CMCF: An Architecture for Realtime Gesture Generation by Clustering Gestures by Motion and Communicative Function

Carolyn Saund
Department of Computing Sciences
University of Glasgow
Glasgow, United Kingdom
carolyn.saund@gmail.com

Andrei Bîrlădeanu
Department of Computing Sciences
University of Glasgow
Glasgow, United Kingdom
a.birladeanu.1@research.gla.ac.uk

Stacy Marsella
Khory College of Computer Science
Northeastern University
Boston, Massachusetts
stacymarsella@gmail.com

ABSTRACT

Gestures augment speech by performing a variety of communicative functions in humans and virtual agents, and are often related to speech by complex semantic, rhetorical, prosodic, and affective elements. In this paper we briefly present an architecture for human-like gesturing in virtual agents that is designed to realize complex speech-to-gesture mappings by exploiting existing machine-learning based parsing tools and techniques to extract these functional elements from speech. We then deeply explore the rhetorical branch of this architecture, objectively assessing specifically whether existing rhetorical parsing techniques can classify gestures into classes with distinct movement properties. To do this, we take a corpus of spontaneously generated gestures and correlate their movement to co-speech utterances. We cluster gestures based on their rhetorical properties, and then by their movement. Our objective analysis suggests that some rhetorical structures are identifiable by our movement features while others require further exploration. We explore possibilities behind these findings and propose future experiments that may further reveal nuances of the richness of the mapping between speech and motion. This work builds towards a real-time gesture generator which performs gestures that effectively convey rich communicative functions.

KEYWORDS

gesture generation; rhetorical parsing; communication; clustering; machine learning

ACM Reference Format:
Carolyn Saund, Andrei Bîrlădeanu, and Stacy Marsella. 2021. CMCF: An Architecture for Realtime Gesture Generation by Clustering Gestures by Motion and Communicative Function. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

Gestures play a powerful role in human face-to-face interaction [21, 30], and moreover reflect the relation between thought, speech, and motion [10, 30]. Gestures are shown to mirror the fine-grained structure of dialogue, such as its underlying architecture comprised of logical and rhetorical units [19]. The complexity of the relationship between gesture and language is also compounded by the multiple levels at which one can observe correlations between motion and speech. Grady [15] found that situated language is frequently used to ground abstract metaphors in concrete physical descriptors ("These fabrics aren't quite the same, but they're *close*", p. 283), while Chiu et al. [7] showed how these conceptual metaphors are readily mapped onto everyday gestures.

One example of the multiple ways in which the communicative intention can be reflected in gestures is when a person presents the option of an "important or trivial idea" using different gestural performances. In one scenario they may emphasize the rhetorical contrast of "important *or* trivial" by holding up their right hand for important and their left hand for trivial. In another situation they may focus on the semantic aspects of the contrast, making a large gestural frame to emphasize "important," and move their hands close together when they utter "trivial" in order to convey the relative significance of the ideas through the metaphorical connection between importance and size.

Because the relationship between speech and gesture is nuanced, gestures generated by virtual agents often lack the same complexity displayed in human performances. Some gesture generators are rule-based [5, 7], and thus have a limited library of both gestures and understandings of when to deploy them. While many rule based approaches use acoustic data to modulate gesture [24, 27, 33] they are still beholden to rules which behavior designers implant in them. These rules, while effective and grounded in theory, are ultimately non-exhaustive and often prescriptive instead of reflective of gestures which occur naturally and spontaneously.

End-to-end machine learning approaches combat this, and have recently gained significant traction [12, 13]. These instead take video and audio data and use it to learn a mapping of speech to gesture. One challenge here is the need for sufficient data to capture the complex multi-faceted mapping between communicative function and gestures. As a result these technique are very good at conveying prosodic elements in the speech such as emphasis through rhythmic beat gestures [13, 27] but they lack the sufficient data to capture more complex relationships; they assume the gesture is solely driven - or at least captured - by the acoustic properties of speech, as opposed to some deeper communicative function that may not be reflected acoustically. In addition, these techniques largely forego designer control, other than limiting the data that is the input to machine learning, to, for example, specific speakers in order to capture that speaker's style [13].

Nevertheless, large language corpora have led to a range of evolving natural language tools, derived using machine learning, that can analyze prosody [27], syntactic structure [6], semantic and metaphoric elements [1, 31, 37] as well as rhetorical structure within text and dialog [18, 26, 34]. In addition, there is considerable

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

un-annotated video data that is available to analyze gestures, for example using tools such as OpenPose [4].

Our work has pursued the following ideas: (1) these different analysis techniques provide a way to extract different elements (semantic, rhetorical and prosodic) from speech while avoiding the limited data problem, (2) if we break analysis down into these elements we may be able to afford more designer control, (3) within a particular analysis element, we assume there will be a difference in gestural motion properties in order for the communicative function to be effectively conveyed, and (4) the breakdown into function and gestural motion also supports driving gestures directly from communicative functions if available. This suggests the following approach to generation: Perform these distinct analyses on speech extracted from video data. Within a particular analysis, such as rhetorical structure, cluster the associated gesture videos based on motion properties to derive clusters associated with different rhetorical elements such as contrast, elaboration, etc. These clusters then provide candidate gestural motions to convey these functions.

