

A Comprehensive Review of Data-Driven Co-Speech Gesture Generation

S. Nyatsanga¹ T. Kucherenko² C. Ahuja³ G. E. Henter⁴ M. Neff¹

¹University of California, Davis, USA

²SEED - Electronic Arts, Stockholm, Sweden

³Meta AI, USA

⁴Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

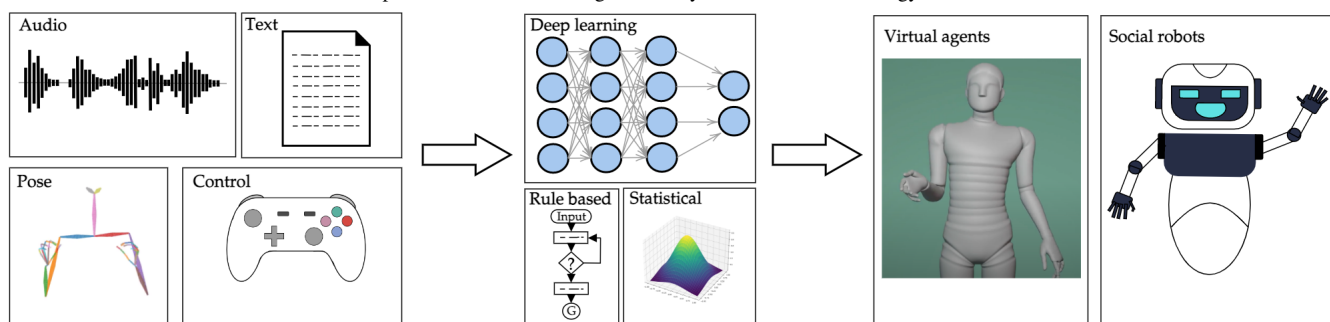


Figure 1: Co-speech gesture generation approaches can be divided into rule-based and data-driven. Rule-based systems use carefully designed heuristics to associate speech with gesture (Section 4). Data-driven approaches associate speech and gesture through statistical modeling (Section 5.2), or by learning multimodal representations using deep generative models (Section 5.3). The main input modalities are speech audio in an intermediate representation; text transcript of speech; humanoid pose in joint position or angle form; and control parameters for motion design intent. Virtual agents and social robotics are the main research applications, although also compatible with games and film VFX.

Abstract

Gestures that accompany speech are an essential part of natural and efficient embodied human communication. The automatic generation of such co-speech gestures is a long-standing problem in computer animation and is considered an enabling technology for creating believable characters in film, games, and virtual social spaces, as well as for interaction with social robots. The problem is made challenging by the idiosyncratic and non-periodic nature of human co-speech gesture motion, and by the great diversity of communicative functions that gestures encompass. The field of gesture generation has seen surging interest in the last few years, owing to the emergence of more and larger datasets of human gesture motion, combined with strides in deep-learning-based generative models that benefit from the growing availability of data. This review article summarizes co-speech gesture generation research, with a particular focus on deep generative models. First, we articulate the theory describing human gesticulation and how it complements speech. Next, we briefly discuss rule-based and classical statistical gesture synthesis, before delving into deep learning approaches. We employ the choice of input modalities as an organizing principle, examining systems that generate gestures from audio, text and non-linguistic input. Concurrent with the exposition of deep learning approaches, we chronicle the evolution of the related training data sets in terms of size, diversity, motion quality, and collection method (e.g., optical motion capture or pose estimation from video). Finally, we identify key research challenges in gesture generation, including data availability and quality; producing human-like motion; grounding the gesture in the co-occurring speech in interaction with other speakers, and in the environment; performing gesture evaluation; and integration of gesture synthesis into applications. We highlight recent approaches to tackling the various key challenges, as well as the limitations of these approaches, and point toward areas of future development.

Keywords: co-speech gestures, gesture generation, deep learning, virtual agents, social robotics

CCS Concepts

• **Computing methodologies** → **Animation**; *Machine learning*; • **Human-centered computing** → *Human computer interaction (HCI)*;

1. Introduction

This paper summarizes research on the synthesis of gesture motion, with a particular emphasis on more recent techniques using deep learning. The focus is on *co-verbal gesture*, gesture that accompanies speech. When considering the problem, a first reasonable question is “Why should we care about gesture at all?” Gesture plays at least three main functions. First, and most simply, it helps artificial agents and robots look more alive and be more engaging (this has been shown multiple times, e.g. [SKW*12, SER*13, BDK16, SN19]). Second, it communicates functional information. This can include pointing or deictic gesture that establish reference; emblems that replace words, and imagistic metaphoric and iconic gestures that illustrate concepts and artifacts. Third, gesture communicates social information, including personality [SN17, NAW10, NTB*11, DKD*16], emotion [VMD*14, XBHN13, GFD*15, NLK*13, DGS13, FP16, CN19] and subtext.

Before summarizing work on gesture synthesis, it is worthwhile to consider how gesture can support a range of applications for virtual agents and robots. First, it is well established that gestures do indeed communicate [GM05, Ken94, Hos11, GM99] Hostetter’s meta-analysis [Hos11] presents three main findings for when gestures communicate: gestures depicting motor actions are more communicative than those depicting abstract topics; gestures that are not completely redundant have a larger impact on communication, and children benefit more from gesture than adults.

Gestures communicate in a different manner than spoken language. They communicate particularly directly when being used to describe spatial concepts or object manipulation because there is a natural iconicity to these concepts, which is well portrayed in gestures. “Gesture permits the speaker to represent ideas that are compatible with its mimetic and analog format (e.g. shapes, sizes, spatial relationships) - ideas that may be less compatible with the discrete and categorical format underlying speech. Thus, when it accompanies speech, gesture allows speakers to convey thoughts that may not easily fit into the categorical system that their spoken language offers.” [GMM99]. The iconicity of gestures makes them more transparent than language, which is purely symbolic. Tversky argues that “[gestures] take advantage of human ability to make rapid spatial judgments and inferences. Neither depictions nor gestures can convey all the information surrounding an idea or set of ideas; this forces them to extract what is essential, to simplify the ideas, making them easier to comprehend and remember.” [Tve07]. They provide an additional code, a motor code for information, and additional codes are known to improve memory. Gestures are particularly congruent with actions or transformations [Tve07]. A more detailed gesture typology is presented in Section 2.

Nonverbal communication appears to be particularly important in providing appropriate social cueing [Whi03]. Feelings, emotions and attitudes are often not made verbally explicit and must be inferred from nonverbal channels. The presence of nonverbal communication can radically change the outcome of an exchange. For example, a study comparing face-to-face and voice only union negotiations showed greater interpersonal communication in the face-to-face setting, whereas the speech-only communication focused more on content, was more impersonal, saw reduced praise, greater

blame, more disagreement and was more likely to end in deadlock [RSD81].

Additionally, gestures can be used to regulate a dyadic or group interaction by managing turn-taking [Whi03, Kip05]. Kipp defines turn-taking as “assigning, holding or yielding a turn” in a dialog [Kip05]. Bavelas [Bav94] identified so-called “turn gestures” in dialogue interactions, with sub-categories of gestures that indicate: giving turn to the other speaker, accepting a turn from the other speaker or offering turn to any speaker in a group.

Gesture can also support particular applications, for example, there is a growing body of evidence showing that gestures help people learn. There are at least three mechanisms by which this happens: learners watching a teacher gesture, learners performing gestures themselves, and teachers adapting their instruction based on information gained from the learner’s gestures. For an excellent summary, see [NGM15]. Gestures can link abstract concepts in the immediate environment, reduce cognitive load and enhance spoken communication [NGM15]. A recent meta-analysis [Dav18] of twenty experiments showed that the gesture already present in pedagogical agents is beneficial for student learning, with positive effects on transfer of knowledge, retention of learning and agent persona, but not on reducing cognitive load.

Given the potential value of nonverbal communication, the question remains as to how to synthesize appropriate behavior in our computational systems. Formally, the problem is to generate an input to output mapping, where the input is some representation of the content to be expressed and the output is some representation of the behavior to perform. Most learning systems to date have assumed audio and/or text as the input, although we will see this has limitations. The output is generally either frames of pose data or a lexicalized representation of the gestures that should appear (e.g. accompany this sentence with a conduit gesture that displays the idea of conveying something to the interlocutor). The latter must then be converted to animation using some secondary system.

Previous surveys have covered co-speech gesture synthesis with varying scope and emphasis [WHTL22, LMSJ21, YWJ21]. Wei et al. [WHTL22] focus on audio-visual learning for human-like machine perception, and briefly cover gesture synthesis. We focus on co-verbal gesture synthesis, including its theory, synthesis techniques with an emphasis on deep learning, and an eye towards application to virtual agents and robots. Ye et al. [YWJ21] surveyed deep-learning-based human motion modeling, and thus is related to our work via the emphasis on deep generative models. Our survey covers a larger scope of deep-learning-based gesture synthesis research and offers a more comprehensive set of key challenges that are specific to the problem. The survey by Liu et al. [LMSJ21] is more closely related to our work, emphasizing co-verbal gesture generation for virtual agents and robots. Their work presents a scoping review of data-driven techniques for speech-to-gesture generation, related datasets, and evaluation metrics. We include a larger set of papers (40 vs. 19) and are able to provide a more in depth treatment of the technical material given the substantially longer STAR format.

Overall, our survey makes the following contributions to the field:

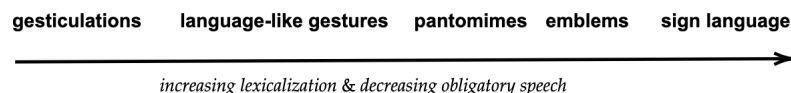


Figure 2: *Kendon’s Continuum of gesture categories, as described by McNeil [McN92a, McN05]*

- A detailed discussion on the theory and motivation for co-verbal gesture synthesis.
- A discussion on rule-based and statistical techniques, illustrating how these approaches can complement the strengths and weaknesses of recent deep-learning approaches.
- An emphasis on deep-learning-based generation systems using input modality as an organizing principle for the research.
- A discussion on the most commonly used speech-to-gesture datasets, collected via motion capture or pose estimation.
- Identifying and detailing a set of key challenges for co-verbal gesture synthesis and potential research directions.

The remainder of the paper begins by providing a deeper background on gesture, followed by a summary of synthesis techniques that have been developed to date and concludes with a discussion of major open problems.

2. Human gesticulation

Manual Gestures are non-verbal, non-manipulative hand/arm movements that occur during speech [McN92a, Tui93, Ken83]. We will refer to manual gestures as simply “gestures” in this work, although gestures can in general be performed by other body parts, such as the head. Gestures aid in the communicative intent and are closely linked to accompanying speech in terms of timing, meaning and communicative function. For instance, gestures can be used for pointing to resolve references to objects (“what is that”) or illustrate concepts that would otherwise be difficult to explain verbally [Kip05]. Therefore, gestures play an important complementary role to speech because they enable broader and more efficient expression of personality, emotion and motivation of the speaker [EF69, Meh68, Ekm92]. Additionally, gesture plays an important cultural role, because members of a community can either identify with or easily understand the emotions and attitudes of those around them through these non-verbal cues [BL93].

Gestures can take many forms depending on the speaker, and the morphological rules governing their construction equally vary. Based on Kendon’s gesture categorization [Ken88], McNeill proposed “Kendon’s Continuum” [McN92a, McN05] where gesture categories are sorted in increasing *lexicalization*, that is the degree to which they adhere to formal, language-like grammatical rules, as illustrated in Figure 2. In this framework, the least lexicalized (conversational gesture) have obligatory speech, while the fully lexicalized (sign language) have little or no obligatory speech as the gestures themselves gain explicit lexical properties. Most importantly, the fully lexicalized end of the spectrum (sign languages) have formal syntactic structure like spoken languages, but this is absent from verbal gesticulations. This lack of structure is one of the challenges in producing verbal gesture as the behavior can be highly idiosyncratic.

There are many different forms of gesture and McNeill [McN92b, McN05] argues for a dimensional view in which the dimensions are iconic (images of the concrete), metaphoric (images of the abstract), deictic (pointing) and beat (flicks of the hand in time to rhythm of the speech). See Figure 3. An iconic gesture might show the size of a box being discussed by drawing it in space, whereas a metaphoric gesture might indicate an abstract concept, such as all ideas are included, by making an umbrella shape. A given gesture may load on multiple of these dimensions, for example displaying both iconicity and deixis. An additional category are adaptors (or self-adaptors), which are self-manipulations such as scratching one's nose or bracing fingers. These are not designed to communicate, but do convey information about personality [NTB* 11].

