant positive correlation with human-annotated form feature similarity and reveal that the similarity between the learned representations is consistent with well-motivated patterns concerning speakers, referents, and the dynamics of dialogue interaction.
- Finally, we conduct a diagnostic probing analysis to investigate whether the model representations implicitly encode interpretable gesture features, finding that several form features can be recovered to some extent from the latent representations learned via self-supervised contrastive learning.

Overall, this study shows that multimodal contrastive learning is a promising approach to learning gesture representations whose properties are well aligned with human judgments and with theoretically motivated expectations. This paves the way for using self-supervised gesture representations to scale up gesture analysis studies that currently rely on small hand-annotated samples.

2 RELATED WORK

2.1 Gesture Representation and Modelling

2.1.1 Gesture Representation and Similarity. Prior to the rise of deep learning, researchers utilized handcrafted features, such as velocity, rhythm, acceleration, and anatomical models of hands, to encode kinematic, physical, and shape characteristics of gestures [53]. Nowadays, the focus has shifted to training neural networks end-to-end for certain downstream tasks, such as gesture recognition [42], detection [18], and generation [64]. The studies exploring gesture generation frequently involve quantitative evaluations for the quality and fidelity of synthetic gestures using custom metrics, such as Fréchet Gesture Distance (FGD) and Beat Alignment Score (BeatAlign) [62, 64]. For example, FGD evaluates generated gestures based on learned visual features through autoencoders [40, 64].

The similarity of gesture or sign pairs has been studied in a range of contexts, for example, in silent gestures, co-speech gestures, co-singing gestures, as well as sign language generation and production [30]. For instance, Kanakanti *et al.* [30] used Dynamic Time Warping (DTW) to evaluate speech-to-sign language generation models. These metrics assess the alignment between the predicted sign language sequences and the ground truth of Indian Sign Language sequences. One drawback of spatio-temporal alignment metrics such as DTW is that they operate at the level of key points and might not be optimal for measuring the more context-dependent similarity of two co-speech gestures. Indeed, a very weak but reliable correlation is generally found between the dissimilarity of gesture kinematics relative to the dissimilarity of what the gestures are about [49]. In our work, we assess the properties of the learned representations by performing pairwise gestural analysis. Hence, FGD and BeatAlign are not suitable metrics for this study. Rather, we calculate the cosine similarity of the learned representations, as often done in Natural Language Processing and speech research, where it is common to use embeddings for representational and similarity analysis [12, 44, 46, 48].

2.1.2 Joint Speech and Gesture Models. Joint modeling of speech and gestures through speech and vision models has been studied through different gesture analysis tasks. For instance, Ghaleb et al. [18] used speech to enhance gesture detection methods. A study by Kucherenko et al. [35] found that the meaning and timing of gestures can be predicted using prosodically relevant acoustic features. In gesture generation, according to a survey by Nyatsanga *et al.* [45], adding information about speech prosody and meaning can improve the quality of generated beat and representational gestures, respectively, making them appear more natural based on evaluations by raters. A typical pipeline representing gesture generation models is the work by Bhattacharya *et al.* [10], who developed a dual-speech model. The first model uses prosodic features, while the second one leverages word embeddings to capture speech's prosodic and semantic characteristics separately. Another example closely related to ours is the study by Lee *et al.* [36], who proposed an approach that exploits spoken language and gestures for gesture generation through cross-modal contrastive learning. The authors cluster speech and gesture embeddings cross-modally. In contrast, we include cross-modal objectives to complement unimodal gesture representations, as explained in Section 4.

In sum, gesture representations are typically learned through supervision signals tailored for a certain task. Such an approach might be suboptimal for learning general-purpose gesture representations. Autoencoders, in turn, are an established technique for this purpose in unimodal settings [41, 64], but are not frequently used when researchers are interested in constructing *multimodal* representations. In this paper, we aim to learn general gesture representations that capture properties that align with expert intuitions,

grounding their visual form-based characteristics in speech signals via unimodal and cross-modal contrastive learning.

2.2 Pre-training through Contrastive Learning

Contrastive learning has emerged as an efficient self-supervised learning strategy for pre-training deep learning models across various fields without using data annotations [28, 51]. The main idea behind this paradigm is to maximize a similarity metric computed between pairs of semantically close instances, i.e., positive pairs, by contrasting them against semantically different or negative pairs. The contrastive learning approach can be flexibly used for both unimodal [14, 24] and multimodal [21, 58] representation learning.

Both unimodal and multimodal contrastive objectives have been applied to skeleton data. Without ground truth annotations, semantic similarity is typically defined by the correspondence of feature representations to instances. In the unimodal case, a positive pair is typically formed by two views augmented from the same instance [14]. Building upon this idea, multiple works [23, 57] propose to use various spatial and temporal augmentations applied to skeletal key points for pre-training of graph- and sequence-based architectures. In turn, multi-view and multimodal contrastive objectives treat an instance represented through different views and modalities as a positive pair [51, 58]. Applied to skeletal data, these approaches aim to pre-train models by maximizing instance-level similarities of representations extracted from different views of skeletons, e.g., joints and motion [37], or modalities such as RGB and depth videos [26] or inertial measurement units [11]. In this paper, motivated by the effectiveness of contrastive self-supervised learning, we adopt both unimodal and multimodal objectives to learn representations of co-speech gestures without using data annotations.