In this paper, we present and assess a potential architectural model of gesture generation which integrates rhetorical, semantic, affective, and acoustic relationships between utterances and their accompanying gestural motions. We first present the overall architecture, and then deeply explore the implementation of the rhetorical branch of this model as a demonstration of this novel method of clustering gestures based on co-speech elements and motion. We present our method of comparing and clustering gestural motion, building off of multiple third-party ontologies and ML-based NLP tools and the motion database found in [13], and provide techniques for evaluating the clustering of these gestures, as well as ways to overlay clusters to provide a complex picture of the relationship between motion and speech. We focus on what this clustering technique can objectively tell us about the relationships between rhetorical structure and gestural motion.

2 ARCHITECTURE OVERVIEW

In this section we present an architecture for a virtual agent which uses a pre-trained model to perform gestures, and which is agnostic about how the animations are realized. We refer to this as Clustering by Motion and Communicative Function (CMCF).

Our proposed architecture attempts to generate gestures which carry the rhetorical, semantic, and affective communicative functions of natural human gestures. While these categories are non-exhaustive, there is reason to believe that these provide an effective foundation for gesture analysis [7, 19, 37]. Since our demonstration and assessment in this paper focuses on rhetorical structure, we provide background on its relevance to gesture generation, with the recognition that all these elements play fundamental roles in non-verbal communication [2, 10, 36].

While the relationship between discourse structure and gesture has been explored in virtual agents, we explicitly explore the relationship between rhetorical structure and gesture with respect to Rhetorical Structure Theory [25]. Lascardes & Stone [23] conduct similar work bridging formal analyses with pragmatic interpretation and generation mechanisms, demonstrating the importance of the shared roles of theoretical and applied work in this area.

Previous studies have used information contained in rhetorical structures to generate nonverbal behaviour in virtual agents.

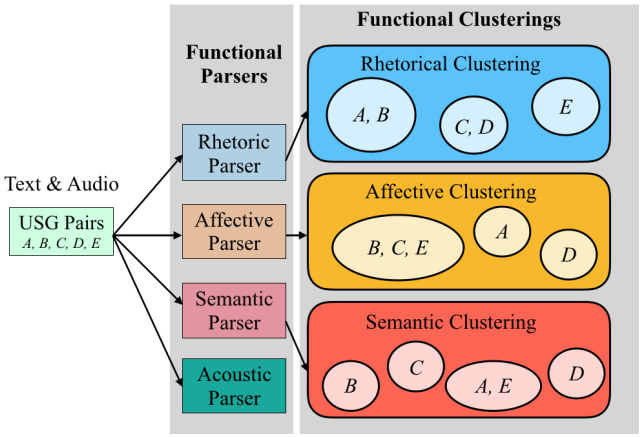


Figure 1: Overall architecture of generative model. During the pre-training step, example USG pairs A through E are tagged and grouped into functional clusters corresponding to the utterance. An elaboration on motion sub-clusters is found in Figure 2.

For example, [27] used a rule-based algorithm to extract semantic and rhetorical content from text and further applied it to generate nonverbal behavior, including gestures. By using the semantic and rhetorical content of discourse in addition to prosody to generate nonverbal behaviour, the character was shown to become more life-like, and was rated more highly on appropriateness compared to either prosodic based or random gestures. An important component of the mapping between speech and gesture thus appears to be the high-level relations between units of speech that might be projected onto specific hand movements during communication.

2.1 Clustering by Motion and Communicative Function (CMCF)

Our framework takes as input a piece of text and optionally an audio performance of that text. Its output can be used as an abstraction to an animation system. This system tags input speech with a variety of linguistic functional (discrete) labels using third-party parsers, which it then uses to derive appropriate gestures. Acoustic input is optional as each functional component of this model acts separately, and all available information is concatenated at the end.

The architectural overview for this model is shown in Figure 1. The pipeline contains three parallel processes for rhetorical, semantic, and affective domains [1]. It clusters gestures together categorically by tags given by the parsers, derived from the gesture's co-speech utterance. This way, when given an utterance, the agent performs a gesture that occurred with a linguistically similar utterance in the past. This works first by creating the clusters (Section 2.2) *offline*, then exploiting these pre-calculated clusters at run-time (Section 2.3).

[1] In this proposed architecture, acoustic information is to be used primarily to generate beat gestures, modulate expressive dynamics of gestures, as well as determine domain priorities over the gestural analyses spanning the utterance in the event multiple relevant domains cannot be co-articulated with a single gesture. In addition, letting acoustic information be optional allows the generator flexibility to use pre-recorded speakers or text-to-speech that may lack interesting prosodic variation.

2.2 Pre-training the model

In pre-training, the model draws from a set of gestures and their associated audio and transcription of utterances. Throughout this paper, we refer to each utterance segment and the gesture that co-occurs with it temporally as an **Utterance-Segment-Gesture (USG) pair.** In this context, the Utterance Segment is the segment of the utterance which is relevant to one particular rhetorical tag. For example, the phrase "I would tell him, but it is too late" would be parsed into multiple USGs: "I would tell him," and "but it is too late," and the relevant motion (specifically the gestural stroke) associated with only the corresponding specific segment of the overall utterance.

2.2.1 Functional Domains. For the purposes of this architecture we define a **Functional Domain** as a level at which natural language can be analyzed. In this case, we refer to the Rhetorical, Affective, and Semantic domains.

The architecture requires an interface to third-party **Functional Parsers**, with the possible outputs from these parsers defining the set of **Functional Tags** that can be applied to USG pairs in the input dataset. This interface makes the architecture agnostic to the parser's implementations. It is thus suitably flexible to accommodate evolving rhetorical, semantic, and affective text parsers popular in NLP communities, as well as acoustic feature extractors[2]. This modularity also allows our architecture to take a communicative intent or function as input, instead of text or audio. This feature gives it flexibility and an advantage over end-to-end machine learning, and makes it compatible with SAIBA guidelines for implementing virtual agents [22].