Kendon introduced a three-level hierarchy to describe the structure of gestures [Ken72]. The top level is the *gesture unit*. Gesture units start in a rest pose, contain a series of gestures, and then return to a rest pose. The starting and ending rest pose need not be the same. A *gesture phrase* encapsulates an individual gesture in this sequence. Each phrase can in turn be broken down into a sequence of *gesture phases*. These include: a *stroke*, which is the main meaning carrying movement of the gesture and has the most focused energy; a *preparation*, which is a motion that takes the hands to the required position and orientation for the start of the stroke; a *prestroke hold*, which is a period of stillness in which the hands are held at the starting point of the stroke, before the stroke begins; a *poststroke hold*, in which the hands are held at the end position of the stroke; and finally, a *retraction*, that returns the hands to a rest pose. All phases are optional except the stroke. The pre- and poststroke holds function to synchronize the gesture with speech.

There are many challenges for automatically synthesizing gesture, for instance to drive virtual agents in human-computer interaction. One theory on the origins of gesture, the growth point hypothesis [McN92a], argues that gesture and language emerge from a common communicative intent. Some communication may take verbal form and some nonverbal, with some being replicated across both. Some agent architectures, such as SAIBA [KKM*06], have tried to model this communicative intent. This allows non-verbal communication to be unique, carrying different information than the verbal channel. Many gesture synthesis approaches, and all deep learning approaches that we are aware of, do not model a communicative intent. Instead, they synthesize gesture from audio, text or both. This necessarily means that the gestures will be redundant with these other channels, and thus more limited than actual human gesture.

Another challenge is that gesture is idiosyncratic [McN00], so different people may gesture in very different manners. The same person may also generate different gesture even when delivering

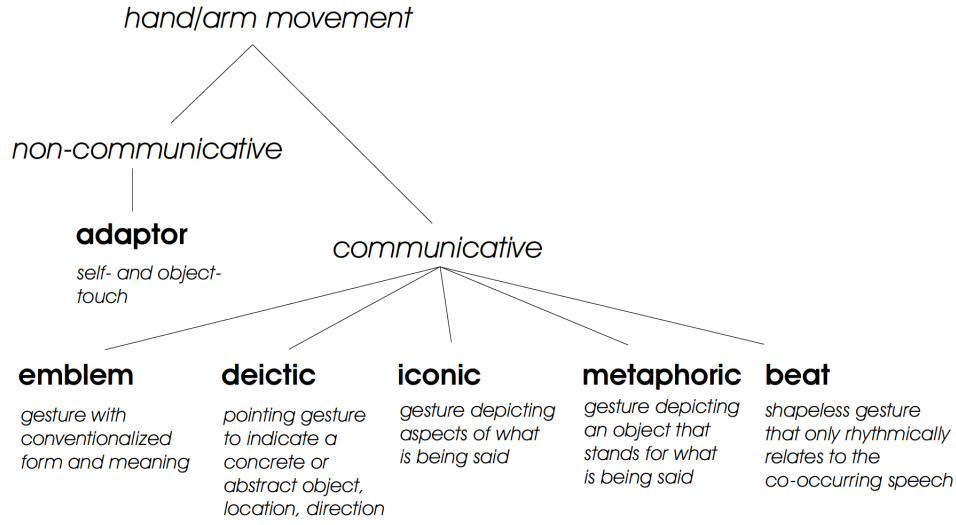


Figure 3: Relational graph of gesture categories and their defining properties. Figure from Kipp [Kip05] (used with permission)

the same text. Finally, gestures are synchronized in time with their co-expressive speech. About 90% of the time, the gesture occurs slightly before the co-expressive speech [Nob00] and rarely occurs after [Ken72]. While the earlier occurrence of gesture is common in human behavior, and research on animated characters also indicates a preference for this slightly earlier timing, it also indicated that people may not be particularly sensitive to errors in timing within a ± 0.6 seconds [WN13].

3. Approaches for gesture synthesis

Synthesizing co-speech gestures is essential for creating interactive and believable virtual characters in human-computer interaction (HCI), graphics and social robotics. Thus significant effort has been applied and progress has been made for applications in virtual agents [CVB01, PB03, KW04, NKAS08, CMM15] and humanoid robots [SKWJ09, YKJ*19, LHP11, GHLY18, NTHLO10, HS16]. Neff identifies the two main sub-problems of generating gesture as the *specification problem* and the *animation problem* [Nef16]. The specification problem is concerned with determining *what* gestures are to be performed by the character, and the animation problem entails *how* to generate the appropriate hand motion. Gesture specification can use a range of inputs including speech, prosody, text and communicative intent, where rule-based, statistical and learning-based models have been used to determine the appropriate gesture. Similarly, gesture animation has used a range of procedural, physics-based or learning-based models to produce the hand motion.

Gesture generation models can be divided into two main categories: rule-based and data-driven. Within the latter, there are two sub-categories, statistical and learning-based. Rule-based systems [CVB01, PB03, KW04, SKWJ09, CPB*94, KW02, PCDC*02, LM06, TMMK08, MXL*13] use carefully designed heuristics to select the appropriate gesture for a given speech input. Data-driven

systems instead learn to associate speech with corresponding gestures from the data, and we expound on the sub-categories next. Statistical systems [NKAS08, Kip05, BK09a, YYH20] typically precompute probabilities or assign a prior distribution over the given gesticulation data and a gesture is sampled from the distribution based on speech input. Learning-based models [CMM15, YKJ*19, LKTK10, CM11, HKS*18, AMMS19, FM18, KHH*19, GBK*19, ALNM20, AHKB20, KJvW*20, LKP*21, FNM19] make the fewest assumptions about the distribution of gesticulations and instead optimize the parameters of a complex non-linear function to map the input speech into the appropriate gesture. This non-linear function is usually implemented as a deep neural network with some form of a gradient-based optimization algorithm, and thus we simply refer to them as deep learning approaches. While animation may be synthesized using a range of methods for each technique, rule-based and statistical approaches have generally predicted a gesture label that is used to index either hand-animated or pre-recorded gesture clips that are then used to synthesize the final sequence. In contrast, deep-learning approaches have tended to synthesize motion on a per-frame basis.

Figure 4 illustrates the development of the gesture generation field, specifically how different approaches handle the trade-off between naturalness and communicative efficacy. The early approaches were intent-driven and hence had high communicative efficacy [CVB01, KW04, TMMK08]. They were not very natural, since they were mainly inserting pre-defined animations. Later approaches used statistics to analyze and retrieve gestures from large databases [NKAS08, Kip05, BK09a]. Statistical approaches improved gesture naturalness, while slightly compromising communicative efficacy. Finally, modern approaches are mainly deep-learning-based, making the fewest assumptions about the underlying distribution of gesture data [KHK*21, ALNM20, YCL*20]. Deep learning-based approaches can generate continuous and fairly natural gestures, but they are significantly less communicative. Mo-

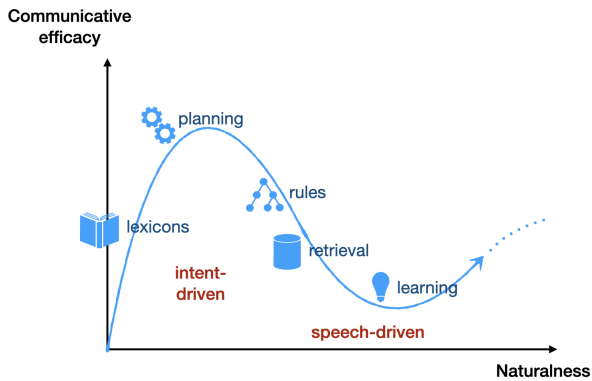


Figure 4: Overview of the development of the gesture generation field, as outlined by Stefan Kopp at the GENE Workshop 2020. Figure by Stefan Kopp. Used with permission.

tivated by this challenging trade-off, recent notable research has proposed hybrid systems for generating natural and semantically meaningful gestures, by combining rule-based and deep learning-based approaches. [KNN*22, ZBC22, HES*22]. We review the seminal approaches in rule-based, statistical and learning-based generation next. In Section 4, we discuss what, in our estimation, are some of the most impactful, rule-based systems. We discuss them in chronological order for ease of understanding and to emphasize how the approaches influenced one another. The selected works have in common that they pioneered approaches for speech-driven hand or facial animation by devising heuristics and domain-specific languages for modeling behavior intent, planning, and realization. Since the focus of the paper is on data-driven systems, we only review these selected works. For a more detailed review of rule-based systems, we recommend the review article by Wagner et al. [WMK14].

4. Rule-based approaches

Cassell et al. [CPB*94] presented *Animated Conversation*, the first rule-based system to automatically generate context-appropriate hand gestures, facial movements and intonation patterns between multiple human-like agents. Notably, this work was one of the first to explore the latent relationship between speech and gesture for generating realistic animation. The system initiated a dyadic interaction between two agents through a dialogue generator and planner. The generated text was transformed into speech through a text-to-speech system [LB85] and deep symbolic representations were used to encode timing, intonation and the corresponding gesture prototypes. The gesture prototypes were used to perform the full gesture. The result was agents with appropriate and well-synchronized speech, intonation, facial expressions and hand gestures. However, the system was limited to domain-specific dialogue generation between two agents, which not only restricted free-form conversation (by restricting the discourse) and gesture animation, but also precluded real-time interaction with a human user.

Thórrisson proposed *Ymir*, [Th696] which improved on the Animated Conversation framework by enabling multimodal input from

a user, including speech, gaze, gesture and intonation. It consisted of multiple modules for input perception, dialogue generation, decision making and action schedulers in order to produce well-synchronized hand animation. However, although this offered more interactivity with a user, the system could only produce limited multi-modal output in real time. The work of Cassell et al. [CBC*00] subsequently improved on the two frameworks by integrating the real-time multi-modal interactivity of Ymir with the symbolic generation and richer multi-modal synthesis capability of Animated Conversation. The result was an embodied conversational agent framework that produced reactive characters that behaved intuitively and robustly in conversations, albeit still limited to dialogue deriving from a static knowledge base.

Another of the seminal works in a rule-based generation was the Behaviour Expression Animation Toolkit proposed by Cassell et al. [CVB01]. BEAT took typed text as input and could synthesize well-synchronized speech, gesture, facial animation and intonation. The system used contextual information latent in the text to choose pre-recorded hand, arm and facial movements by relying on a set of carefully designed heuristics from previous nonverbal conversational behavior research. BEAT was highly extensible allowing animators to insert new rules that parameterize personality, movement characteristics, scene constraints and desired animation style.

Alternatively, Kopp et al. [KW02] proposed a model-based approach for generating complex multimodal utterances (i.e. speech and gesture) from XML specifications of their form. Instead of relying on pre-recorded gestures, as the previously discussed approaches did, the system applied non-uniform cubic B-Splines to form a gesture trajectory that satisfies all velocity and position constraints. The authors demonstrated the multimodal capabilities of the system through Max: a virtual reality-based agent that interacts and assists a human user through construction tasks, by using prosodic speech, deictic and iconic gestures, gaze and emotive facial expressions [KJLW03].

Facial expressions, gaze direction and head movements are essential non-verbal behaviors that communicate the intent and emotional state of a speaker. They can also act as facial gestures e.g. “raised eyebrow” or gaze direction utilized to resolve a referent object or direction. Therefore, endowing virtual agents with such qualities can make them more anthropomorphic. Pelechaud et al. [PCDC*02], developed Greta: a 3D virtual agent whose facial gestures communicated the agent’s emotional state. The system was designed as a BDI agent (i.e. prior Beliefs, Desires and Intentions) [RG91]. A Dynamic Belief Network (DBN) modelled Greta’s constantly evolving emotions and computed the triggering thresholds and evolution of her emotions, resulting in emotive verbal and non-verbal behaviour.

The development of new rule-based systems often necessitated the development of a new domain-specific language (DSL), usually based on XML. Examples of these include an XML processing pipeline in the BEAT system [CVB01], MURML for multimodal behavioral planning and animation of Max [KJLW03, KKW02], APML for representing the agent’s behaviour semantically [CPPS04], and RRL for representing simulations of multimodal dialogue [PKS*04]. However, these DSLs were often incompatible with each other even as the systems solved similar

or overlapping objectives. As a result, a group of researchers developed a unified language for multimodal behaviour generation for virtual agents, called the Behavior Markup Language (BML) [KKM*06, VCC*07]. BML was designed in the context of a comprehensive framework with intent planning, behavior planning and behavior realization stages. Within this framework, BML described the desired physical realization and thus connected behavior planning to behavior realization. BML became the standard format for rule-based systems, finding use in open-source frameworks like SmartBody [TMMK08], and other agent embodiments like humanoid robots [LHP11].