3 DATASET AND PRE-PROCESSING

3.1 Dataset

We use a dataset collected by Rasenberg *et al.* [52] that consists of 19 naturalistic face-to-face dialogues. In these dialogues, Dutch-speaking participants play a referential game where they need to identify different objects or *referents* that do not have a conventional label. The game setup is shown in Figure 1. Each pair of participants plays the game for six rounds. In each round, one speaker acts as the 'director,' describing one of the target objects, while the other speaker acts as a 'matcher' who attempts to identify the object among several candidates. Each round includes 16 trials (the total number of objects), and the speakers exchange the director-matcher roles after each trial. The speakers were recorded from three angles, and we use video recordings from semi-frontal views (as shown in Figure 1). The dataset includes 38 subjects and more than 8 hours of recordings and is part of the larger CABB dataset [17].

The speakers are free to communicate as they please while playing the game; they were not given any instructions on the use of gestures. However, due to the nature of the referential task, the dataset is rich with representational iconic gestures, i.e., gestures used to depict objects. In this dataset, gesture strokes were manually identified and segmented. In addition, gestures were annotated according to what subpart of an object they refer to (see an example object and highlighted subpart in Figure 1). This annotation

resulted in 4949 gesture segments with an average duration of 610 milliseconds.
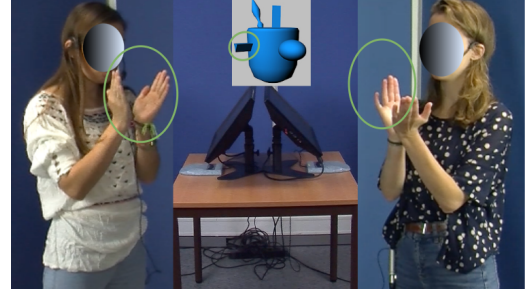

Figure 1: Pair of gestures referring to the highlighted subpart of the non-conventional object. The gesture pair is coded as similar in all form features: handedness (speakers use both hands), shape, position, rotation, and movement.

3.2 Manual Similarity Coding of Gesture Pairs

A subset of the dataset was annotated by Rasenberg *et al.* [52] to study the extent to which speakers mimic each other through gestures. The researchers coded the similarity of 419 semantically related pairs of gestures (i.e., gestures referring to the same subpart as shown in Figure 1) made by two speakers within a dialogue. The annotation codes, in a binary manner, whether two gestures are similar or not with respect to five form features: shape, movement, rotation, position, and handedness (whether both speakers use the left hand, the right hand, or both hands). This set of features is well-motivated and close to the set used by Bergmann & Kopp [9], who studied the similarity of temporally related gestures. As an example, the two gestures in Figure 1 were coded as similar for all features since both speakers use both hands with the same shape, movement, position, and rotation.
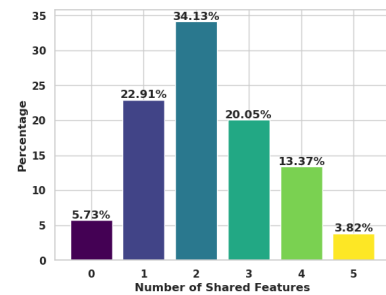

Figure 2: Distribution of the number of shared form features in the gesture pairs manually annotated by Rasenberg *et al.*

The distribution of the number of form features for which the gesture pairs are similar can be found in Figure 2. The majority of referentially related gesture pairs by two speakers are similar with respect to at least one feature, and around 38% are similar with respect to at least 3 features out of 5. Handedness similarity occurs the most frequently, followed by orientation, shape, movement, and

position. However, similarity with respect to all form features is rare (only 3.82% of gesture pairs). Section A of the Supplementary Materials provides further details on the inter-coder reliability of the annotation and the value distribution of each feature.

3.3 Data Samples

Using the manually segmented 4949 gesture strokes as an anchor, we sample one-second windows around these strokes. We apply a sliding window with an offset of 2 frames and take any time windows that overlap > 50% with gesture strokes. We do so because the segmented gesture strokes exclude other gesture phases, such as preparation, retraction and hold [43], which can provide very useful information to represent gestures [19]. This procedure resulted in 70153 gestural windows. Since speech and co-speech gestures are semantically and temporally coordinated but not perfectly aligned, we take a larger speech window following ghaleb *et al.* [18]. For each one-second gesture window, we include an extra half-second before and after, resulting in a two-second speech window.

For the speech modality, we use the underlying raw audio waveform of each speech time window. For the visual modality, we use MMPose [55] to obtain body poses from the gestural time windows. MMPose extracts the 2D positions (i.e., the x and y coordinates) of body joints and the detection confidence of 133 body joints. We utilize only 27 hands and upper body joints, which are the most relevant for sign language recognition [29] and hand gesture detection [19]. As proposed by Yan *et al.* [63], we create spatio-temporal graphs from the detected poses. Each graph contains j joints (vertices) and e edges and is represented as $x = (j, e)$. The resulting graph has two types of edges: spatial edges that correspond to the natural connectivity of joints and temporal edges that link the same joints across successive frames. As a result, a time window of a skeleton is represented with a tensor as follows: $X^G \in \mathbb{R}^{c \times t_v \times j}$, where c is a joint data point (x, y, confidence), t_v is the number of frames in a visual time window, and j is the number of body joints.

4 MODELS AND CONTRASTIVE OBJECTIVES

In this study, we use contrastive self-supervised learning to refine the embeddings extracted by two pre-trained backbone ("tower") models: a speech model (Section 4.1.1) and a skeleton model (Section 4.1.2). Given that contrastive objectives aim to align representations, our goal here is to bring representations of speech and co-speech gestures closer in a latent space. More formally, given a dataset $\{x_l^S, X_l^G\}_{l=1}^N$ of size N containing matching speech and gestural signals, their encoded representations can be written as $\{f_S(x_l^S), f_G(X_l^G)\}_{l=1}^N$, where $f_S(\cdot)$ and $f_G(\cdot)$ are feature encoders for the speech and gesture modality, respectively. We also use the projection heads $g_S(\cdot)$ and $g_G(\cdot)$ (Section 4.1.3), that map features from two modalities $g_S(f_S(x_l^S)), g_G(f_G(X_l^G))\}_{l=1}^N$ to the same size in a joint latent space [14, 51]. The proposed framework is illustrated in Figure 3.