2.2.2 Clustering by Function and Motion. The architecture uses the functional parsers to assign functional tags to all USG pairs. Each USG pair thus has at least one functional tag within each functional domain. With these tags, it establishes a **Functional Clustering** by grouping USG pairs together with others with the same functional tag. This defines different clusterings for each functional domain, with different USG pairs grouped together in different domains.

For each cluster in each functional domain, it then creates a **Sub-clustering** based on the motion of the gesture (Figure 2). Each USG pair thus appears in exactly one (motion-derived) sub-cluster in at least one (functional) cluster, for each functional domain (Figure 1).

The motion sub-clusters are further refined through pruning and combining. It is necessary to prune out sub-clusters which are significantly larger than the rest. These can occur due to noise, because not every gesture within a USG pair with a particular functional tag is necessarily relevant to that functional domain. In the "important or trivial" example, in the *Size* cluster, this USG pair may not cluster neatly with others, instead clustering into a messier sub-cluster which can be avoided at runtime as it is unlikely to contain gestures that are meaningfully associated with the "Size" semantic aspect of speech. Accordingly, the architecture works by assigning multiple functional tags, forming a categorical clustering for each functional domain. This explicitly recognizes that the gesture may be relevant to, for example, the rhetorical structure of the utterance, but not to the semantic content.

[2]Although switching parsers would require an interface to be defined between the parser output and input to this model. For example, the output of rhetorical parsers differ according to the underlying theory on which they are based.

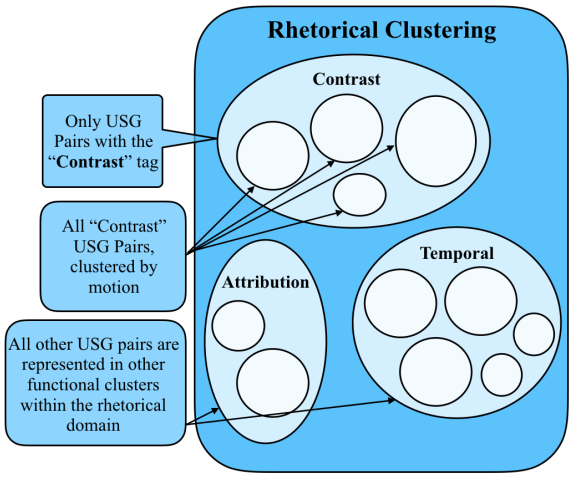


Figure 2: An illustration of rhetorical motion sub-clusters within tagged clusters in the functional Rhetorical domain.

Large motion sub-clusters also form because speakers are not constantly in motion as they speak, so many gestures have little to no motion at all and cluster together (motion does not take the speaker's static pose into account).

Following pruning of each sub-clustering, each sub-cluster is compared to each sub-cluster in the other domains to create a distance matrix for all sub-clusters that spans across functional domains. This matrix describes the difference in motion between one sub-cluster and all other sub-clusters (across all functional domains), providing crucial further information with which an agent can select which sub-cluster to execute at runtime. We discuss the runtime use of this distance matrix below.

2.3 Runtime execution

At runtime, an agent uses the same parsers from pre-training to analyze an incoming utterance to perform. The agent may then select one of these functional components to emphasize according to its context, or perform a beat gesture. How an agent can choose a domain to emphasize is explored in [9, 27, 32]. If the agent chooses to perform a gesture according to one of these functions, it selects a sub-cluster from the appropriate functional cluster to retrieve motion information and perform a gesture.

The agent must select between motion sub-clusters of its assigned functional cluster. To do this, the agent accesses the information in the distance matrix for each sub-cluster in the functional clusters. This is used to compare sub-clusters across the functional clusters that the utterance belongs to. For example, our "important or trivial," example with a rhetorical "Contrast," tag and a semantic "Size," tag (illustrated in Figure 3). The agent can compare the sub-clusters within these functional clusters to determine the nearest-neighbor sub-cluster, indicating that the particular motion described by these sub-clusters may be salient to multiple functional components of that utterance segment[3].

[3]Conversely, to reduce potential communicative ambiguity of a gesture, the agent could select a sub-cluster maximally different from other potentially relevant sub-clusters. The specifics of motion sub-cluster selection and its impact on subjective interpretation of gestures is not explored here.

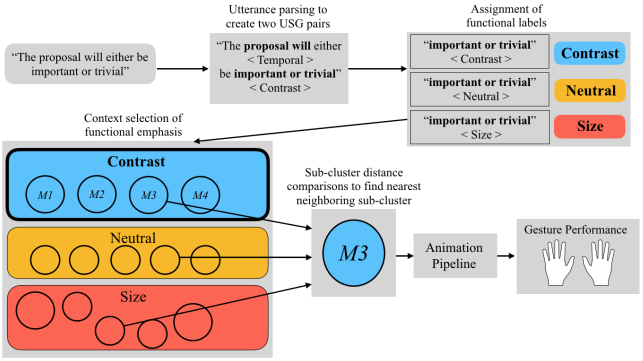


Figure 3: An illustration of how the architecture can select a gesture performance for the "important or trivial" example utterance.

Once a sub-cluster is selected, the agent may choose to perform a gesture from it in any number of ways. Our architecture does not prescribe a specific animation but rather a family of motions which the agent's overarching architecture may interpret in a manner appropriate to that specific agent (explored in section 2.4).