The development of BML led to continued advances in rule-based systems, even as some research started to explore learning-based systems. For instance, Marsella et al. [MXL*13] generated facial expressions and behaviors (including gestures, head movements, eye saccades, blinks and gaze), for a 3D virtual character, by analyzing the prosodic and acoustic features of speech, as well as shallow analysis of the utterance text to determine rhetorical and pragmatic content. Ravenet et al. [RPCM18] generated metaphorical gestures by leveraging BML to extract metaphorical properties from input speech. Their system leveraged BML annotations to synchronize speech audio and gestures, and configure gesture characteristics (e.g., hand shape, movement, orientation) to convey the desired representational meaning during behavior realization. Overall, BML continues to be a standard domain-specific language for behavior planning and realization in rule-based gesture generation systems.

Rule-based gesture generation systems can produce high-quality gestures that are well synchronized with speech. Due to their reliance on pre-recorded motion, hand animation or carefully engineered systems for generating gestures, rule-based systems can have better motion quality than learning-based systems. Hand-tuned rules may also better preserve semantics within their limited domain. However, the gesture distribution is often not diverse. Moreover, the carefully designed rules require significant expert knowledge which is laborious and not scalable. Such systems are inflexible in that they can only produce a small set of plausible gestures for a particular speech input or scenario. Therefore, the inability to produce diverse gestures in a non-deterministic manner means the resulting virtual agents (or any other embodiments) can only behave in an expressive and naturalistic way for limited examples. Data-driven methods were proposed to try to overcome these limitations. Given the overall advances in deep learning, they may eventually also produce the highest quality motion. We review the two data-driven sub-categories next, statistical and deep learning-based methods.

5. Data-driven approaches

5.1. Speech-Gesture Datasets

Any data-driven method is fundamentally limited by the data it is trained on. The number of datasets suitable for machine-learning on human gesture data has been steadily rising, as has their size. Table 1 provides an overview of major datasets for gesture generation, and their characteristics such as size, motion format (2D or 3D), modalities, included annotation, and more. It is seen that

the dataset sizes have reached new heights of 100+ hours in recent years, and there is also greater diversity in terms of the number of speakers, thanks to 3D pose estimation from video. Unfortunately, only a small fraction of datasets contain high-quality finger motion, which is of great importance for generating expressive and meaningful gestures.

There are two main methods for obtaining motion data for gesture synthesis: optical motion capture [MYL*16, TKS*17, LDM*19, JSCS19, FM18] or pose estimation from monocular video [YKJ*19, ALIM20, JKEB19, KNN*22, HXM*21].

Existing datasets recorded using motion capture are usually smaller, since that method of data collection is much more expensive and labor-intensive, and generally takes place in a controlled studio environment. Emotion is often acted. The main advantages of the resulting data is that movements are in 3D and have high quality. This method is also the best at capturing finger motion.

Datasets instead obtained from pose estimation can be an order of magnitude larger, as they can be sourced from online videos. This enables finding genuine, in-the-wild gestures, and the material can be large enough to include much more diversity. The downsides are relatively lower motion quality (fingers being especially hard) and being limited to 2D motion only. Recent monocular video work has lifted the skeleton motion to 3D [HXM*21].

In practice, the amount of data needed is likely to depend on the application at hand. While gathering data from a specific target speaker of interest is usually better than having an equivalent amount of data from non-target speakers, the gesture manifolds of different speakers nonetheless often have a significant overlap. It has been found that, starting from a generative model trained on one individual style, one requires only two minutes of data to fine-tune a gesture generation model for another style [ALM22]. Recent work has also demonstrated the possibility of learning to embed different gesture styles, which then can be used for zero-shot adaptation to the style of an unseen target speaker with no training data of the target speaker [FGPO22, GFH*]. Techniques for augmenting gesture data so as to increase the amount of motion data for training have also been studied, especially mirroring [WTGM22b, WTGM22a].

5.2. Statistical and early machine learning approaches

In statistical systems, the latent relationship between speech and gesture is modeled by the statistics of the underlying gesture distribution, instead of being encoded by an expert. Compared to rule-based systems, statistical approaches make fewer assumptions about the speech-gesture association, instead either pre-computing conditional probabilities for the gesture data or assigning a prior probability distribution. Similar to our approach in Section 4, we focus on a subset of statistical approaches that, in our estimation, are some of the most impactful in the field. The works are described in chronological order to illustrate advances in statistical systems.

Kipp proposed one of the earliest statistical systems, which modeled an individual's gesture by analyzing an annotated co-speech dataset and producing a *gesture profile* [Kip05]. The data was annotated using the video annotation tool ANVIL [Kip01] to define a gesture profile consisting of individual properties such as handedness, timing and communicative function. The gesture profiles were

Name	Size	# of sp.	Mot. format	Modalities	HQ fing.	Dialog?	Link
IEMOCAP [BBL*08]	12h	10	mp4 video	Ges., Audio, Text, Emotion		Dialog	sail.usc.edu/iemocap/
SaGA [LBH*13]	1 h	6	mp4 video	Ges., Audio, Gest. properties		Dialog	www.phonetik.uni-muenchen.de/Bas/BasSaGAeng.html
Creative-IT [MYL*16]	2h	16	3d joint rot.	Ges., Audio, Text, Emotion		Dialog	sail.usc.edu/CreativeIT/ImprovRelease.htm
MPI-EBEDB [VDLRBM14]	1.43h	8	3d joint rot.	Ges., Text		Monolog	ebmdb.tuebingen.mpg.de
Gesture-Speech Dataset [TKS*17]	5h	2	3d joint rot.	Ges., Audio	✓	Monolog	bit.ly/2Q4vSnT
CMU Panoptic [JSL*17]	5.5 h	50	3d joint rot.	Ges., Audio, Text		Dialog	domeb.perception.cs.cmu.edu/
Trinity Speech-Gesture I [FM18]	6 h	1	3d joint rot.	Ges., Audio		Monolog	trinityspeechgesture.scss.tcd.ie/data/TrinitySpeechGestureI/
Speech-Gesture [GBK*19]	144 h	10	2d coords.	Ges., Audio		Monolog	github.com/amirbar/speech2gesture
TED Dataset [YKJ*19]	106 h	1,295	2d coords.	Ges., Audio		Monolog	github.com/youngwoo-yoon/youtube-gesture-dataset
Talking With Hands [LDM*19]	50h	50	3d joint rot.	Ges., Audio	✓	Dialog	github.com/facebookresearch/TalkingWithHands32M
PATS [ALIM20]	250 h	25	2d coords.	Ges., Audio, Text		Monolog	chahuja.com/pats
Trinity Speech-Gesture I GENEA Extension [KJY*21]	6 h	1	3d joint rot.	Ges., Audio, Text		Monolog	TrinitySpeechGestureI/GENEA_Challenge_2020_data_release/
Trinity Speech-Gesture II [FNM21]	4 h	1	3d joint rot.	Ges., Audio, Gest. segment.		Monolog	trinityspeechgesture.scss.tcd.ie/data/TrinitySpeechGestureII
Speech-Gesture 3D extension [HXM*21]	144 h	10	3d coords.	Ges., Audio		Monolog	nextcloud.mpi-klb.mpg.de/index.php/s/7LzxGSeprzndg2x
Talking With Hands GENEA Extension [YWK*22]	20h	17	3d joint rot.	Ges., Audio, Text	✓	Dialog	zenodo.org/record/6998231
SaGA++ [KNN*22]	4 h	25	3d joint rot.	Ges., Audio, Gest. properties		Dialog	svito-zar.github.io/speech2properties2gestures
ZEGGS Dataset [GFH*]	2 h	1	3d joint rot.	Ges., Audio	✓	Monolog	github.com/ubisoft/ubisoft-laforge-ZeroEGGS
BEAT Dataset [LZI*22]	76 h	30	3d joint rot.	Ges., Audio, Text, Gest. properties, Emotion	✓	Both	pantomatrix.github.io/BEAT

Table 1: Speech-gesture datasets ordered from oldest to newest. The following abbreviations were used: “sp.” for “speakers”, “mot. format” for “motion format”, “HQ fing.” for “high-quality fingers”, “coords.” for “coordinates”, and “rot.” for “rotations”.

then modeled using statistical models inspired by work in speech recognition and dialogue act recognition [RK97]. The plausibility of a gesture was estimated using conditional probabilities on gesture bi-grams, and the occurrence of the gesture given semantics from input text. The result was statistical models for an individual’s gesture properties like handedness, transitions and timing, forming an individual’s gesture profile. The profiles were then used to generate plausible gestures from annotated input speech. The generation process had distinct stages: 1) Assigning semantic tags to input text; 2) Generating all possible gestures, adding them to an intermediate graph representation, and assigning probability estimates to the graph; 3) Filtering and temporal placement of gestures using text-gesture associations and timing profiles, respectively. The final output was an XML action script that could be used in a downstream animation system.

Extending this approach, Neff et al. [NKAS08] proposed a statistical system that learned gesture profiles but also added a character-specific animation lexicon. The system had two distinct phases. The pre-processing stage started with a video corpus of a character, hand-annotated in ANVIL [Kip01]. The annotation process was similar to that of Kipp [Kip05], but with an additional English-speaking character. Given the annotated data, a gesture profile (a statistical model) and animation lexicon were created, where the latter consisted of hand orientation, torso posture and data for after-strokes (i.e. subsequent repeated hand movements after a prominent stroke), for each gesture lexeme. The full-automated generation phase had two distinct paths: 1) *re-creation* that took in an annotated video as input and could re-create the gestures (observed in the video) in the animation system, useful for validating the annotations; 2) *gesture generation* that could generate gesture from novel annotated text without the need for video input. Either path leveraged the character’s gesture profile to generate a gesture script. The gesture script was used in the animation engine to generate the final animation either through a kinematic or dynamic simulation algorithm.

Bergman and Kopp proposed a different statistical approach for

modeling the transformation of speech that describes objects into iconic gestures that resemble said objects [BK09b]. The proposed system generated coordinated speech and gesture by leveraging propositional and imagistic knowledge representations for content planning and concrete speech and gesture formulation. Their work involved dyadic conversations where one speaker gives spatial directions to another after exploring a virtual environment in VR. The study investigated what contextual factors are important in the formation of speech and gesture describing physical objects. As part of their framework, they developed a Bayesian network for gesture formulation. The Bayesian network defined a probability distribution over gesture properties such as indexing, placing, shaping, drawing and posturing. The probability distribution also took into account the idiosyncratic patterns for mapping visuospatial referents onto gesture morphology, i.e. the specific way an individual might index, shape or draw a gesture when describing a referent object. Gesture formulation resulted in fine-grained features including hand shape, wrist location, palm direction, extended finger direction, movement trajectory and direction. For the final animation, the framework leveraged the rule-based Articulated Communicator Engine (ACE) [KW04] to realize synchronized speech and gesture.

Bergman and Kopp closely followed up with a hybrid framework combining data-driven and model-based techniques to model iconic gestures using Bayesian *decision* networks [BK09a]. They used a similar corpus of dyadic interactions with spontaneous speech and gesture employed for direction giving and landmark description. The corpus was richly annotated with temporal segments, gesture morphology and references to objects for iconic gestures. Extending their earlier work that used Bayesian networks [BK09b], they used Bayesian decision networks, supplemented by decision network nodes [HM05]. Bayesian decision networks enabled them to formulate gesture generation as a finite sequential decision problem by combining probabilistic and rule-based components. For example, the decision to include a certain gesture or the morphology of the gesture could be encoded as a decision node (activated by a rule) or chance node (activated by probability) with a specific

probability distribution. To generate a gesture, the Bayesian network defined a probability distribution over gesture morphological features, based on object referent features, discourse context and the previously performed gesture.