The following subsections describe the backbone models, projection heads, and contrastive learning objectives. Section B of the Supplementary Materials details the implementation and data augmentation methods. Additionally, all corresponding data and code needed to reproduce the results are publicly available on GitHub.[1]

[1] https://github.com/EsamGhaleb/Learning-Co-Speech-Gesture-Representations

4.1 Backbone Models and Projection Heads

4.1.1 Speech Model. For the speech stream, we use the pre-trained WAV2VEC-2 model, which is trained with a masked language-modeling objective on large amounts of speech data [2]. Concretely, we use the `wav2vec2-xlsr-300` version, pre-trained on the raw speech waveforms in multiple languages. The model first applies Convolutional Neural Networks (CNNs) to 1-D waveforms as a feature extraction layer, followed by 24 Transformer encoders. The CNN block extracts representations of fine-grained speech segments (20 milliseconds long). Sequences of these representations are then used as separate tokens in the 24 self-attention blocks. WAV2VEC-2 has shown state-of-the-art performance across different speech-based tasks, such as automatic speech recognition [27] and speech emotion recognition [13]. We expect that the information related to speech prosody and content captured by this model will be beneficial for co-speech gesture representations.

In our implementations, we build a downstream network on top of the output sequences corresponding to two-second speech windows, obtained from all 24 self-attention layers. We do not simply use the features of the last layer because outputs of different layers contain different amounts of semantic and prosodic information [38, 56], which can be influential for guiding co-speech gestures [45]. As a result, the obtained representations have high dimensionality. To map information from these features into a lower-dimensional vector, we apply a weighted average with learnable weights for embeddings from each network layer. Furthermore, two pointwise CNN layers are implemented to fuse signals along the temporal dimension [47]. As a result, the output of the employed architecture is a one-dimensional vector containing 128 dimensions.

4.1.2 Skeletal Model. To encode the spatio-temporal graphs of body joints for each one-second gesture window, we use Spatio-Temporal Graph Convolutional Networks (ST-GCNs) [63], which generalize traditional convolution operations to graph neural networks. ST-GCNs embed the bodily movements of each gestural time window, capturing their spatial-temporal dynamics. We use a pre-trained model for Sign Language Recognition (SLR) by Jiang *et al.* [29]. The model was trained for Turkish Sign Language recognition using only 27 hand and upper body joints' coordinates. We refer to this model as the **SLR baseline model**. This model takes a spatio-temporal graph of a skeleton window as input and produces an embedding vector with 256 dimensions.

4.1.3 Projection Heads. The projection head for each modality consists of a three-layer MLP. The first layer is linear and has 128 dimensions for the speech head and 256 dimensions for the skeletal head. This layer is followed by a ReLU activation, which then leads to a final linear layer that projects both modalities into a 128-dimensional space. The objective of both heads is to project encoded features into a space that can be effectively compared through the contrastive learning method.

4.2 Unimodal Contrastive Objective

We use the SimCLR [14] for training the encoder from the SLR baseline. This framework crafts the positive pairs for the contrastive objective by applying skeletal augmentations that generate multiple
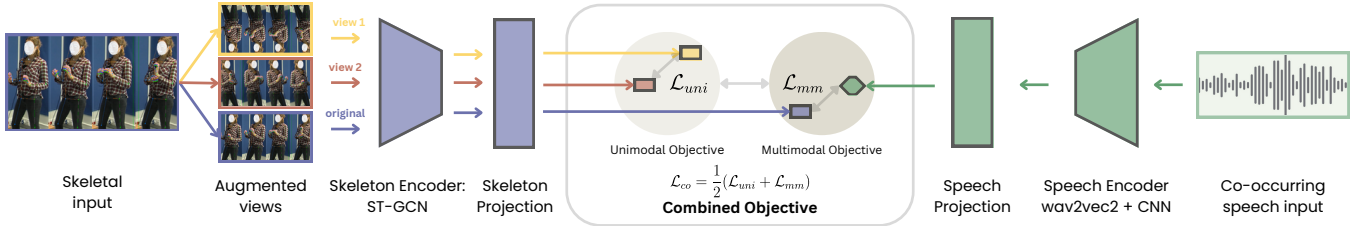
Figure 3: The proposed contrastive learning framework utilizing both unimodal and multimodal objectives.

views of the same instance, training the model to build representations of the underlying gesture, regardless of the augmentations. We use a combination of mirroring, scaling, random moving, jittering, and shearing, each with a 50% probability. Thus, a gesture instance X_l^G is transformed into two views with two random skeleton augmentations $t_i(X_l^G)$ and $t_j(X_l^G)$. Then, the transformed views are passed through the encoder and projection to obtain the view representations $z_i^G = g_G(f_G(t_i(X_l^G)))$ and $z_j^G = g_G(f_G(t_j(X_l^G)))$ for the loss calculation. For this positive pair, the contrastive loss function treating z_i^G as an anchor can be written as follows:

$$l_{uni}(i, j) = -log \frac{exp(\frac{s(z_i^G, z_j^G)}{\tau})}{\sum_{k=1}^{2N_b} \mathbb{I}_{[k \neq i]} exp(\frac{s(z_i^G, z_k^G)}{\tau})}, \quad (1)$$

where τ is a temperature hyperparameter and $s(\cdot)$ is a cosine similarity, and N_b is a batch size. The **unimodal contrastive objective** for the whole batch, then, is formulated as follows:

$$\mathcal{L}_{uni} = \frac{1}{2N_b} \sum_{\{i,j\} \in N_b} (l_{uni}(i, j) + l_{uni}(j, i)). \quad (2)$$