The labeling, clustering of the input dataset, and distance calculation of sub-clusters is done in pre-training. Because of this, the speed, and therefore feasibility of using this model in real time, is determined by the speed of the functional parsers in tagging an incoming utterance, as well as by the algorithm's contextual analysis in choosing a functional domain to emphasize.

2.3.1 Example Usage. We described the structure of this algorithm of pre-training to cluster a large corpus of gestures and select candidate gestures at runtime. We will now go step-by-step through our implementation of this architecture with an example utterance to illustrate how this model generates gestures in real time.

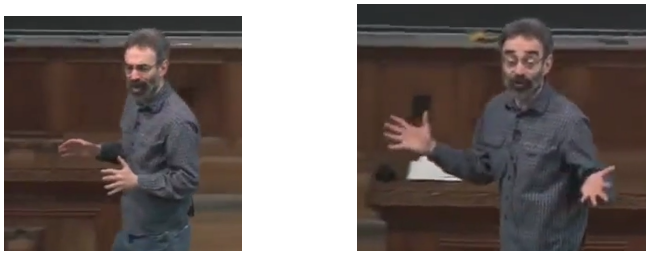
Let us give our example utterance "important or trivial" to an agent using this model to perform (visually illustrated in Figure 3). This incoming utterance is analyzed and broken up by the functional parsers and given the "Contrast" rhetorical tag, the "Neutral" affective tag, and the "Size" semantic tag. For purposes of this example, let us assume context tells the agent to emphasize the rhetorical domain of speech. We then look at all sub-clusters within the rhetorical *Contrast* cluster, and use the distance matrix calculated in pre-training to find the closest motion sub-cluster within the *Size* or *Neutral* functional clusters. The sub-cluster within the *Contrast* functional cluster which has the smallest distance to a motion sub-cluster of another domain - in this case either *Size* or *Neutral* - will be selected to perform. Potential examples of runtime animation using this pipeline are shown in section 2.4.

Figure 4 shows two specific candidate gestures from the *Contrast* motion sub-cluster obtained using this process in our specific implementation, described below in Section 3. Notice how despite different starting positions, angles, and even speakers, these two gestures follow a similar path, resulting in similar motions.

2.4 Animation Options

Although our architecture does not specify an animation pipeline, here we put forth several alternatives given the purpose of the model. As the intended use is a dynamic gesture generator to be

(a) Starting and post-stroke poses for gesture by Speaker 1



(b) Starting and post-stroke poses for gesture by Speaker 2

Figure 4: Two gestures from the same motion sub-cluster within the "Contrast" rhetorical cluster using our implementation and input dataset.

used for on-the-fly gesture generation, these options mainly occur after pre-training and prior to runtime usage.

One option may be that an animation is created to represent each sub-cluster, with sub-clusters providing reference video for the virtual agent designer or animator. Alternatively, one can analyze a pre-defined library of animations to determine which sub-cluster they each belong to, and use this to map sub-clusters to animations. In both cases, at runtime the agent would simply perform the animation associated with its chosen sub-cluster. This solution ensures the animation is appropriate and appears natural for the agent's form. This is feasible because the sub-clusters are clustered according to motion, and thus a sub-cluster can be represented by a single animation.

Another option could be to use the motion of USG pairs within the sub-cluster to define a "centroid" gesture. This has the added benefit of being able to be altered dynamically at runtime, although the architecture itself does not specify these alterations. However, this relies on the motion of the USG pairs to be transformed into an animation compatible format (e.g. BVH). This would therefore not be feasible with datasets that do not specify 3D motion.

3 MODEL IMPLEMENTATION FOR THE RHETORICAL DOMAIN

In this section we describe a method of parsing, comparing, and clustering gestures to determine their relationship to rhetorical communicative functions. First, we describe the tools used to compile our dataset. Then, we discuss our specific methods of characterising and comparing motion between gestures. We then describe how we compare the physical characteristics of gestures to feed into a clustering algorithm. Finally, we state our hypotheses underlying

the use of this technique, which combines research on human gestures with computational analysis. For brevity, we only describe our implementation and objective analysis of the rhetorical domain in detail, using it to illustrate the overall approach to the different analysis pathways.

3.1 Input data

We used the pre-segmented motion and video of gestures found in [13] as our gesture dataset. We then used Google Cloud Speech API to retrieve the transcripts that accompany each gesture. We analyzed audio from the entire video to provide context for better speech recognition, then matched the transcript section to each individual gesture based on timestamps of the original gestures and words received by the speech API. We then parsed these transcripts using the CODRA rhetorical parser [18] which is powered by the Charniak re-ranking parser [29] [4]. To do this, we also sent the entire transcript at once - as opposed to an individual sentence or short paragraph that would correspond to a single gesture - to achieve better rhetorical parses. Together, these tools provided a rich dataset of gesture movement and accompanying verbal communication, comprised of 11 speakers and over 500,000 minutes of frontal video.

3.1.1 Splitting gestures. Although in actual behavior, there is ambiguity over what constitutes an individual gesture, we can break them into the phases of the individual gesture and phrases comprised of multiple gestures [30]. For our analysis purposes, the key phase of a gesture is its stroke, which carries the meaning. The stroke phase can vary in length [21]. Gestures in this dataset were between 2 and 250 seconds (60-7500 frames), with longer gestures naturally containing more varied movement. It was therefore necessary to break these into a shorter, more standardized length.