Levine et al. proposed a hidden Markov model (HMM) to select the most suitable motion clip from a motion capture database, by using prosody-based features extracted from the original speech [LTK09]. The trained HMM used prosody cues to select the most appropriate gesture sub-units from the motion capture, ensuring that the chosen sub-units transition smoothly and are appropriate for the tone of the current utterance. However, directly associating prosody with gesture sub-units created a dependence on the quality and amount of training data, which made the system susceptible to overfitting. Levine et al. [LKTK10] improved upon the previous system by proposing “gesture controllers” that decoupled the kinematic properties of gestures (e.g. velocity, spatial extent) from their shape. Gesture controllers inferred gesture kinematics using a conditional random field (CRF) that analyzed the acoustic features in the input speech and learned a distribution over a set of hidden states. The hidden states encoded the latent structure of gesture kinematics without regard for the morphology of the gesture which reduced the number of false correlations and thus alleviated overfitting. Finally, a Markov Decision Process (MDP) took the hidden states and the distribution over them as input and used an optimal policy (learned via the reinforcement learning algorithm) to select the appropriate gesture clips.

Chiu et al. [CM11] maintained the use of prosodic features to learn a probabilistic model for gesture generation. They restricted their study to learning gesture types that are associated with prosody, i.e. rhythmic movements (beats). The gesture generator was based on a modified Hierarchical Factored Conditional Restricted Boltzmann Machine (HFCRBM) [TH09]. They first built a compact motion representation by training a conditional Restricted Boltzmann Machine (CRBM) using an unsupervised learning algorithm. Then the HFCRBM generator autoregressively took in the previous gesture representation and a sequence of audio features extracted from the original speech to generate the gesture representation for every time step, until the full motion sequence was completed. Finally, they smoothed discontinuities between frames by reducing the acceleration of wrist joints if they exceed a set threshold. However, their approach was restricted to rhythmic gestures and thus did not consider other commonly occurring gesture types such as iconic, pantomimes, deictic, emblematic and metaphoric.

Recently, Yang et al. [YYH20] proposed a statistical motion-graph-based system that generated gestures and other body motions for dyadic conversations, that were well synchronized with novel audio clips. They constructed a motion graph that preserved the statistics of a database of recorded dyadic conversations. During generation, the graph was used to search for the most plausible motion sequence according to three constraints of audio-motion coordination in human conversations: 1) coordination to phonemic clause; 2) listener response; 3) partner’s hesitation pause. The system adapted motion graphs, successfully employed in locomotion [LCR*02, KGP02, AF02, TLP07], for free-form conversation gestures with a lot more stylistic variation. Their conversational motion graph was significantly larger than that for locomotion due

to the richness of conversational gestures. Given such a large graph, the system balanced search efficiency and style diversity by leveraging a stochastic greedy search algorithm to find a high-quality animation, well synchronized with the audio.

Statistical models provide more flexibility than rule-based systems and capture the non-determinism found in conversational gestures. In fact, a lot of the statistical principles (i.e. learning a probability distribution over the gesture data, through maximum likelihood estimation (MLE)) are still useful and relevant to most state-of-the-art methods, currently dominated by deep learning-based models. However statistical systems usually considered a limited number of independent variables based on painstakingly annotated gesture data. Deep learning-based models provide more flexibility through great representative capacity as well as making even fewer assumptions about the statistics of the underlying data. We describe this family of models next.

5.3. Deep learning approaches

Deep learning-based generative models recently gained interest because of their ability to synthesize data from abstract representations of training datasets. They are increasingly prominent in character animation applications, including character control in games, and facial or gesture animation conditioned on speech and text in virtual agents. Such models typically make few assumptions about the underlying data distribution (except useful inductive biases), and learn their parameters to fit the data through gradient-based optimization of an objective function.

The use of deep-learning approaches has moved the field forward substantially in terms of perceived naturalness, but arguably represents a step backwards in terms of communicative efficacy with respect to previous methods, as illustrated in Figure 4. Instead, the main targets of systems based on deep learning have been human-likeness and appropriateness for speech audio and semantic content. The former is the degree to which the generated gesture motion visually resembles believable human behavior, while the later is how suitable it is for a given speech audio, text input, or other contextual information. Early deep-learning systems ignored semantics, instead focusing on improving human-likeness [AHKB20, KHH*19, FNM20]. Later approaches have tried to incorporate semantics in order to generate meaningful gestures. The first attempts could generate only a handful of such gestures [KJvW*20, YCL*20, ALIM20]. Although more recent work suggest that progress can [KNN*22] and has been made [AGL*22], appropriateness remains a challenge. This can be seen from the GENEA Challenge (where GENEA stands for Generation and Evaluation of Non-verbal Behavior for Embodied Agents), which is a recurring large-scale comparison of gesture synthesis systems, whose most recent iteration [YWK*22] found that the human-likeness of motion can now reach the level of human motion capture, while appropriateness is still barely above chance.

The proliferation of deep learning in conversational gesture generation has led to a large number of approaches that can be grouped based on the input modalities, i.e. audio, text, audio and text, or audio with other non-communicative modalities, and control parameters. We employ this taxonomy to organize our exposition and give

a summary of the models and their respective categories in Table 2. We only include approaches that produce hand gestures and were published before the submission deadline for our review. In sections 5.3.1, 5.3.2, and 5.3.3 we discuss generation approaches that use audio-only, text-only, and a combination of audio and text input, respectively. Section 5.3.4 focuses on approaches that use non-linguistic input, i.e. input other than speech audio or text. Finally, Section 5.3.5 explores approaches that employ control input. The approaches within each modality section are presented in chronological order to reflect the evolution of the field.

5.3.1. Audio input

Hasegawa et al. [HKS*18] proposed an autoregressive approach to generate gesture from audio utterances using a bi-directional LSTM [HS97]. The bi-directional LSTM learned audio-gesture relationships with both backward and forward consistencies over a long period of time. The model was trained with a then novel audio-gesture dataset, collected using a headset and marker-based motion capture [TKS*17]. The model predicted a full skeletal human pose from the utterance features input at every LSTM timestep. Temporal filtering was then used to smooth out discontinuities in the generated pose sequences.

Kucherenko et al. [KHH*19] extended the work of Hasegawa et al. [HKS*18], removing the need for temporal smoothing through representation learning of an autoencoder. The proposed model transformed audio input into a gesture sequence in the form of 3D joint coordinates. They achieved this by (i) learning a lower dimensional representation of human motion using a denoising autoencoder consisting of a motion encoder (called MotionE) and a motion decoder (called MotionD) and (ii) training a novel speech encoder (called SpeechE) to transform speech to the corresponding motion representation with reduced dimensionality. During inference, the SpeechE predicted the motion representations, based on a given speech signal, and the MotionD decoded the motion representations into gesture sequences. However, their approach was deterministic and thus unable to capture the commonly observed phenomena where a person gesticulates differently at different points of the same utterance.

Deterministic generative approaches usually learn their parameters using a regression objective, e.g. L1 (Mean Absolute Error) or L2 (Mean Squared Error). Optimizing with either of those objectives typically forces the model toward learning to generate the mean representation of the data, producing averaged motion for different inputs, and resulting in undesirable results; usually called regression to the mean. Several approaches avoided this by incorporating probabilistic components into their objectives. Probabilistic components can increase the range of gesture motion in multiple ways, namely: (i) greater range of motion for different inputs, or (ii) stochastic motion for the same input. The most prominent are implicit log-likelihood evaluation via adversarial learning with Generative Adversarial Networks (GAN) [GPAM*14], explicit log-likelihood evaluation via variational inference with Variational Autoencoders (VAE) [KW13], and exact log-likelihood evaluation via invertible transformations with Normalizing Flows [KPB20, PNR*19].

GANs aim to do implicit density estimation of the underlying

distribution through the interplay of a generator that tries to produce samples that are representative of the data, and a discriminator that strengthens the generator by classifying samples as real (from the distribution) or fake (not from the distribution). Multiple gesture generation approaches added an adversarial objective as a term in a composite loss function, which increased the range of gesture motion although still deterministic for a given audio input [SB18, GBK*19, FNM20, YCL*20, ALNM20, RGP21, WLII21b, WLII21a, ZRMOL22, HES*22]. We discuss some notable examples below.

Ferstl et al. [FNM19, FNM20] added multiple adversarial objectives to a recurrent neural network, known to be susceptible to regression to the mean for long sequences. The adversaries accounted for gesture phase structure, motion realism, displacement and diversity of the minibatch. Ahuja et al. [ALNM20] learned to implicitly estimate a mixture of densities, each representing a speaker to enable the transfer of one speaker’s style onto the speech input of another. The mixture of densities was estimated by instantiating a generator (per speaker) that is responsible for generating gestures that are representative of that speaker’s underlying gesture distribution. All the generators were trained in unison using the adversarial objective. Most recently, Habibie et al. [HES*22] used an adversarial objective in the form of a conditional GAN to refine gesture clips that were selected using a k-Nearest Neighbour (kNN) search that is conditioned on speech and a control signal.

Normalizing flows learn to describe highly complex distributions by applying invertible sub-transformations to a simple initial distribution, where invertibility allows one to optimize the exact likelihood of the deep generative model via gradient descent [KPB20, PNR*19], unlike in GANs or VAEs. In the context of gesture generation, Alexanderson et al. [AHKB20] extended a normalizing flow-based model for locomotion [HAB20] to apply to speech audio-driven gesture synthesis with style control. Their normalizing flows learned invertible transformations from simple Gaussian distributions to the distribution of upper- or full-body motion capture data. These transformations were conditioned on speech acoustics and, optionally, arbitrary style parameters. Specific style parameters considered included properties such as the average hand height, gesture speed, and gesture radius in a 4-second interval. The resulting model produced gestures that scored among the best in terms of naturalness and appropriateness in the GENE Challenge 2020 [KJY*20].

VAEs aim to do explicit density estimation of the underlying data by optimizing a combination of a reconstruction loss for an autoencoder (usually an L1 or L2 loss) and the Kullback-Leibler (KL) divergence for distribution matching between a prior distribution and approximate posterior distribution of the data. The prior distribution is usually instantiated as a Gaussian for simple parameterization. The learned stochastic variables can then be sampled and decoded into diverse outputs. Recently, Li et al. [LKP*21] used a conditional VAE to translate speech into diverse gestures. They explicitly modeled a one-to-many speech-to-gesture mapping by splitting the cross-modal latent code into a shared code (audio + gesture) and motion-specific code (gesture only). The shared code modeled the correlation between speech and gesture e.g. synchronized speech and rhythmic gestures. The motion-specific code at-