4.3 Multimodal Contrastive Objective

For multimodal training, we use Contrastive Multiview Coding [58], also frequently referred to as the CLIP-like objective [51]. This framework allows us to align and ground representations of co-speech gestures in corresponding speech without using data annotations. Concretely, contrastive learning aims to maximize the similarity between representations of matching gesture-speech pairs $\{z_l^S = g_S(f_S(x_l^S)), z_l^G = g_G(f_G(X_l^G))\}$ by contrasting them against other instances in mini-batches. In this case, the loss function for the l-th pair of instances contains two terms, $l_{mm}(G, S)$ and $l_{mm}(S, G)$. The former considers gestural features as anchors and mines positive and negative speech representations to form pairs, whereas the latter takes speech representations as anchors. More formally, the **multimodal contrastive objective** $l_{mm}(G, S)$ can be written as follows:

$$l_{mm}(G, S) = -log \frac{exp(\frac{s(z_l^G, z_l^S)}{\tau})}{\sum_{k=1}^{N_b} exp(\frac{s(z_l^G, z_k^S)}{\tau})}, \quad (3)$$

Similarly to Equation 2, the loss for the whole batch with multimodal representations is formulated as follows:

$$\mathcal{L}_{mm} = \frac{1}{2N_b} \sum_{l=1}^{N_b} (l_{mm}(G, S) + l_{mm}(S, G)). \quad (4)$$

4.4 Combined Objective

Finally, we propose an architecture that employs unimodal and multimodal objectives jointly to map related augmented skeleton representations closer while also grounding them in the co-occuring spoken language. We achieve this **combined objective** as follows: $\mathcal{L}_{co} = \frac{1}{2}(\mathcal{L}_{uni} + \mathcal{L}_{mm})$.

5 GESTURE SIMILARITY ANALYSIS

In this section, we first evaluate the representations obtained with our four models: the SLR baseline model and the three contrastive models with unimodal, multimodal, and combined objectives. For the three proposed contrastive models, the gestural representations are extracted from the last layer of the skeletal projection head. For the SLR baseline model, the representations are obtained from the last FC layer. We then use the best model's gestural representations to study gesture similarity within and across speakers in face-to-face referential dialogue.

5.1 Evaluation of Model Representations

We assess the effectiveness of the models' representations by examining the extent to which pairwise gesture similarity correlates with the number of form features shared by gesture pairs. This kind of *intrinsic* evaluation of model representations against human similarity judgments is common in natural language processing [44] and vision-and-language research [12, 48]. Here, we apply it for the first time to gesture representation learning.

For this evaluation, we use the 419 gesture pairs manually coded for similarity in the dataset by Rasenberg *et al.* [52]. As explained in Section 3.2, these gesture pairs have been annotated with 5 binary form features indicating whether the two gestures are similar or not with respect to handedness, shape, rotation, movement, and position. For each model, we compute the Spearman correlation coefficient between the pairwise cosine similarity of model representations and the number of form features these pairs share. The number of shared form features ranges from 0 (no similarity) to 5 (similar with respect to all features)—see Figure 2. Cosine similarity of model representations has a range of $[-1, 1]$.

The SLR baseline model shows a correlation coefficient of $\rho = 0.23$. Surprisingly, the unimodal model trained with contrastive learning on our dataset gives a slightly lower correlation of $\rho = 0.20$. In contrast, the multimodal model (i.e., the model trained with the multimodal contrastive objective) displays a correlation of $\rho = 0.24$, i.e., slightly higher than the baseline. The highest correlation, $\rho = 0.31$, is obtained with the representations learned

by the model with the combined training objective. All correlation coefficients are significant, with $p < 0.001$.
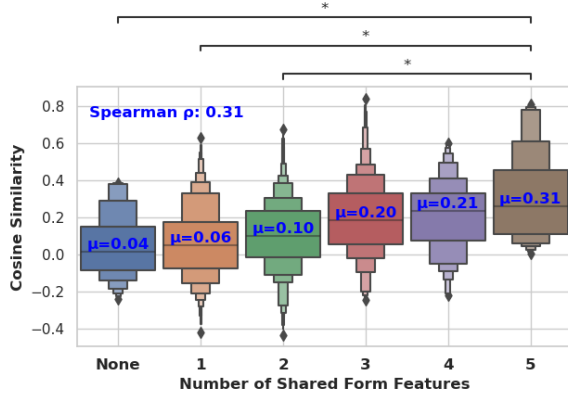

Figure 4: The cosine similarity scores distribution between pairs of gestures' representations, based on the number of shared form features in each pair. The similarity scores of gesture pairs sharing 5 form features are significantly higher than the similarity scores of gesture pairs sharing 2, 1 or 0 form features, respectively (t> 3.8, p-value< 0.001)

Figure 4 shows the distribution of similarity scores by the best-performing model with the combined objective for gesture pairs that share different amounts of form features. As can be observed, when there are no shared features (i.e., zero similarity according to human judgments), the mean average similarity score is approximately 0.04, significantly lower (according to an independent t-test with Bonferroni correction, t=$-3.855, p < 0.001$) than the mean similarity when the two gestures share all form features, which is approximately 0.31. The plots corresponding to the other models can be found in Section C.1 of the Supplementary Materials.