It is common to split gestures based on motion quality [8], however we found this led to splitting gestures in the middle of spoken phrases. This created abruptly segmented and consequently confusing co-speech context for resulting gestures.

As an alternative method, we split the gestures based on the rhetorical parses of the transcripts. This preserves context in particular phrases, however it relies on gestures occurring with their relevant speech at the same time, whereas in reality gestures often precede speech [20]. This also introduces the problem of biasing the gestures towards being relevant for rhetorical parses, as we purposefully split gestures with respect to the rhetoric aspects of speech, as opposed to detecting a shift in semantics or changes in pitch. Furthermore, this segmentation relies on the quality of the rhetorical parser. Alternative implementations could also break gestures syntactically or affectively based on the functional parser, perhaps even splitting differently for different domains.

Splitting large gesture phrases into smaller units fails to incorporate important large-scale rhetorical structure that takes place at the paragraph (or higher) levels. For example, often in speech we reference an idea and give it "space," in our physical surroundings [14]. We may then elaborate on that idea in various ways, referencing the physical space we created for it utterances later [28]. By analyzing gestures as individual units, we knowingly fail to detect high-level rhetorical structure. We consider losing out on high-level rhetorical structure by effectively shortening our average gesture an acceptable trade-off to better cluster motion, as the purpose of this model is to generate relevant and meaningful gestures given distinct utterances, such as a turn of dialog, as opposed to, for example, a speech or lecture.

3.2 Motion Sub-Clustering

After obtaining rhetorical parses and clusters from the input data, we then create sub-clusterings based on motion for each of the rhetorical clusters. This includes determining how best to characterize the motion from keyframe values, as well as how to cluster these gestures once a suitable distance metric is determined.

3.2.1 Characterizing Gesture Motion. In order to cluster gestures by their motion, we developed a distance metric to determine how similar or dissimilar the motions of gestures are, which necessarily works on gestures of differing lengths. We use high-level features to create a descriptive Feature Vector of a gesture[5].

We created a 12-dimensional feature space of motion. This consists of: the maximum and minimum distance of the palms from each other, the maximum and minimum velocity and acceleration of each palm, the distance the hands move together and apart throughout the gesture, the maximum and minimum vertical and horizontal orientation of each palm, and the extent to which the hands cycle, oscillate, and change hand position over the course of the gesture. Note that these features are agnostic to the absolute position of keyframes, instead focusing on relative position between hands, and also put emphasis on two-handed gestures. We then normalized each feature across gestures, and performed K-Means clustering using the Euclidean distance of these feature vectors. The use of these features reduces each gesture down to a single feature vector, which allows gesture comparison across different video conditions.

Although the features chosen are well-documented as meaningful in gesture literature [3] they are by no means exhaustive. This technique also relies on assumptions by the implementer on the relative importance of various features of the gesture, which may be given weights (which may themselves fluctuate based on functional domain). We discuss these limitations further in the Discussion.

3.2.2 Clustering Algorithm. We performed K-Means clustering for each rhetorical functional cluster. We used the scikit-learn [35] implementation of K-Means clustering, and combined each cluster with only one gesture with its next-closest cluster. We also broke apart all clusters in the bottom 10% of silhouette scores, reassigning gestures to the next closest cluster. We determined the optimal number of clusters by running the clustering multiple times to observe the highest silhouette scores.

3.3 Hypotheses

We have illustrated a model with which to generate gestures for virtual agents in real time. The quality of these gestures relies on the assumption that the movements from USG pairs can be captured and meaningfully sub-clustered to obtain a group of gestures with

[4] Current additional implementations not expanded upon in this paper use the VADER sentiment analysis parser [17] for affective parses and Spacy [16] feeding into the TRIPS ontology [38] for semantic parses.

[5] We focus on hand and arm gestures, but this architecture does not preclude analyzing full-body poses or facial gestures if such information was available.

similar motion profiles according to our selected features. Therefore, our analysis of the rhetorical element of the model tests the assumption that after creating functional clusters using tags obtained from the parser, we are able to effectively sub-cluster USG pairs. The alternative to this would be that despite breaking gestures into functional categories, they do not cluster meaningfully using the selected features.

The other hypothesis to be tested is that sub-clustering by motion is in fact necessary and effective to produce communicatively meaningful gestures. It may be the case that gestures may be immediately clustered according to communicative function and naturally form families of similar motion.

4 ANALYSIS AND RESULTS

In this section we discuss our methods of determining the efficacy of our clustering techniques, and the necessity of performing motion sub-clustering in order to generate communicatively meaningful gestures. We define objective metrics with behavioral correlates that evaluate to what extent we can expect the architecture defined above to perform gestures that are relevant to a given utterance.

Using the rhetorical splicing technique described in 3.1.1, we achieved a dataset of 66,529 gestures across 8 speakers. Of these, there were 226 unique rhetorical tags. However, as the parser only provided 20 tags, some of these are sequences of tags. For simplicity, we dropped all gestures with multiple tags (the impact of this is discussed further in Section 5). Additionally, some of these tags do not carry gestural significance (such as the "Nucleus" tag). All such tags were grouped into one cluster with no tag. In the end this produced 43,683 gestures with 15 rhetorical clusters.