Model	Dataset	Inputs					Training Objective	Sequential	Output Representation				Stochastic
		Audio	Text	Pose	Control	Other			Joint pos.	Joint rot.	Pre-rec.	Other	
DCNF [CMM15]	DIAC [GAL*14]		✓				MLE	✓			✓		
Hasegawa et al. [HKS*18]	Gesture-Speech [TKS*17]	✓					MSE	✓	✓				
Ishi et al. [IMM18]	Ishi et al. [IMM18]	✓	✓				MLE	✓			✓		✓
Kucherenko et al. [KHH*19]	Gesture-Speech [TKS*17]	✓					MSE	✓	✓				
Yoon et al. [YKJ*19]	TED [YKJ*19]		✓				MSE + MAE - Var	✓	✓				
DRAM [AMMS19]	Ahuja et al. [AMMS19]	✓		✓			MSE	✓			✓		
Speech2Gesture [GBK*19]	S2G [GBK*19]	✓					MAE + Adv		✓				
Ferstl et al. [FNM19]	Trinity I [FM18]	✓					MSE + Adv	✓	✓				✓
CDBN [SB19]	MSP-AVATAR [SLB15]	✓				✓	EM	✓				✓	✓
Mix-Stage [ALNM20]	PATS [ALNM20]	✓					MAE + CCE + Adv		✓				✓
Yoon et al. [YCL*20]	TED [YKJ*19]	✓	✓			✓	Huber + Adv + KL	✓	✓				✓
Bozkurt et al. [BYE20]	Creative IT [MYL*16]	✓				✓	MLE	✓				✓	
Gesticulator [KJvW*20]	Trinity I [FM18]	✓	✓				MSE	✓		✓			
StyleGestures [AHKB20]	Trinity I [FM18]	✓			✓		MLE	✓	✓				✓
ASLE [ALIM20]	PATS [ALNM20]	✓	✓				MAE + CCE + Adv		✓				✓
Habibie et al. [HXM*21]	S2G 3D [GBK*19, HXM*21]	✓					MSE + MAE + Adv		✓				
Korzun et al. [KDZ21]	Kucherenko et al. [KJY*20]	✓	✓				MSE	✓		✓			
Audio2Gestures [LKP*21]	Trinity I [FM18]	✓					MAE + GeoD + KL		✓	✓			✓
Body2Hands [NGDJ21]	S2G [GBK*19]			✓		✓	MAE + Adv		✓				
Lee et al. [LAM21]	PATS [ALNM20]	✓	✓				MAE + Adv + CC-NCE	✓	✓				
Text2Gestures [BRB*21]	MPI-EBEDB [VDLRBM14]		✓			✓	MSE	✓		✓			
ExpressGesture [FNM21]	Trinity A + B [FM18, FNM21]	✓					MSE	✓			✓		
CMCF [SBM21]	S2G [GBK*19]	✓	✓				WCSS	✓			✓		
Qian et al. [QTZ*21]	S2G [GBK*19]	✓					MAE + KL		✓			✓	
Rebol et al. [RGP21]	Rebol et al. [RGP21]	✓					MAE + Adversarial		✓				
Flow-VAE [TWGM21]	Taylor et al. [TWGM21]	✓					NLL + MSE	✓		✓			✓
Wu et al. [WLIJ21b]	Gesture-Speech [TKS*17]	✓		✓		✓	Adv + WGAN-GP + Huber	✓		✓			✓
Wu et al. [WLIJ21a]	Gesture-Speech [TKS*17]	✓				✓	Adv	✓	✓				✓
Kucherenko et al. [KNN*22]	SaGA++ [KNN*22]	✓	✓				CCE					✓	
Nguyen et al. [NC22]	JESTKOD [BKK*17]	✓		✓			Adv	✓			✓		
GestureMaster [ZBC22]	TWH GENE [YWK*22]	✓	✓				L2 + Hamm + ETC	✓			✓	✓	
ZeroEGGS [GFC22]	Ghorbani et al. [GFC22, GFH*]	✓			✓		MAE + KL	✓	✓	✓		✓	✓
ZS-MSTM [FGP22]	PATS [ALNM20]	✓	✓				MSE + Adv	✓	✓				
Habibie et al. [HES*22]	S2G 3D [GBK*19, HXM*21]	✓		✓			MAE + Adv + WGAN-GP		✓				✓
SEEG [LFZ*22]	TED [YKJ*19]	✓	✓				MAE + Adv + KL						
Zhuang et al. [ZQZ*22]	Zhuang et al. [ZQZ*22]	✓	✓				MSE + SIMM + ETC				✓		
Zhou et al. [ZYL*22]	Personal Story [ZYL*22], TED [YKJ*19]	✓				✓	MAE	✓				✓	
Deichler et al. [DWAB22]	Deichler et al. [DWAB22]			✓			IR + TR	✓	✓				
DiffGAN [ALM22]	PATS [ALNM20]	✓	✓	✓			MAE + MSE + Adv		✓				
Rhythmic Gesticulator [AGL*22]	Trinity I [FM18], TED [YKJ*19]	✓	✓	✓	✓		MSE + KL	✓		✓			✓

Table 2: Summary of deep learning-based models in chronological order. The “Sequential” column indicates frame-by-frame generation otherwise frames are generated in parallel. The “Stochastic” column indicates varied gesture output for any input otherwise the output is the deterministic. See Table 3 for elaborations of the “Training Objective” abbreviations.

tempted to capture the diversity in gesticulation, independent of audio information.

In a similar vein, Ghorbani et al. [GFC22, GFH*], used a VAE-based framework for style controllable co-speech gesture generation conditioned by a zero-shot motion example i.e., an instance of a motion style unseen during training. Given an audio input and a motion example, they generated an encoding of the audio and a style embedding from the motion, and the two latent codes were used to guide the generation of stylized gestures. The variational nature of the style embedding enabled them to easily modify style through latent space manipulation or blending and scaling of style embeddings. Moreover, the probabilistic nature of the model enabled the generation of varied gestures for any audio and exemplar motion input. The resulting model performed favorably against state-of-the-art probabilistic techniques [AHKB20] in terms of naturalness of motion, appropriateness for speech, and style portrayal.

Taylor et al. [TWGM21] adapted the conditional Flow-VAE framework [BHF*19], combining the advantages of the VAE and

normalizing flow architectures, to generate spontaneous gesture movement for speaker and listener roles in a dyadic interaction. They used the Flow-VAE framework for modeling expressive gesture because of its ability to improve generative capacity of the VAE by estimating the latent space with a highly complex distribution using a normalizing flow, instead of the standard Gaussian. Their autoregressive framework was trained on a set of previously generated gestures, an audio input window and the dyadic role, i.e. speaker or listener, as input. The preceding gestures were encoded into a latent variable then transformed into a complex distribution using a normalizing flow, conditioned by the audio window and role in the dyad. Their decoder then generated the next gesture based on the latent variable sampled from the complex distribution. The resulting model could generate expressive co-verbal gestures in a dyadic setting based.

Hybrid systems that combine deep learning and database matching components can also help tackle the regression to the mean problem [KNN*22, ZBC22, FNM21, HES*22]. Indeed, this approach has been used effectively in motion synthesis problems, e.g.

Objective	Full name
Adv	Adversarial Loss
CCE	Categorical Cross Entropy
CC-NCE	Cross-modal Cluster Noise Contrastive Estimation
ETC	Edge Transition Cost
EM	Expectation Maximization
GeoD	Geodesic Distance
WGAN-GP	Wasserstein-GAN Gradient Penalty
Hamm	Hamming Distance
Huber	Huber Loss
IR	Imitation Reward
KL	Kullback–Leibler Divergence
L2	L2 Distance
MAE	Mean Absolute Error
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error
NLL	Negative Log-likelihood
SIMM	Structural Similarity Index Measure
TR	Task Reward
Var	Variance
WCSS	Within-cluster Sum of Squares

Table 3: Training objective abbreviations in Table 2

game animation where high fidelity motion is crucial [HKPP20]. In the context of conversational gesture, the intuition is that modeling the association between high dimension audio input and gestures, represented by exact joint positions or angles, using standard regression objectives (L1 or L2 loss) discourages the model from producing otherwise plausible gestures that do not exactly match the ground truth, thus greatly reducing the variety of generated gestures. Alternatively, the audio-gesture association can be modeled by predicting higher-level parameters for gesture motion. Ferstl et al. [FNM21] realized this idea by learning to map audio to gesture via higher-level expressive parameters, specifically gesture velocity, acceleration, size, arm swivel angle, and extent of hand opening. First they pre-trained a model to associate audio prosodic features to the expressive parameters. Then they predicted gesture timing by extracting pitch peaks in the audio signal. At inference time, the prosodic features were used to estimate the expressive parameters that were in turn used to search for a matching gesture in the database, and the pitch peaks were used to temporally position the matching gesture. Finally, synthetic preparation and retraction phases were added to connect the gestures in the sequence.

Another interesting approach for preserving gesture form is through audio-based search in a video gesture database where the gestures are represented by video frames. Zhou et al. [ZYL*22] explored this idea in a gesture reenactment task by generating a gesture video for an unseen audio input, using gesture frames from a reference video. They first encoded the reference gesture video as a “video motion graph” - a directed graph where each node represented a video frame and corresponding audio features, and the edges represented transitions. The graph encoded how the reference video can be split and re-assembled in difference graph paths. In order to increase graph connectivity, i.e. diversity of plausible path, they added synthetic edges based on a frame pose similarity

threshold computed using the SMPL pose parameters [LMR*15]. Given unseen audio input as a guide, they traversed the graph using a beam search algorithm [RR77] to find the most optimal path or order of gesture frames that best matches the speech audio. For graph paths that contain temporally disjoint frames, they trained pose-aware video blending network to synthesize smooth transitions between the frames.

5.3.2. Text input

Approaches that used audio as the primary modality produced well-timed hand movements that tend to be highly associated with acoustics, largely corresponding to beat gestures. However, the lack of text transcript means they were not informed by the structure and context inherent in the text, for example, semantic meaning and punctuation. Such structure can help produce more meaningful and communicative gestures. Therefore, next, we describe some approaches that used text as the primary input modality.

Ishi et al. [IMM18] proposed a text-based gesture generation approach for controlling a humanoid robot. They modeled the text-to-gesture motion translation by associating words to concepts, concepts to gesture categories (i.e. iconic, metaphoric, deictic, beat, emblem and adapter), and gesture categories to gesture motions. Further, they estimated conditional probabilities to model the association between word concepts and gesture categories, and between gesture categories and gesture motion clusters that were pre-computed with the k-means clustering algorithm.

Yoon et al. [YKJ*19] proposed an encoder-decoder approach that transformed speech text, from a dataset based on TED talks, into a sequence of gestures. They created the TED video dataset by picking video segments with the speaker’s upper body and hands. Then they performed pose estimation using OpenPose [CHS*18] and removed segments containing noisy or no estimations. Speech text was converted into a sequence of 300-dimensional word vectors using pre-trained GloVe embeddings [PSM14]. Similarly, poses estimations were converted to 10-dimensional vectors using principal component analysis (PCA). Therefore, co-speech gesture generation became a sequence-to-sequence translation problem from word embeddings to human poses encodings. The encoder part of the network was a bi-directional GRU [CVMG*14] taking in speech text one word (vector) at a time and capturing bi-directional context. The last hidden state of the encoder was passed into the decoder, also a bi-directional GRU. The decoder also took previous pose estimations to condition the prediction of the next pose, in addition to using soft-attention [BCB14] to focus on specific words when predicting the next pose. Finally, the generated 2D poses were mapped to 3D and executed on the NAO humanoid robot.

Recently, Bhattacharya et al. [BRB*21] used text transcripts to produce expressive emotive gestures for virtual agents in narration and conversation settings, using MPI-EBEDB, a dataset of actors performing multiple emotion categories (amusement, anger, disgust, fear, joy, neutral, pride, relief, sadness, shame, surprise) [VDLRBM14]. Their approach consisted of Transformer-based encoders and decoders [VSP*17], where the encoder took in the text transcript sentences (encoded as GloVe embeddings [PSM14]) to produce an encoding which was concatenated with the agent at-

tributes such as narration/conversation, intended emotion, gender and handedness. The previous pose’s encoded concatenation and 3D joint positions were passed as input to the Transformer decoder to generate the next pose’s joint positions. The process was repeated in a recurrent manner until the full pose sequence was generated.

5.3.3. Audio and text input

An interesting trade-off exists between audio-based and text-based gesture generation systems. audio-based generators have access to intonation and prosody which helps generate rhythmic or kinematic gestures (e.g. beats) but lack semantic context. Conversely, text-based generators have access to semantic context which helps generate meaning-carrying gestures (e.g. iconic or metaphoric), but lack intonation and prosodic information. Therefore, combining the audio and text modalities enables a gesture generator to learn to produce semantically relevant and rhythmic co-speech gestures.

Although generating meaning-carrying gestures using audio only is theoretically possible, it is unlikely since prosody is suitable for kinematics, but not sufficient to infer shape which is associated with meaning [LWH*12]. As far as we know, meaningful gestures from speech audio alone have not been empirically demonstrated. Instead, combining audio with text appears to be the most promising approach to generating meaningful gestures to date. We, therefore, focus on approaches that combine these two modalities for generating meaning-carrying, communicative gestures.

Chiu et al. [CMM15] proposed an approach that combined the text and prosody of the speech to generate co-verbal gestures. Their model, called the Deep Conditional Neural Field (DCNF), was a combination of a fully-connected network, for representation learning, and a Conditional Random Field (CRF), for temporal modeling. For the gesture prediction task, the model took in a text transcript, part-of-speech tags and prosody features as input, and predicted a sequence of gesture signs which were a set of predefined hand motions.

Leveraging the representation power of deep learning models for multimodal input (i.e. audio and text) for co-speech gesture generation was the next logical step. In fact, three groups of researchers independently proposed the first deep-learning-based gesture generators that used both audio and text to generate continuous gestures, namely Yoon et al. [YCL*20], Ahuja et al. [ALIM20] and Kucherenko et al. [KJvW*20]. We discuss their pioneering work combining audio and text next, followed by subsequent efforts in the area.

Yoon et al. [YCL*20] proposed a gesture generation approach that combines the tri-modal context of speech, text and speaker identity to produce gestures that were human-like, and matched the content and rhythm of the speech. The model processed input speech and text with a speech and text encoder, respectively. The speaker identity was used to sample the intended speaker from a learned style embedding space. Together the three features (i.e. speech encoding, text encoding, and style) were passed to a gesture generator to produce the sequence of poses. Closely related was Liang et al. [LFZ*22] whose framework utilized audio and text information in order to generate meaningful gestures by disentangling semantic and beat gestures. Their system consisted of two

encoders, one that took in audio and text to encode semantics, and another that took in audio volume and beat to encode non-semantic information. The encoded information from both encoders ensured the disentanglement of semantic and beat gestures, while decoder took this information and was trained to encourage generation of meaningful semantic gestures.