In sum, we conclude that the model with the combined contrastive objective is able to learn robust gesture representations that are well aligned with manually coded representations to a larger extent than the representations learned by the other models we test. In Section 6, we will dive deeper into the model representations to investigate whether they encode specific features.

5.2 Gesture Similarity in Referential Dialogues

In this section, we investigate to what extent the learned gesture representations comply with well-motivated expectations regarding gesture similarity within and across speakers. We report the results obtained with the best model's representations according to the evaluation carried out above, but the same pattern of results holds for the four models we test (see Section C.2 in the Supplementary Material). For these analyses, we include the model representations for all gestures in the dataset by Rasenberg *et al.* [52]. Recall that all gestures are annotated with their referent.

5.2.1 Referent vs. Speaker Driven Similarity. In the first analysis, we focus on gestures that take place within each two-participant dialogue and investigate the interplay between two variables regarding gesture pairs: the *speaker* (whether two gestures are produced

by the same speaker or not) and the *referent* (whether two gestures refer to the same object or not). We formulate the following hypotheses:

H1 a. Representations of gestures by the same speaker will be more similar if the gestures have the same referent than if they refer to different objects.

   b. Representations of gestures made by different speakers will be more similar if the gestures have the same referent than if they refer to different objects.

H2 Representations of gestures with the same referent will be more similar if the gestures are produced by the same speaker than if they are made by different speakers.

The two branches of hypotheses H1 are motivated by the iconicity of the gestures present in the dataset we analyze: iconic gestures resemble what they depict; in addition, the use of an iconic gesture is often concurrent with a referential expression articulated in speech. For example, to refer to the object part highlighted in Figure 1, the participants produce a gesture resembling a book while uttering *"you mean with that open book, so to speak", "yeah a bit like you've opened a book"*. Hence, we hypothesize that our model representations should be well-equipped to capture the visual and speech similarities that tend to characterize same-referent gestures, both when they are produced by the same speaker (H1a) as well as when they are produced by different speakers (H1b).

Hypothesis H2 is motivated by the assumption that there are individual idiosyncrasies in how speakers gesture and speak. For example, the relationship between systematic and idiosyncratic patterns in the production of iconic gestures has been studied by Bergmann and Kopp [7, 8], where both factors were found to impact gesture formation. Systematic factors and dialogue context affect the production of iconic gestures across speakers, but individuals may differ in how they realize their gestures, e.g., regarding the amount of shaping or drawing motions they use.

To test whether the learned representations comply with these hypotheses, we extract four sets of gesture representations: *same-referent-same-speaker* (7K pairs), *same-referent-different-speaker* (5.5K pairs), *different-referent-same-speaker* (455K pairs), and *different-referent-different-speaker* (358K pairs).[2]

Results. Figure 5 shows the distribution of cosine similarity scores for each of these sets of gesture pairs. We observe that gesture pairs produced by the same speaker to refer to the same referent exhibit the highest degree of similarity ($\mu = 0.34$), while gesture pairs produced by different speakers to refer to different objects are the least similar ($\mu = 0.07$). In compliance with H1a, gestures by the same speaker have lower similarity when they are not referentially related ($\mu = 0.34$ vs. $\mu = 0.17$). The same trend holds for gestures by different speakers, in line with H1b: their similarity is higher when the gestures refer to the same object ($\mu = 0.10$ vs. $\mu = 0.07$). Finally, we find that gestures with the same referent are more similar when they are made by the same speaker than by different speakers ($\mu = 0.34$ vs. $\mu = 0.10$), which complies with H2. All mean

---

[2]We use the total number of pairs available for each of these sets; the results we present next and their statistical significance remain stable when the number of pairs is downsampled (e.g., to 5.5k for each set).
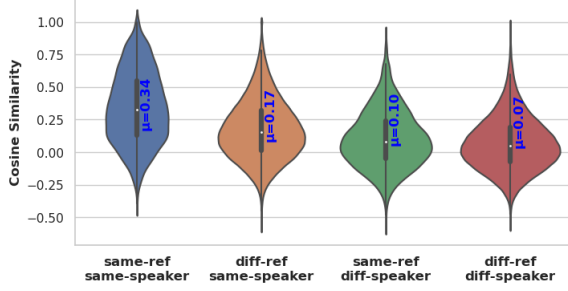
Figure 5: Distribution of cosine similarity scores of self and across-speaker gesture pairs in a dialogue when referring to the same or different referents. The labels 'same-ref' / 'diff-ref' indicate whether the gestures in a pair refer to the same object or not. According to the independent t-test with Bonferroni correction, distributions of similarity scores in all sets are significantly different.

differences are significantly different according to an independent t-test with Bonferroni correction.

In conclusion, the similarity of gesture pair representations learned by the models follows expected patterns that reflect both the interplay of speech and vision in iconic gestures that are referentially related, as well as individual characteristics of speakers.

5.2.2 Referent vs. Interaction Driven Similarity. In the previous analysis, we observed that referentially related gestures by different speakers are significantly more similar than cross-speaker gestures that do not refer to the same object ($\mu = 0.10$ vs. $\mu = 0.07$ in Figure 5). Is the similarity between referentially related gestures across speakers exclusively due to the shared referent (i.e., to the iconic character of the gestures), or is it also influenced by dialogue participants synchronizing their gestures through interaction? In studies of face-to-face dialogue, there is evidence that interlocutors tend to mimic each other's behavior—regarding both speech and gestures—when interacting [1, 20, 25, 52]. We therefore formulate the following hypothesis:

H3   Representations of gestures by different speakers will be more similar when the two speakers are interlocutors within a dialogue than when the speakers are from different dialogues.