4.1 Analysis Technique

We measured sub-clustering quality with the silhouette score with respect to the Feature Vector distance metric described in 3.2.1. The silhouette score measures how well a USG pair fits within its own cluster, compared to others around it. The silhouette score s_i for one USG pair i is defined as:

$$s_i = \frac{b(i) - a(i)}{max(a(i), b(i)}, if \|C_i\| > 1 \quad (1)$$

Where $a(i)$ is the mean distance between i and all other points in its cluster, C_i. This is a measure of how well i fits into its cluster (the smaller the value, the better the assignment). $b(i)$ is the mean value between i and all other points in the next best fit cluster for i (with a higher value meaning a worse fit). This is a proxy for how dissimilar the next-closest cluster is (the between-cluster distance). This score necessarily falls between -1 and 1, with 1 being the best fit. We compute this value for all points in cluster C_i, and use the mean to describe the silhouette score $s(C_i)$ for the cluster. We further describe the score of a clustering as the average silhouette score for all clusters within that clustering.

High silhouette scores indicate a USG pair fits well within its own cluster, and not with the next-closest cluster. While this does lose some nuance of explicitly measuring the distance between clusters and cluster density, it is a useful proxy that is a well-established metric to measure cluster quality in the field of machine learning. This metric is also comprised of behaviorally relevant measurements: between-cluster distance, and within-cluster similarity.

A large between-cluster distance is indicative that the motions in a cluster are distinct from others near it. That is, the motion of those gestures map exclusively to their corresponding rhetorical tag, suggesting such a gesture should only be used when that tag is present or risk being confusing for the viewer.Since no other sub-clusters contain similar motions, the motion will be highly communicatively distinct within that categorical tag. Put another way, when accompanied by an utterance that falls within that functional category, the motion of a cluster that is highly distinct from other sub-clusters is likely to carry meaning.

Sub-clusters with high within-cluster similarity indicate a low variance in performance: a well-defined, specific motion. Such characteristics are relevant in the virtual-agent space because a collection of gestures with very similar movement profiles indicates the potential use of a pre-crafted library of gestures, which can be pre-loaded and run without the heavy computation of generating a completely novel gesture on-the-fly.

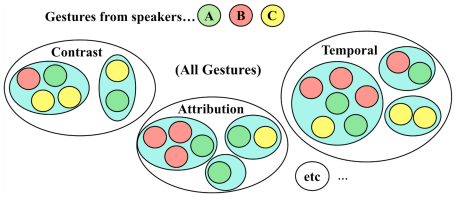
4.1.1 Functional only vs. Functional with sub-clustering. Measuring the quality of the clusters created by only using the functional co-speech elements of the gestures can indicate how well the particular motion of a gesture is relevant for the functional domain. We explored the possibility that it may be possible to skip motion sub-clustering and exclusively use functional clusters to define motion. We obtain silhouette scores for these clusters by using the Feature Vectors of each USG pair in a functional cluster to create a centroid, then measure cluster overlap using distances of these Feature Vectors to their own and other clusters' centroids.

We present two alternatives when collecting metrics: evaluating the quality of the Motion Sub-Clusterings (for example, the sub-clusters only within the *Contrast* cluster, Figure 5a), or evaluating the quality of the Functional Clusterings without sub-clustering (Figure 5b). If silhouette scores are high in the initial functional clustering, then there would be no need for motion sub-clustering as the motions defined in each category may be sufficiently distinct.Notably, these two analyses compare different sets of gestures: the former compares the motion sub-clustering only of USG pairs with a specific functional tag, while the latter compares the motion of all USG pairs by determining cluster quality using clusters defined by functional labels.

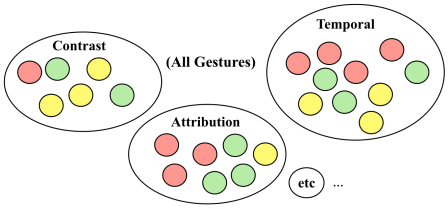
4.1.2 Individual vs. aggregated speaker sets. Finally, we compared the silhouette scores for both clusterings (Functional only, and Functional with Sub-Clustering) with those of individual speakers (Figure 5c). For this, we ran the model on all 8 speakers and found the average silhouette score for motion sub-clusterings and functional clusterings. This allows us to see trends which emerge in individuals that may be obfuscated in a dataset that aggregates all speakers. We then compare this to an aggregated dataset which contains the gestures of all speakers.

We present three evaluations of these two possible clusterings using the silhouette scores: The average silhouette score of individual speakers for functional and motion sub-clustering, the average silhouette score of the aggregated gesture set for function and sub-clustering, and the breakdown of silhouette values for sub-clusterings using the aggregated speaker dataset.

(a) Visual representation of Sub-clustering analysis (aggregated).



(b) Visual representation of Functional analysis (aggregated).



(c) Visual representation of Sub-clustering analysis (individuals).

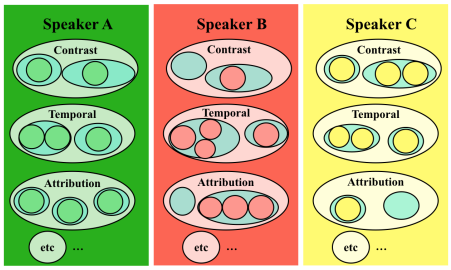


Figure 5: Demonstrations of Sub-clustering or Functional Clustering, and Individual and Aggregated speaker set analysis. Notice how sub-clusterings for individual speakers may result in different numbers of sub-clusters for a functional tag, and that USG pairs for the same speaker may be in the same sub-cluster when analyzed on the aggregate level but not on the individual level. Regardless, the functional tags remain constant.

	Sub-clustering	Functional clustering
Individual speakers	0.317 (0.289)	0.018 (0.153)
Aggregated speakers	0.280 (0.304)	0.009 (0.151)

Table 1: Average and (standard deviation) of silhouette scores of clusterings for individuals and aggregated speakers, for sub-clustering and functional-only clustering (no motion sub-clustering).