Ahuja et al. [ALIM20] identified two key challenges in an attempt to learn the latent relationship between speech and co-speech gestures. First, the underlying distributions for text and gesture are inherently skewed and therefore necessitated the need to learn their respective long tails, accounting for rarely occurring text or gestures. Second, gesture predictions are made at the sub-word level, which necessitated the need to learn the relationship between language and acoustic cues that may give rise to, or be accompanied by, a particular gesticulation. So motivated, they proposed the Adversarial Importance Sampled Learning (AISLe) framework, that combined adversarial learning with importance sampling to balance precision and coverage. The model took in speech and text transcripts and performed encoding and alignment between sub-words and acoustics, using a multi-scale Transformer [VSP*17]. The resulting alignment was passed to the model’s generator to predict the pose sequence and an adversarial discriminator was used to determine if the pose was real or fake. For optimizing the adversarial objective, the AISLe framework scaled the loss function such that rarely occurring gesture samples, the long tail of the distribution, were weighted more than those that are more likely to occur.

Kucherenko et al. [KJvW*20] proposed an autoregressive generative model that combined speech acoustics and semantics to produce arbitrary acoustically-linked or semantically-linked gestures. The key insight of their approach was to envision a gesticulation system that encompasses so called “representational” gesture types (i.e. iconic, metaphoric and deictic) that convey semantics, and beats that are synchronized with acoustics. Their approach took a concatenation of semantic features that were extracted using BERT [DCLT18] and acoustic features represented as log-power mel-spectrograms as input into an encoder. Then they integrated past and future context for each gesture pose frame via a sliding window operation over the encoded speech features. The model generated each pose autoregressively where each was conditioned on the information of three preceding frames to ensure motion continuity. Their extensive evaluation indicated that autoregression for continuous motion and combining audio and text had the most significant positive impact on the quality of the generated gesticulations.

Equally inspired by autoregressive generative models, Korzun et al. [KDZ21, KDZ20] reimplemented the text-only recurrent framework by [YKJ*19] to accommodate both text and audio input. The proposed model was a combination of a recurrent context encoder, inspired by [KHH*19], that generated hidden states for 3-second audio and text context windows and a recurrent encoder-decoder that took in the concatenated results of the context encoder and used an attention mechanism to condition the generation of the final gesture motion. Similar to Yoon et al. [YKJ*19], they trained the model using the continuity and variance objectives to ensure fluid and natural-looking gestures. The resulting model produced gestures that were deemed natural and appropriate as part of the GENE Challenge 2020 [KJY*20].

Designing generation systems that produce meaningful gestures is one of the major goals in non-verbal behavior research. Spurred on by this question, Kucherenko et al. [KNN*22] investigated whether contemporary deep learning-based systems could predict gesture properties, namely phase, type and semantic features, as a way to determine if such systems can consistently generate gestures that convey meaning. Their model used both audio and text for predicting gesture properties, through two distinct components that predicted the probability for gesticulation and probabilities for the aforementioned set of gesture properties. They conducted their experiments on a direction-giving dataset with a high number of representational gestures [LBH*13]. Their experiments showed that gesture properties related to meaning such as semantic properties and gesture type could be predicted from text features (encoded as FastText embeddings [BGJM17]), but not from prosodic audio features. Conversely, they found that rhythm-related gesture properties (e.g. phase) could be better predicted from audio features.

In order to mimic the communicative intent of co-speech gestures, it is crucial to understand and model the complex relationship between speech acoustics, text, and hand movements. An interesting approach is to group gestures with distinct movement properties in order to find emergent rhetorical categories. Saund et al. [SBM21] investigated this approach by modeling the rhetorical, semantic, affective and acoustic relationships between gestures and co-occurring speech audio and text, for a hypothetical gesture generation system. They first used k-means clustering to cluster speech-gesture pairs into functional domain clusterings (i.e. rhetorical, affective and semantic) based on functional tags generated from third-party natural language parsers. The speech-gesture pairs were refined into sub-clusters based on gesture motion. Therefore each speech-gesture pair belonged to at least one sub-cluster (based on motion), within one functional cluster (based on its assigned functional tags). At run-time, a hypothesized virtual agent would leverage the same pre-trained parsers and clusters to analyze an input speech and text transcription, select a functional cluster and from that a motion sub-cluster. The agent could then either choose an appropriate gesture from a pre-recorded library or the centroid gesture in the motion sub-cluster.

Motion graphs, commonplace in conventional animation systems (e.g. [KGP02, AF02, LCR*02]), can be effective at producing realistic non-verbal behavior because they rely on databases of high-quality motion capture or RGB video. As we discussed before, they were effectively employed for audio-driven gesture reenactment using video-based motion graphs [ZYL*22]. Zhou et al. continued this trend for audio and text by adapting a motion-graph-based music-to-dance system [CTL*21] for co-speech gesture generation [ZBC22]. They first built a database of audio, text and gesture clips from 3-tuples of (audio, text transcript, gesture), using a splitting algorithm. For each audio clip, they generated a style signature using StyleGestures [AHKB20], and a rhythm signature using a binary encoding scheme that denotes the presence of words by leveraging the word-level timing information in the text transcript. For the corresponding gesture motion, they generated a style signature, parameterized by the same attributes as StyleGestures [AHKB20] (e.g. wrist speed, radius and height), and a rhythm signature using a similar binary scheme that denoted the presence of pausing, sharp turning or a stroke phase of the gesture. During

synthesis, they computed the rhythm and style signatures for input audio and text and used a graph-optimization algorithm to find gesture clips that closely matched the generated style and rhythm in terms of Hamming distance, and minimized the motion transition in the graph. This model performed on par or better than motion capture data in terms of Naturalness in the GENE Challenge 2022 [YWK*22].

Style transfer is a widely adopted optimization technique in deep learning for blending visual content and style, e.g. given a content image and a reference image that specifies the style, adjust the content image to match the style [GEB15]. In the context of co-speech gesture generation, it might be desirable to transfer the speaking style of one speaker to the predicted gestures of another. Ahuja et al. [ALNM20] learned unique style embeddings for multiple speakers that enabled either generation of gestures consistent with the original speaker in the audio input, or style transfer by combining the audio input of one speaker with the style embedding of a different speaker. Although they proposed the PATS data where multiple modalities such as audio, gesture pose and text have style and content, they focused on gesture pose style to learn *unimodal* speaker-specific style embeddings. Fares et al. [FGPO22] leveraged the multiple modalities in the PATS dataset to learn *multimodal* style embeddings based on audio, text and gesture pose input. Their framework consisted of a speaker-style encoder that used speaker audio, text and gesture pose to learn a multimodal style embedding, and a sequence-to-sequence decoder that generated gestures based on audio and text, and conditioned on the desired speaker's style embedding. Furthermore, unlike the work of Ahuja et al. [ALNM20] that required the entire speaker's gesture data to learn the speaker's style embedding, their trained speaker-style encoder could generate style embeddings in a zero-shot manner i.e., for speaker styles not seen in the training set.

A key tenet of semantically meaningful gestures is that they are appropriate for the given utterance. To achieve this, there needs to be a greater emphasis on generating precise gestures using audio and text as grounding (i.e. the appropriateness of the gesture to the utterance), versus generating diverse gestures. Lee et al. [LAM21] investigated this approach and made an interesting observation about human gesticulation, that multiple semantically different utterances are often accompanied by the same gesture. They thus proposed a contrastive-learning framework that constrained the mapping of semantically different utterances to a smaller subset of relevant high-quality gestures. They introduced a novel contrastive learning objective that preserved similarities and dissimilarities of gestures in the latent representation. The objective ensured that latent language representations of two semantically different utterances were close together if they were accompanied by the same gesture. They first clustered gestures based on similarity or dissimilarity, then created positive (similar gesture poses) and negative (dissimilar gesture poses) required for the standard contrastive learning objective. Finally, they learned gesture-aware embeddings via a contrastive and adversarial objective. The resulting embedding space was used to generate gestures that were semantically relevant and closer to the ground truth.

When designing 3D avatars, it may be desirable to have a holistic animation system that includes facial and full-body movement.

Combining audio and text modalities can be effective at achieving this goal because of their rhythmic and semantic properties. Zhuang et al. [ZQZ*22] investigated this approach by proposing a hybrid system consisting of Transformer-based encoder and decoder modules, and motion-graph retrieval module to generate facial motion and full-body motion that included gestures. Their encoder used both audio and text, in the form of phoneme labels and Mel Frequency Cepstral Coefficients (MFCC) and Mel Filter Bank (MFB) features, to generate 3D facial parameters for synchronous lip movement. Simultaneously, the decoder used speech features, previous expression motion and semantic tags to generate 3D facial parameter for expression. The motion-graph retrieval sub-system used speech audio and text to find the most appropriate body motion segments, including gesture, that correspond to the text semantics and rhythm in the audio. Finally the facial and body motion were used to drive a skinned polygonal model.

5.3.4. Non-linguistic modalities

Several deep learning-based systems complemented input audio or text with additional information that could reasonably be deemed relevant to co-speech gestures. This included speech context, speaker style, discourse or an interlocutor's movements. Sadoughi and Busso [SB19] proposed a system that bridges rule-based and learning-based techniques in order to select gestures that are communicative and well synchronized with speech. They proposed a Dynamic Bayesian Network (DBN) which took in speech and two constraints to condition the generation. The constraints were: 1) discourse function, which restricts the model to behaviors that are characteristic of that discourse class (e.g. questions); 2) prototypical behaviors, which restricted the model to certain target gesticulations (e.g. head nods). Given constraints on prototypical behaviors, the approach could be embedded in a rule-based system as a behavior realizer creating head and hand trajectories that are temporally synchronized with speech.

In a dyadic conversation between interlocutors, there can be a lot of spontaneous non-verbal behavior that is influenced by the nature and tone of the interaction. Leveraging the co-adaptation of non-verbal behavior between interlocutors present in human-to-human interactions, cf. [BK12, CBK14, OB16], can enable virtual agents to be naturally conversational and collaborative. Ahuja et al. [AMMS19] proposed the Dyadic Residual-Attention Model (DRAM), a framework that could interactively generate an avatar's gesticulation conditioned on its speech and also the speech and gesticulation of a human interlocutor in a telepresence setting. In order to generate natural behavior, the avatar had to consider its own speech as well as the speech and gesticulation of the human. The DRAM model generated natural dyadic behavior by taking in the speech and pose history of the avatar as well as the speech and pose history of the human to adapt the avatar's gesticulation accordingly.

The idea of conditioning the motion of a deep-learning-based agent on interlocutor speech and motion has subsequently been used in several other works. Jonell et al. [JKHB20] used a model based on normalizing flows for generating head motion and facial expression, while Nguyen and Celiktutan [NC22] used conditional adversarial learning to drive full-body skeletons. Both of these works found that statistically significant improvements in generated behaviors were achieved by being interlocutor-aware.

A similarly interesting dyadic scenario is human-robot interaction where one of the interlocutors is a social robot. In this case, the robot must exhibit natural non-verbal behavior in order to be engaging and interesting. Therefore, it is desirable for the robot to mimic human non-verbal motion with gestures that are natural and communicative. Deichler et al. [DWAB22] investigated this idea by proposing a combination of a data-driven and physically-based reinforcement learning (RL) framework to generate pointing gestures learned from motion capture data. Given a diverse motion capture dataset of pointing gestures and corresponding targets, they trained RL control policies adapted from [PALvdP18, PMA*21] to imitate human-like pointing motion while maximizing the reward based on pointing precision.

Automatic synthesis and animation of gestures that accompany affective verbal communication can endow virtual agents with emotional impetus. Bozkurt et al. [BYE20] directly mapped emotional cues in speech prosody into affect-expressive gestures. They investigated the use of three continuous affect attributes (i.e. activation, valence and dominance) for the speech-driven synthesis of affective gesticulation. They proposed a statistical model based on hidden semi-Markov models (HSMM) where states were gestures, and observations were speech prosody and continuous affect attributes. They first estimated the affective state from speech prosody and then used the state and speech prosody to predict gesture clusters. The gesture segments were animated using a unit selection algorithm [BEY15], and discontinuities were smoothed using an exponential smoothing function. Finally, the smoothed sequence was animated in Autodesk MotionBuilder.