To test whether the learned model representations confirm this hypothesis, we consider the two sets of gesture pairs by different speakers from the same dialogue we had extracted for the previous analysis (*same-referent-different-speaker* and *different-referent-different-speaker*) and extract two additional sets of pairs from different dialogues: *same-referent-different-speaker-diff-dialogue* (137K pairs) and *different-referent-different-speaker-diff-dialogue* (10M pairs).

Results. Figure 6 shows the mean similarities for the sets of pairs mentioned above. The results show that referentially related gesture pairs are most similar if they are produced by speakers interacting within a dialogue rather than by speakers who refer to the same object but do not directly interact with each other. Furthermore, even for gestures that do not refer to the same object, the representations across speakers are more similar when they are interlocutors in a dialogue. Taken together, these results lend evidence to H3: the
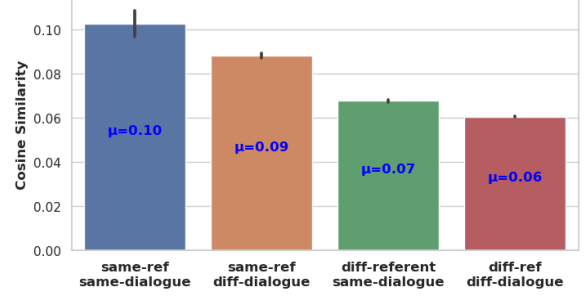

Figure 6: Similarity scores of gesture representation pairs by different speakers. The labels 'same-ref' / 'diff-ref' indicate whether the gestures in a pair refer to the same object or not; the labels 'same-dialogue' / 'diff-dialogue' indicate whether the speakers are interlocutors within the same dialogue or are participants from different dialogues. An independent t-test with Bonferroni correction shows that all sets' distributions of similarity scores are significantly different.

model representations reflect that gesture similarity is partially driven by alignment processes due to dialogic interaction.

6   PROBING ANALYSIS

The analysis presented in Section 5.1 showed that the similarity between model gesture representations significantly correlates with manually coded gesture similarity. However, that analysis does not reveal to what extent the latent representations may encode interpretable features. In this section, we shed light on this issue via diagnostic probing [4], a technique widely applied to decode linguistic properties present in textual representations [5, 15, 54] that has also been used to interpret features of other modalities, including visual signals [3] and spoken utterances [16].

Diagnostic probing is frequently formulated as a classification task. Specifically, a classifier is trained on top of the learned representations to predict a certain property. Here, we use the 419 gesture pairs manually annotated with 5 binary features and build 5 classification models, one per feature. For each pair of representations, the probing classifiers predict whether the representations are similar in handedness, shape, rotation, movement, or position.

6.1   Probing Models

We probe the representations learned by the model with the combined contrastive objective, in particular, the representation extracted from the last layer of the ST-GCN encoder. For comparison, we also probe the representations of the baseline SLR-skeleton model, taken from the same layer.

Probe architecture. Our probing models need to operate on top of pairs of representations as the binary labels are provided for pairs of gestures. Given that the gestures in each pair are by different speakers and there may be contrasts in input skeletal joints due to different camera viewpoints (e.g., mirrored views), we pass each representation in a pair through a simple linear layer with 32 neurons, followed by ReLU activation. Subsequently, the projected vectors are concatenated and used as input to a linear probe with
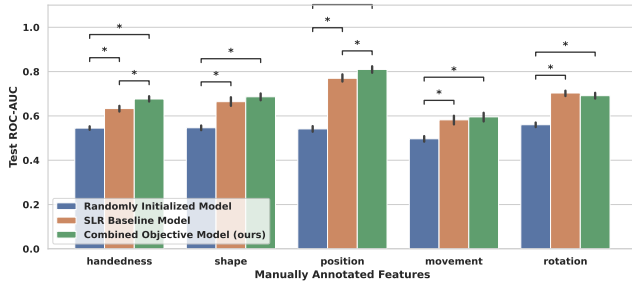
Figure 7: Test ROC-AUC scores of the probing models for each type of gestural similarity feature. The statistically significant differences are highlighted with * when observed.

sigmoid activation. Finally, the binary cross-entropy loss function is computed between the output of the probing model and the ground truth binary label indicating the similarity of gestures.

Random baseline. While the idea of diagnostic probing is simple, interpreting the results of a probing classifier requires careful consideration, as probing models may capture irrelevant patterns [4]. To account for this, we also train our probing model using random features. Specifically, these features are generated by an encoder following the same ST-GCN architecture but initialized with random weights. If the performance of the probe with the actual gesture representations is significantly higher than its performance with random features, we can conclude that the model representations encode, to some extent, the property being probed for.

Implementation details. The model is trained for 50 epochs using the Adam optimizer with a learning rate of $5e^{-4}$ for parameter tuning. For each property and type of gesture representation, 100 experiments were conducted using train, validation and test splits sampled with different random seeds and containing 60, 20 and 20% of gesture pairs, respectively. The probing models are evaluated using the ROC-AUC metric on the test set. For measuring the statistical significance of the observed differences, we employ the Mann-Whitney U-test ($\alpha = 0.05$) and apply Benjamini-Hochberg correction for multiple testing.