4.2 Interpretation of results

The improvement in scores when measuring cluster quality for motion sub-clusters instead of functional clusters (Table 1) indicates that clustering by motion is necessary after functional clustering. This firmly confirms our hypothesis that functional tag by itself is not enough to define a gesture, and rejects the alternative proposed in 3.3 that motion sub-clustering may be unnecessary. This makes intuitive sense from the "important or trivial," example as this is reflective of a phrase that may be gestured in very different ways depending on communicative function. By clustering by motion after clustering by functional tag, we separate out these very different motions and begin to establish correlational links across and within functional domains.

Rhetorical Tag	$N_{gestures}$	$N_{clusters}$	Silhouette score
Span	20622	88	-0.25 (0.420)
Elaboration	7269	26	**0.621** (0.265)
Attribution	5795	17	**0.680** (0.275)
Joint	3404	11	0.413 (0.186)
Temporal	2081	15	0.314 (0.276)
Same-Unit	1988	88	**0.713** (0.249)
Cause	667	9	**0.685** (0.278)
Enablement	472	3	0.099 (0.391)
Background	418	27	0.280 (0.263)
Condition	360	10	0.176 (0.347)
Contrast	250	12	0.436 (0.251)
None	183	3	-0.018 (0.320)
Comparison	96	8	0.045 (0.282)
Manner-Means	52	5	0.019 (0.305)
Explanation	26	2	-0.014 (0.331)

Table 2: The breakdown of sub-clustering scores for each rhetorical tag when using aggregated speaker set. Number of gestures, number of motion sub-clusters, and mean and (standard deviation) of silhouette scores for sub-clusters. Selected scores over threshold of 0.6 in bold.

Furthermore, the relatively high silhouette scores for some rhetorical categories (Table 2) indicate that the selected features do effectively distinguish between motions within a particular rhetorical structure for some rhetorical tags, and that these motions are distinct within a rhetorical structure. These results establish moderate support for our hypothesis that clustering gestures by rhetorical structure of corresponding co-utterances, then sub-clustering by motion properties within those categories creates well-structured clusters that are relevant to the rhetorical category.

The high variation in clustering quality between rhetorical tags (Table 2) suggests that some structures do not have consistent canonical forms that are neatly captured by the features we used. Some of these are surprising (such as the *Comparison* cluster) and challenge assumptions about what features may be relevant for a particular gesture.

The higher average sub-clustering silhouette scores for individual speakers (Table 1) suggests this method is somewhat better suited to modeling gestures of individuals than an aggregated group, although this difference is small. This is consistent with individuals having strong tendencies to gesture in a particular style[13]. Still, the small differences in average sub-clustering scores (Table 1) and per tag values between 0.50-0.70 (Table 2) together suggest that this method is still effective when applied to aggregated speakers.

Another possibility is the higher scores for individual speakers are an artifact of different speaking conditions. While all videos are frontal views of the speaker, slightly varying angles as well as the relative size of the speaker in view makes even relational changes in position of keypoints imprecise across different video conditions. For example, if one speaker's videos lead them to be larger in relation to the overall space, their relative hand variation will look larger as well. However, this result is surprising as we

would expect that if individuals gesture in a consistent manner, those motions would form a distinct sub-cluster within the larger aggregated set, and furthermore we would expect this effect to be exacerbated by variation in video settings.

5 DISCUSSION

That some sub-clusterings achieve high silhouette scores (Table 2) indicates not only that there are many ways in which individuals gesture during any particular phrase, but also that rhetorical structure is indeed relevant to those motions. While there are many ways to gesture for a particular phrase, there is a limited number of families of gestures that many individuals tend to use. This is an argument for the appropriateness of gesture libraries, as one could create a specific animation for each sub-cluster (as discussed in 2.4).

One finding was similar silhouette scores for individuals and the aggregated group of speakers (Table 1). This suggests that although individuals have their own distinct and precise style of gesturing, resulting in a more effective clustering of their motions, there are some linguistic circumstances under which individuals tend to use similar motions to convey certain communicative functions.

One improvement that may further improve scores for motion sub-clustering is to explore a wider variety of motion features to map to rhetorical tags. Clusters with poor scores may perform better if we calculated motion by different features. This highlights how this architecture is not only a functional mechanism to produce gestures but may in the future also be used to test hypotheses of which features correlate to which rhetorical structures – or other high-level linguistic dimensions. Furthermore, a hybrid or weighted-feature system (particularly with automated techniques to derive weights) to determine feature vectors may lead to improvements in this domain, as certain features play a role more heavily in some communicative functions than in others. Some features may be used in the clustering of one functional domain and not in another. Further analysis must be done within each domain to determine how these features may interact to distinguish the roles a gesture plays with respect to each communicative function.

We encountered a variety of challenges which may be overcome to achieve better clustering and model performance. Constructing a dataset which is appropriate for this mapping presents the largest obstacle. There are currently few large-scale datasets of natural social motion. Although development of recent technologies has made scraping motion from video data easier [4], these are still too noisy to effectively track certain meaningful aspects of gesture, such as precise changes in hand shape. Current datasets also do not have transcripts which accompany motion, leaving the transcription task to other third-party programs which can be error-prone and lead to difficulty for rhetorical and semantic parsers. Parsers themselves may also be improved through being trained on social conversation.