Text encodes important semantic information, potentially useful for conveying meaningful emotion through gesture, although it encodes fewer cues about emotional state compared to audio e.g., intonation and speech pauses. An interesting approach is to combine text with an intended emotion for affective gesture generation. Bhattacharya et al. [BRB*21] pursued this approach by combining text transcripts associated with narrative or conversational acting and emotion labels, to produce expressive emotive gestures for virtual agents. The emotions represented were amusement, anger, disgust, fear, joy, neutral, pride, relief, sadness, shame and surprise [VDLRBM14]. Their approach consisted of a Transformer [VSP*17] encoder and decoder, where the encoder took in the text transcript sentences, intended emotional state and agent attributes (e.g. narration/conversation, intended emotion, gender, handedness). The previous pose's encoded concatenation and 3D joint positions were passed as input to the Transformer decoder to generate the next pose's joint positions. The process was repeated in a recurrent manner until the full affective pose sequence was generated.

A speaker's identity or style can affect how they gesticulate, as some speakers gesture a lot while others rarely do. Moreover, they may also prefer particular gesture forms, and use different hands or gesture sizes. Modeling such variation in non-verbal behaviour can help make virtual agents seem unique and have a personality. To this end, Yoon et al [YCL*20] used speaker identity to guide gesture generation that matched the speaker's style. Their adversarial approach combined the tri-modal context of audio, text and speaker identity to produce gestures that were human-like, and matched

the content and rhythm of the speech. The model processed input audio and text with an audio and text encoder, respectively. The speaker identity was used to sample the intended speaker from a learned style embedding space. Together the three features (i.e. audio encoding, text encoding, and style) were passed to a gesture generator to produce the sequence of poses. Similarly, Ahuja et al. [ALNM20] learned a mixture of adversarial generators, representing diverse gesticulation styles of speakers from talk-show hosts, lecturers and televangelists. Learning speaker-specific generators enabled one speaker's style to be aligned with, or transferred to, the audio of another speaker.

Developing robust deep-learning-based gesture generators requires large amounts of diverse gesture data from real world scenarios, captured either via motion capture or pose estimation from videos. However, capturing or estimating hand gestures is very challenging because of the intricate finger motion, relatively small size of hands with respect to the whole body and frequent self-occlusions [HLW*18, LDM*19, JYN*20]. In contrast, capturing body motion (up to and including the arms) is less error prone because the joints are further apart and the articulations are relatively simpler. Therefore, the "upper body" motion as a modality can be an informative prior for generating conversational hand gestures. Ng et al. [NGDJ21] investigated this idea while making the observation that body motion is highly correlated with hand gestures. Their proposed approach took in 3D upper-body motion (up to the wrist) and predicted 3D hand poses. In addition to upper-body motion, the model could take in 2D images of hands and produce the corresponding 3D hand pose estimations. Similar to [GBK*19], they used a combination of a L1 regression loss for the model training signal and an adversarial loss to ensure realistic motion. The learned body-motion-to-hands correlation was versatile enough for several use-cases, namely conversational hand gesture synthesis, single-view 3D hand-pose estimation and synthesizing missing hands in motion capture data and image-based pose estimation data.

5.3.5. Control input

Although control can take either linguistic or non-linguistic forms, it is distinct because it can convey the explicit design and execution intent of an animator. Multiple works in motion synthesis use control as an additional input either during the training phase or the inference phase of learning-based models (e.g. [HKS17, LZCvdP20]). Typically, during training, the control signal is used to train the system to generate animations with certain biomechanical constraints such as posture, gait, etc. During inference, control may be introduced to impose style-related constraints [SCNW19] or user input [HKS17, HAB20, LZCvdP20].

In the context of conversational gesture, Alexanderson et al. [AHKB20] trained a probabilistic model that generated spontaneous co-verbal gesture that was conditioned on control constraints such as wrist height, radial extent and handedness. However, the constraints were introduced at training time, meaning modeling a new constraint required re-training the entire model. Habibie et al. [HES*22] provided a more flexible approach. They first learn a speech-to-gesture motion search through a kNN algorithm, and then refine the motion using conditional GAN. Style control can

be exerted at runtime by dynamically restricting the portion of the database that the kNN algorithm is run on, allowing style variation even within an extended utterance without the need to retrain.

Control can also be imposed by implicitly specifying the desired gestures by learning emergent prototypes of gesture shape or form. Qian et al. [QTZ*21] explored this idea by learning conditional vectors, so-called "template vectors", that could determine the general appearance and thus narrow the potential range of plausible gestures. Their framework took in audio and a zero initialized condition vector, through a 1D UNet-based autoencoder, in order to generate the corresponding gestures as 2D joint positions. During training, they periodically updated the condition vector, through back-propagation, using the gradients computed on the L1 regression loss between the generated and ground-truth gestures. They regularized the template vector space through the KL-divergence between the vectors and a normal distribution. They also separately pre-trained a VAE to reconstruct ground truth gestures and used the resulting latent space to encode gestures into template vectors. At test time, they sampled arbitrary template vectors, either learned through back-propagation or extracted by the pre-trained VAE, to generate diverse gestures.

Animators typically want to specify high-level style parameters to convey design intent e.g. energetic oratory gesticulations or subdued gestures to convey sadness. Additionally, it is desirable to specify the style once in the workflow and for the animation system to generate arbitrarily many motions for that specification. However, there is a gap between desired abstract design intent and existing deep-learning-based style control systems that tend to rely on biomechanical constraints such as wrist speed, radius or height [AHKB20, HES*22]. Style specification is also not data efficient, requiring as many samples as the size of the training set for the model to learn a style [AHKB20, ALNM20]. We conclude this section by discussing several works that proposed approaches for data-efficient style specification [GFC22, GFH*, FGPO22, ALM22].

Ghorbani et al. [GFC22, GFH*] proposed a framework that improves on high-level style portrayal by using exemplar motion sequences that demonstrate the intended stylistic expression of gesture motion. Their framework was able to efficiently extract style parameters in a zero-shot manner, only requiring a single example motion and was able to generalize to example motions (and therefore styles) unseen during training. Fares et al. [FGPO22] used an adversarial framework to learn a speaker-style encoder that could generate speaker-specific style embeddings from novel multimodal inputs – audio, text and gesture pose – not seen during the training phase. The framework generated co-speech gestures in a style that is either consistent with the original speaker in the audio or a different speaker, depending on the chosen style embedding. Ahuja et al. [ALM22] proposed an adversarial domain-adaptation approach for personalizing the gestures of a source speaker with plenty of data, with the style of a target speaker with limited data, using only 2 minutes of target training data. Given a model pretrained on a large co-speech gesture dataset, their framework could adapt the model's parameters using a smaller target dataset by modeling the cross-modal grounding shift, i.e., the change in distribution of speech-gesture associations, and the distribution shift in the target gesture space. The approach's ability to identify distributions shifts

between the source and target domain for parameter updates, enabled the model to extrapolate to gestures in the target distribution without having seen them in the source distribution during pretraining.

6. Key Challenges of Gesture Generation

Animating co-verbal gestures is still a very challenging problem because gestures are spontaneous, highly idiosyncratic and non-periodic. Rule-based approaches generate well-formed gestures by leveraging recording motion, but are inflexible and lack gesture diversity. Additionally, the hand-designed rules are non-exhaustive and often prescriptive, and hence may not be reflective of gestures which occur naturally and spontaneously. Data-driven approaches improve on diversity and flexibility but tend to produce marginally natural gestures that appear more like well-timed hand waving, are not communicative and have little meaning. Although state-of-the-art systems employ speech and/or text information, they still do not handle semantic grounding of gestures properly, evidenced by gestures that seem to lack meaningful information when compared to the ground truth. Furthermore, due to the probabilistic nature of gestures, its idiosyncrasies, rich semantic content makes the evaluation process especially challenging and subjective. In this section, we discuss the limitations of the current work and possible future directions in context of what we view as the key challenges of gesture generation, namely:

1. evaluation (in Section 6.1),
2. data (in Section 6.2),
3. human-like gestures (in Section 6.3),
4. multimodal grounding (in Section 6.4), and
5. multimodal synthesis (in Section 6.5).

6.1. Evaluation

Evaluation is of central importance to gesture generation, both for developing co-speech gesture generation systems and for assessing their performance and capabilities in various aspects, as well as those of the field as a whole. However, evaluating gestures is challenging due to the stochastic nature of gestures and the highly subjective nature of human gesture perception. A comprehensive review of evaluation practices in gesture generation can be found in [WRB22]. We recommend that readers consult that review regarding best practices, but also provide an overview of key open challenges in gesture evaluation here.

6.1.1. Subjective Evaluation

One important aspect to evaluate for gesture-generation systems is the human-likeness of the generated gestures, which is measured and compared through human perceptual studies, often with comparable stimuli presented side by side as in e.g. [JYW*21, KJY*21, WGKB21]. On the other hand, evaluating the other aspects such as the appropriateness and/or specificity of generated gestures in the context of speech and other multimodal grounding information (see Section 6.4) is quite challenging, especially since differences in the human-likeness of the motions being compared tends to interfere with perceived gesture appropriateness (cf. the results in [KJY*21]). To alleviate this challenge for appropriateness, a new

evaluation paradigm of matched vs. mismatched gesture motion has recently been proposed [JKHB20, RGP21, YWK*22]. In this setup, human participants are asked to choose between two motion clips that both were generated by the same system, and therefore have similar appearance and human-likeness, but where one clip is intended to be appropriate to the situation (e.g., the motion in it corresponds to the actual speech audio in the video) whereas the other is chosen at random (e.g., it was generated by feeding unrelated speech audio into the same system instead, and does not match the actual audio track). The extent to which humans are able to identify the video that matches the situation can be used both to probe the strength of grounding in different modalities, and to assess gesture appropriateness for speech, rhythm, interlocutor behavior, etc., while controlling for human-likeness. We expect this methodology to gain wider adoption and advance the state of the art in subjective assessment of different aspects of co-speech gestures.

Another compelling area for future work is to evaluate gesture generation in actual interactions, since the ultimate goal of embodied conversational agents is to enhance human-computer communication and interaction. Initial studies [NKM*21, HPK22] have found that embodied agents that perform gestures generated by data driven models as opposed to performing no gestures, attract more attention from the audience. A larger attention span on a gesticulating agent is indicative of a more engaging communicative quality of gestures and opens doors to evaluating gesture generation in a more natural setting. Although the situated and time-demanding nature of such interactions, coupled with their reliance on many non-gesture components necessary to create interactivity (e.g. human wizards or automatic speech recognition, dialogue systems, and text-to-speech), make proper interactive evaluation challenging and seldom done, it is an important long-term goal for evaluations in the field. Given the difficulties in comparing different research papers in the field, we think that controlled, large-scale comparisons [KJY*21, YWK*22] with open data and materials are going to play an important role to develop the co-speech gesture field and its evaluation practices in the shorter term. This is similar to the role challenges have played in the development of text to speech [Kin14] and the wide use of leaderboards and benchmarks across deep learning today.

6.1.2. Objective Evaluation

While subjective metrics from appropriately designed user-studies are the gold standard in co-speech gesture evaluation [WRB22], they are expensive and time consuming, and thus lack scalability. There is therefore interest in objective metrics to automatically assess synthetic motion, for example its human-likeness. Objective metrics are useful to measure progress during model development in a heavy compute, data-driven learning setup. A natural metric is accuracy of prediction (i.e., how often the predicted position of a joint is within some tolerance of the joint position in a human motion capture clip), which is often called the Probability of Correct Keypoints (PCK). However, this quantity is often not indicative of performance due to the one-to-many nature of the gesture-generation problem. Two examples of human motion for the same speech might involve very different joint positions, and thus have low mutual agreement. Measuring the mean squared error (MSE)

between generated motion and human motion capture suffers from the same issue.

Statistics of motion properties such as acceleration and jerk have been used as an alternative for quantifying and comparing generated gesture distributions [KHK*21], but there is no compelling evidence that these metrics correlate with subjective assessments of motion human-likeness. To improve the measurement of distributional similarity of gestures, new objective quality metrics based on innovations from image processing, namely the Fréchet Inception Distance (FID) [HRU*17] and the Inception Score [SGZ*16], were proposed in [ALIM20, YCL*20] and [ALNM20] respectively. Among these proposals, only [YCL*20] computes the Fréchet distance in a learned space. There has also been work in learning to estimate the human-likeness of gestures from databases of gesture motion and associated subjective ratings data [He22]. However, learning to predict human preference can be difficult even from relatively large training databases, as seen in similar research into predicting the subjective ratings of synthetic speech [HCT*22].