6.2 Results

The results of the probing analysis are summarized in Figure 7. The first result that stands out is that, for all the manually annotated features, the probes operating on model representations (by both the SLR model and the combined objective model) perform significantly better than those operating on randomly initilized representations. This suggests that information about all these properties—handedness, shape, position, movement, and rotation—is somewhat encoded by the models. Yet, the relative increase with respect to the random baseline differs per property, with position exhibiting the largest increase (more than 0.25 in mean ROC-AUC) and movement the smallest (0.1 mean ROC-AUC). In addition, we observe that the combined objective model representations tend to yield higher results than those by the SLR model for all features except rotation. Concretely, this advantage is statistically significant for handedness and position. This suggests that speech may

be more connected to these features of gestures when they are represented through the skeletal joints.

In sum, our analysis indicates that the models, in particular our combined objective model, learn gesture representations that encode different properties identified by experts through manual annotation. The results also suggest that information about movement appears to be the most challenging for the models to learn.

7 CONCLUSION

We proposed an approach for learning co-speech gesture representations using contrastive self-supervised learning. This approach employs loss functions to bring two views of gestural movements closer in a latent space while also grounding them in spoken language. To train the proposed models, we used segmented gestures obtained from a dataset of face-to-face referential dialogues without any additional labeling. While most work on gesture representation learning has focused on task-specific evaluation, in this paper, we performed a thorough intrinsic evaluation of the learned representations by analyzing the extent to which they exhibit desirable properties that align with expert intuitions. Our results showed that a model that combines both unimodal and multimodal contrastive objectives has advantages over a Sign Language Recognition baseline model and other contrastive variants. Concretely, we observed that (1) the representations learned by the combined objective model achieve the highest correlation with pair-wise similarity of form features annotated by experts, and (2) these form features, particularly handedness and position, can more accurately be decoded from the latent representations learned by this model, as demonstrated by a probing analysis. This analysis also suggests that information about gesture movement is challenging for the models to learn. Future work could address this by encoding architectures that can exploit additional aspects of skeletal data, such as bones and motions.

We also examined how different factors, such as gesture iconicity, individual speaker characteristics, and dialogue coordination, influence the similarity between gesture pairs. Our analyses showed that model representations comply with theoretically-motivated expectations regarding these factors. In particular, they appear to capture three major sources of variance: degree of (1) idiosyncratic personal style, (2) common ground building in conversation, and (3) conventional iconic gesture use. We believe that this makes them a valuable tool for gesture analysis studies. When analyzing the data in more detail, beyond the hypotheses investigated in Section 5.2, we observed that for the gesture pairs where referents are different and the speaker is the same, the similarity of the representations learned by the multimodal models is higher than when the referent is the same, and the speakers are different (more details are available in Section C.2 of the Supplementary Materials). This suggests that the inclusion of speech information during training may amplify the encoding of individual speaker differences. We leave an in-depth investigation of this aspect to future work.

REFERENCES

[1] Sho Akamine, Esam Ghaleb, Marlou Rasenberg, Raquel Fernández, Antje Meyer, and Aslı Özyürek. 2024. Speakers align both their gestures and words not only to establish but also to maintain reference to create shared labels for novel objects in interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[3] Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michał Górszczak, Barbara Rychalska, Tomasz Trzcinski, and Bartosz Zieliński. 2021. Explaining Self-Supervised Image Representations with Visual Probing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 592–598. https://doi.org/10.24963/ijcai.2021/82 Main Track.

[4] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* 48, 1 (2022), 207–219.

[5] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology?. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 861–872. https://doi.org/10.18653/v1/P17-1080

[6] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. 2021. IPN hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 4340–4347.

[7] Kirsten Bergmann and Stefan Kopp. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks.. In *AAMAS (1)*. 361–368.

[8] Kirsten Bergmann and Stefan Kopp. 2010. Systematicity and idiosyncrasy in iconic gesture use: Empirical analysis and computational modeling. In *Gesture in Embodied Communication and Human-Computer Interaction: 8th International Gesture Workshop, GW 2009, Bielefeld, Germany, February 25-27, 2009, Revised Selected Papers 8*. Springer, 182–194.

[9] Kirsten Bergmann and Stefan Kopp. 2012. Gestural alignment in natural dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

[10] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2027–2036.

[11] Razvan Brinzea, Bulat Khaertdinov, and Stylianos Asteriadis. 2022. Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 01–08.

[12] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research* 49 (2014), 1–47.

[13] L. W. Chen and A. Rudnicky. 2023. Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1–5.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[15] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2126–2136. https://doi.org/10.18653/v1/P18-1198

[16] Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022-23rd INTERSPEECH Conference*.

[17] Lotte Eijk, Marlou Rasenberg, Flavia Arnese, Mark Blokpoel, Mark Dingemanse, Christian F. Doeller, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, Wim Pouw, Iris van Rooij, Herbert Schriefers, Ivan Toni, James Trujillo, and Sara Bögels. 2022. The CABB Dataset: A Multimodal Corpus of Communicative Interactions for Behavioural and Neural Analyses. *NeuroImage* 264 (Dec. 2022), 119734. https://doi.org/10.1016/j.neuroimage.2022.119734

[18] Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Ivan Toni, Peter Uhrig, Anna Wilson, Judith Holler, Aslı Özyürek, and Raquel Fernández. 2024. Leveraging Speech for Gesture Detection in Multimodal Communication. *arXiv:2404.14952v1* (2024).

[19] Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Peter Uhrig, Judith Holler, Ivan Toni, Aslı Özyürek, and Raquel Fernández. 2024. Co-Speech Gesture Detection through Multi-phase Sequence Labeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. CVF/IEEE, Hawaai, USA.