5.1 Future Work

While it is reasonable to expect that semantic and affective domains will see similar results to this domain, that assumption must first be tested. This process will also help identify relevant motion features to these different functional domains. Whereas we have implemented analyses of these domains, their mapping to motion has not been explored. This implementation also purposefully excludes rhetorical structures that occur at the paragraph or conversational level. Future analyses may explore this using different functional parsing mechanisms in combination with new motion features to provide a more holistic analysis of a wider range of input utterances.

Another avenue would be exploring the specificity of allowing clusters with multiple rhetorical tags. While these could potentially create more relevant or cleaner clusters, our initial analyses found that in practice this created hyper-specific clusters with only one gesture. Allowing multiple tags also raises a question of confidence in domain parsers; multiple tags may reveal ambiguity of the parser's analysis as opposed to specificity, and counter-intuitively lower the silhouette scores.

Although we have described one method for selecting gestures and quantitatively assessed it, a subjective analysis remains for future work. This will specifically involve crowd-sourcing opinions on traditional metrics such as naturalness of a gesture, clarity of the gesture's message, and the perceived meaningfulness of the gesture, in order to determine how well this algorithm does at selecting gestures which should then, more importantly, be tested in real-world virtual agent implementations. This inspires a variety of questions, including how well functional vs. sub-clustering selection perform on subjective metrics, such as naturalness, coherence, and appropriateness with respect to speech. While motion sub-clustering objectively produces a more consistent family of motions, whether or not human observers understand and enjoy viewing those motions remains to be seen.

The dataset used was also across a wide variety of subjects and speaker types. A subsequent experiment would be to train this model on a domain-specific set of videos in a controlled conversational setting, such as found in [11]. The creation of such a dataset could further ensure high-quality motion and speech capture through use of motion capture technologies, high-quality audio equipment, and human transcription quality control.

6 CONCLUSION

In this paper we demonstrated a new approach through which to view the relationship between gestures and their associated utterances. We presented a novel method to map gestural motion to the gesture's co-speech properties by forming clusters based on motion properties within clusters based on communicative function. We described how an agent could make use of such a model as a generative mechanism to create socially appropriate gestures on-the-fly in conversation, and described our implementation and evaluation of the rhetorical functional domain. Our analysis finds that some rhetorical structures are often accompanied by similar gesture performances, while others are not well-defined by simple motion features. Finally, we discussed the challenges and limitations of our architecture and suggest future improvements to address them, and propose subjective evaluations building on these findings.

ACKNOWLEDGMENTS

The work in this article has been supported by United Kingdom Research and Innovation through the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents (EP/S02266X/1) and by the Royal Wolfson Society Award (WRM/FT/170004).

REFERENCES

[1] James Allen, Myroslava O Dzikovska, Mehdi Manshadi, and Mary Swift. 2007. Deep linguistic processing for spoken dialogue systems. In *ACL 2007 Workshop on Deep Linguistic Processing*. 49–56.

[2] Janet Beavin Bavelas. 1994. Gestures as part of speech: Methodological implications. *Research on language and social interaction* 27, 3 (1994), 201–221.

[3] Geneviève Calbris. 2011. *Elements of meaning in gesture*. Vol. 5. John Benjamins Publishing.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

[6] Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

[7] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*. Springer, 127–140.

[8] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 781–788.

[9] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: a deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*. Springer, 152–166.

[10] Alan J Cienki and Jean-Pierre Koenig. 1998. Metaphoric gestures and some of their relations to verbal metaphoric expressions. *Discourse and cognition: Bridging the gap* (1998), 189–204.

[11] Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. 2010. Seeing is believing: body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 1–9.

[12] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 93–98.

[13] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[14] Susan Goldin-Meadow, Howard Nusbaum, Spencer D Kelly, and Susan Wagner. 2001. Explaining math: Gesturing lightens the load. *Psychological science* 12, 6 (2001), 516–522.

[15] Joseph Grady. 1997. Foundations of meaning: Primary metaphors and primary scenes. (1997).

[16] Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 1373–1378.

[17] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[18] Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics* 41, 3 (2015), 385–435.

[19] Adam Kendon. 1972. Some relationships between body motion and speech. *Studies in dyadic communication* 7, 177 (1972), 90.

[20] Adam Kendon. 2000. Language and gesture: Unity or duality. *Language and gesture* 2 (2000).

[21] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[22] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217.

[23] Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics* 26, 4 (2009), 393–449.

[24] Margaux Lhommet and Stacy C Marsella. 2013. Gesture with meaning. In *International Conference on Intelligent Virtual Agents*. Springer, 303–312.

[25] William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

[26] Daniel Marcu. 1997. The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. 96–103.

[27] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual Character Performance from Speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Anaheim, California) *(SCA '13)*. ACM, New York, NY, USA, 25–35. https://doi.org/10.1145/2485895.2485900

[28] Steven G McCafferty. 2004. Space for cognition: Gesture and second language learning. *International Journal of Applied Linguistics* 14, 1 (2004), 148–165.

[29] David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 337–344.

[30] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[31] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[32] Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*. Springer, 222–235.

[33] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2009. Greta: An interactive expressive ECA system. *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2* 2, 1399–1400. https://doi.org/10.1145/1558109.1558314

[34] Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajic, and Zdenka Uresova. 2015. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 915–926.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[36] Frank E Pollick, Helena M Paterson, Armin Bruderlin, and Anthony J Sanford. 2001. Perceiving affect from arm movement. *Cognition* 82, 2 (2001), B51–B61.

[37] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9 (2018), 1144.

[38] Naushad UzZaman and James Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. 276–283.