[20] Esam Ghaleb, Marlou Rasenberg, Wim Pouw, Ivan Toni, Judith Holler, Aslı Özyürek, and Raquel Fernández. 2024. Analysing Cross-Speaker Convergence in Face-to-Face Dialogue through the Lens of Automatically Detected Shared Linguistic Constructions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46.

[21] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[22] Lin Guo, Zongxing Lu, and Ligang Yao. 2021. Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 300–309.

[23] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. 2022. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 762–770.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[25] Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior* 35 (2011), 133–153.

[26] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. 2022. Versatile multimodal pre-training for human-centric perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16156–16166.

[27] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[28] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.

[29] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3413–3423.

[30] Mounika Kanakanti, Shantanu Singh, and Manish Shrivastava. 2023. MultiFacet: A Multi-Tasking Framework for Speech-to-Sign Language Generation. In *Companion Publication of the 25th International Conference on Multimodal Interaction*. 205–213.

[31] Adam Kendon. 2004. Gesture units, gesture phrases and speech. In *Gesture: Visible Action as Utterance*. Cambridge University Press, Chapter 7, 108–126. https://doi.org/10.1017/CBO9780511807572.007

[32] WW Kong and Surendra Ranganath. 2014. Towards subject independent continuous sign language recognition: A segment and merge approach. *Pattern Recognition* 47, 3 (2014), 1294–1308.

[33] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. 2020. Online dynamic hand gesture recognition including efficiency analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 2, 2 (2020), 85–97.

[34] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 97–104.

[35] Taras Kucherenko, Rajmund Nagy, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2022. Multimodal analysis of the predictability of hand-gesture properties. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*. ACM Press, 770–779.

[36] Dong Won Lee, Chaitanya Ahuja, and Louis-Philippe Morency. 2021. Crossmodal clustered contrastive learning: Grounding of spoken language to gesture. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. 202–210.

[37] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 2021. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4741–4750.

[38] Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2023. Exploration of a self-supervised speech model: A study on emotional corpora. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 868–875.

[39] D. Liu, L. Zhang, and Y. Wu. 2022. LD-ConGR: A large RGB-D video dataset for long-distance continuous gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3304–3312.

[40] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, ..., and B. Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*. Springer Nature Switzerland, Cham, 612–630.

[41] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis: Supplementary

Materials. *Gesture* 9, 10s (2022), 11s.

[42] Xin Liu, Henglin Shi, Xiaopeng Hong, Haoyu Chen, Dacheng Tao, and Guoying Zhao. 2020. 3D skeletal gesture recognition via hidden states exploration. *IEEE Transactions on Image Processing* 29 (2020), 4583–4597.

[43] David McNeill. 1992. Hand and mind. *Advances in Visual Semiotics* 351 (1992).

[44] Roberto Navigli and Federico Martelli. 2019. An overview of word and sense similarity. *Natural Language Engineering* 25, 6 (2019), 693–714.

[45] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*, Vol. 42. 569–596.

[46] Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics* 12 (2024), 372–391.

[47] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. *Proc. Interspeech 2021* (2021), 3400–3404.

[48] Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation. *Transactions of the Association for Computational Linguistics (TACL)* (2021). https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00443/1979754/tacl_a_00443.pdf

[49] Wim Pouw, Jan de Wit, Sara Bögels, Marlou Rasenberg, Branka Milivojevic, and Asli Ozyurek. 2021. Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics. In *International Conference on Human-Computer Interaction*. Springer, 269–287.

[50] Wim Pouw, James P Trujillo, and James A Dixon. 2020. The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior research methods* 52 (2020), 723–740.

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[52] Marlou Rasenberg, Asli Özyürek, Sara Bögels, and Mark Dingemanse. 2022. The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes* 59, 3 (2022), 209–236.

[53] Siddharth S Rautaray and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review* 43 (2015), 1–54.

[54] Samuel Ryb, Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. AnaLog: Testing Analytical and Deductive Logic Learnability in Language Models. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, Vivi Nastase, Ellie Pavlick, Mohammad Taher Pilehvar, Jose Camacho-Collados, and Alessandro Raganato (Eds.). Association for Computational Linguistics, Seattle, Washington, 55–68. https://doi.org/10.18653/v1/2022.starsem-1.5

[55] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. 2020. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sensors Journal* 20, 17 (2020), 10032–10044.

[56] Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah. 2021. What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *arXiv preprint arXiv:2101.00387* (2021).

[57] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. 2021. Skeleton-contrastive 3D action representation learning. In *Proceedings of the 29th ACM international conference on multimedia*. 1655–1663.

[58] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 776–794.

[59] Patrick C Trettenbrein and Emiliano Zaccarella. 2021. Controlling video stimuli in sign language and gesture research: The OpenPoseR package for analyzing OpenPose motion-tracking data in R. *Frontiers in Psychology* 12 (2021), 628728.

[60] James P Trujillo, Julija Vaitonyte, Irina Simanova, and Asli Özyürek. 2019. Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior research methods* 51 (2019), 769–777.

[61] Huogen Wang, Pichao Wang, Zhanjie Song, and Wanqing Li. 2017. Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision workshops*. 3138–3146.

[62] Zunnan Xu, Yachao Zhang, Sicheng Yang, Ronghui Li, and Xiu Li. 2024. Chain of generation: Multi-modal gesture synthesis via cascaded conditional control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6387–6395.

[63] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[64] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

[65] Guangming Zhu, Liang Zhang, Peiyi Shen, Juan Song, Syed Afaq Ali Shah, and Mohammed Bennamoun. 2018. Continuous gesture segmentation and recognition using 3DCNN and convolutional LSTM. *IEEE Transactions on Multimedia* 21, 4 (2018), 1011–1021.