Since the above approaches depend on motion data only, they can only give an indication of whether or not generated motion is statistically similar to the human motion capture in the database, but not how appropriate the motion is for the context in which it occurs (whether it is *grounded* in that context). The methods can therefore not assess whether or not the motion is synchronized with the co-occurring speech, whether the motion is semantically relevant, etc. In general, unlike human-likeness, not many techniques have been proposed for objectively quantifying properties like gesture diversity or different kinds of motion appropriateness. One exception is the recent Semantic Relevance Gesture Recall (SRGR) metric from [LZI*22], which proposes to quantify the semantic relevance of gesture by using semantic scores, annotated in the speech text data, to weight the probability of correct keypoints between the predicted and ground-truth gestures higher when the ground-truth gesture has a high semantic score. This is a step in the right direction for evaluating semantic appropriateness, but may suffer from the same issues as regular PCK due to the idiosyncratic, one-to-many nature of gesticulation. Given the impact that the Inception Score and the Fréchet Inception Distance have had in driving progress in image generation, reliable metrics that estimate gesture human-likeness and especially appropriateness for e.g. the rhythm and semantics of co-occurring speech are an important continuing challenge, where recent and future innovations are likely to have significant impact on the field.

6.2. Data

Compared to machine-learning applications in text, speech, and images, gesture-generation is currently a data-limited field. A particular bottleneck is finger motion, which is difficult to capture accurately even through motion capture; cf. Table 1. When finger motion is unreliable or unavailable, a possible mitigation might be to predict finger motion from other information, for example the rest of the body as in [NGDJ21]. In general, motion capture data is high quality, but laborious to capture, particularly when considering large scale data corpora. Other issues arise due to the high variation in gesture behavior. It can vary based on the individual, the environment, the number of people interacting, their emotional state and

the topic of the conversation. Some of this variation is grounded in information that cannot be effectively recorded because it, e.g., is internal to a speaker (such as their emotional state), or that is rarely captured, such as properties of the space in which an interaction is taking place. But even if one were to capture or control for many of these these sources of variation, a great diversity in gesture behavior and realization would persist, which will be difficult to cover in any database we can record.

In the long term, if we can achieve sufficiently reliable 3D gesture extraction from monocular, in-the-wild online video, that will be a game-changer for the field of co-speech gesture generation. It promises to have a transformative impact on both perceived authenticity and model capabilities, similar to how very large datasets for deep learning has powered recent advances in generative models for text and images, such as GPT-3 [BMR*20], DALL-E [RPG*21, RDN*22], and Stable Diffusion [RBL*22]. At present, works that study the use of in-the-wild data for gesture synthesis exist, for example [YKJ*19, GBK*19, YCL*20, HXM*21, ALIM20], but the quality of the data and the gestures do not yet amount to such a leap forward.

6.3. Human-Like Gestures

The most prominent research target in deep-learning-based co-speech gesture generation has long been perceptual quality. This is similar to the focus on perceptual surface quality in other areas such as image generation [KLA*20, RPG*21, RDN*22, RBL*22] and speech synthesis [WSRS*17, vdODZ*16]. One reason for this focus might be that perceptual surface quality is easier to estimate using standardized procedures, compared to quantities such as “gesture appropriateness for speech”. See, especially, the rapid quality improvements in the image-synthesis field, once reasonable objective metrics such as the Inception Score [SGZ*16] and the Fréchet Inception Distance [HRU*17] became available.

Just like deep generative methods in general have advanced greatly in recent years, there is strong evidence from large evaluations that the human-likeness of the best gesture-generation systems is improving as well [KJY*21, YWK*22]. The better the visual quality of the avatar and greater range of expressive motion, the easier it should be to spot differences between natural and synthetic motion. From this perspective, head motion (which only has three degrees of freedom) might for example be easier to make indistinguishable from human head motion, than it is to generate convincing arm and finger motion. In this light, the achievement of GestureMaster [ZBC22] in the GENEA Challenge 2022 [YWK*22] is particularly noteworthy, since the synthesized upper- and full-body gestures produced by this model were rated higher than the original motion capture from the human speaker. Although a very impressive result, this may partly be attributed to the presence of some motion clips with motion-capture artifacts, especially for the fingers, that may reduce the perceived human-likeness of the notional human reference motion.

At the same time, even “high quality” gesture motion on a high-fidelity avatar is still judged as being far from human: in the GENEA Challenge 2022 [YWK*22], neither the human motion capture nor the best performing system came near the rating of 100 that

would correspond to being “completely human-like”. More specifically, the median human-likeness of the best performing synthesis system were 69 for upper-body motion and 71 for full-body motion, with scores of 63 and 70 for human motion capture, respectively. Our statement comes with several caveats. Some of the gap up to a score of 100 might be attributable to shortcomings of motion capture when it comes to capturing the full range of human expression. For example, how an avatar moves and its lack of face, mouth, gaze and lip motion behavior can impact the visual qualities of the avatar. Even in the case of speech synthesis, where recreating human behavior is as easy as playing back an audio recording, it is well known that humans tend to rate the human-likeness of high-quality recordings of human speech as around or below 4.5 on a 5 point scale; see for example the naturalness scores in the large and careful evaluation in [?]. Complete human-likeness may thus in practice be achieved at a score below the maximum on the any given ratings scale. All that said, we believe that human-likeness can and will be improve further in the future, especially with more accurate motion capture and more lifelike avatars to display motion on.

As for the path that the gesture generation will take towards achieving new heights in human-likeness, we can look to history, and to other fields. Data-driven generative modeling like [GBK*19, FNM20, ALIM20, KJvW*20, YKJ*19] took over as the state of the art in co-speech gesture generation with the advent of publicly available motion capture datasets suitable for training deep-learning architectures. Since then, a variety of deep generative approaches have been applied (see Table 2), and human-likeness keeps improving [KJY*21, RGP21, YWK*22]. There is no doubt interesting work to come in applying recent diffusion models [SD-WMG15, SE19, HJA20], already considered for general motion synthesis [TRG*22], to gesture generation. While generated gestures from data-driven machine learning models are convincing, a lack of large scale gesture datasets currently limit the human-likeness of these approaches. Hence, in the short term, we may expect hybrid systems such as GestureMaster [ZBC22, AGL*22] to be the leaders in human-like gesture generation. Specifically, these are systems where machine-learning decides which general properties are needed of the gestures, but the actual gesture motion is primarily realized by assembling pre-recorded motion clips and frames, like in motion graphs [LCR*02, KGP02, AF02, TLP07] and motion matching [Cla16]. In the long-term, however, purely deep learning models are likely to take over. This would match the trajectory followed by text-to-speech synthesis, where hybrid systems once gave the best perceptual quality [Kin14], but pure deep-learning-based approaches trained on very large speech databases have recently taken the crown [TQSL21].

6.4. Multimodal Grounding

Visually human-like gesticulation is not the only goal of gesture generation. As discussed in the introduction to this article, a key goal with generating co-speech gestures is to facilitate communication, in much the same way as gestures enrich human communication. This requires gestures that not only exhibit human-like movement on the surface but also are appropriately *grounded* in the context of the interaction, so that they can contribute to it. In

more engineering-oriented terms, systems must take many relevant modalities as input, and make use of this information in an adequate way, to obtain synthetic gestures that can fulfill the same communicative roles as human gesticulation does. It can be difficult to capture this information both in training data and at synthesis time, as well as to make meaningful use of it in the gesture generation.

Grounding information can take many forms. Consequently, this section discusses challenges in grounding gesture-generation in a variety of relevant multimodal aspects (system inputs), beginning with aspects internal to the speaking agent, and then discussing grounding in other parties in the conversation as well as in the surrounding space. More specifically, we cover grounding in

1. temporal information (Section 6.4.1);
2. semantic content (Section 6.4.2);
3. speaker identity, personality, emotion, and style (Section 6.4.3);
4. interlocutor behavior (Section 6.4.4); and
5. spatial information (Section 6.4.5).

We also discuss some derived challenges posed by the often weak correlation between grounding information and the gesture motion (Section 6.4.6), and how gestures may be grounded in the creative intent of a system designer (Section 6.4.7).

6.4.1. Temporal Grounding

Gestures are temporal, which is a result of their correlation with a heavily temporal acoustic modality, along with the fact that they might depict occurrences or trace out paths or shapes over time. The rhythmic nature of the gestures (i.e. beat gestures) in context of acoustic prosody has been studied heavily since the era of rule based gesture synthesis [CMM99, MXL*13]. Fast forward to approaches with data-driven synthesis, some explicitly rely on extracted prosodic features [FNM21], while others [GBK*19, ALNM20] learn implicit embeddings from acoustics which prosody is one of the key components. It seems clear that gesture production must be grounded in the rhythm of audio data, and appropriate beat gestures will be challenging to achieve from text transcriptions alone, without timing information [KNN*22]. Alternatively, both audio and gesture must be synthesized to have comparable rhythmic structure.

6.4.2. Semantic Grounding

Beyond the rhythmic nature of gestures, there is often a semantic meaning associated with the performed gesture. The small size of gesture-generation databases, and the complicated relationship and weak correlation between speech semantics and gesture form (see Section 6.4.6), mean that it is unrealistic to expect systems to learn to generate semantically appropriate gestures driven by speech acoustics alone. Text, on the other hand, is a compact way to represent much of the semantic content behind co-speech gestures, and has been heavily studied since the era of rule-based gesture synthesis [CVB01] as well as in data-driven synthesis [SB19, LAM21, ALIM20, KJvW*20, YCL*20, ZYL*22, LFZ*22]. Current data-driven approaches typically attempt to gain semantic awareness by relying on deep-learning based language models trained on large amounts of text, such as [MCCD13, DCLT18]. Recent large language models based on large amounts of text

[BMR*20] have indeed been capable of generating text with surprisingly coherent semantics, suggesting that they can capture lexical meaning to a significant extent. While the inclusion of text has improved human perception of automatically generated gestures [KJvW*20, ALIM20, YCL*20, AGL*22], it is still not trivial to measure the semantic content of gestures (see the discussion in Section 6.1). Hence, it is unclear how much (if any) of the improved human perception can be attributed to the semantic awareness created due to the use of language models, nor how much of the bottlenecks that exist may be removed with continuing progress in neural language models. More broadly, there is a need for gesture synthesis models to perform better with regard to semantics. Gesture is most powerful when it conveys information, and doing this effectively has been a challenge for most deep learning systems; cf. Figure 4.

6.4.3. Identity, Style, Emotion, and Personality

Co-speech gestures are idiosyncratic. The manifold of gestures performed by a speaker are not just a function of the content of the speech, but are also dependent on the identity, emotional state and the context of the speaker. Generating personalized gestures based on speaker identity became possible with the influx of large scale multi-speaker datasets [YCL*20, ALNM20]. Several GENE Challenge 2022 [YWK*22] systems also make use of speaker identity. A deeper analysis of the impact of speaker identity input [KNN*22] shows that different speakers have different gesture-property prediction certainty, evoking even more interest in the idiosyncrasies of co-speech gestures. More recently, it was also shown that a short motion clip can be used for style control in “zero-shot style adaptation” [FGPO22, GFC22, GFH*]. For many applications, it is desirable for the designer to be able to control the nature of the motion. This goes beyond replicating idiosyncratic motion recorded of an individual to being able to specify novel characters. We are far from having ways to author a character with an imagined personality for a particular application.

Apart from the speaker identity, the emotional or affective state of a speaker also impacts the gestures performed by them. A striking example of this is the large range of expressive motion variation with the same lexical message explored in the Mimebot data [AONB17]. Building emotionally aware embodied agents is a common research direction [CBFV16, SZGK18]. More recently, data-driven models have been explored where affective cues were learned using a dedicated encoder in an adversarial setup [BCRM21] to imitate these patterns of affective behavior. It is important to be able to drive these emotions in a way that is consistent with a character’s personality and to be able to shift mood and emotion over time. One way forward might be to leverage findings from the literature of gesture and motion perception, which has identified many useful properties of gesture motion that correlate with the perception of personality [Lip98, KG10, SN17] and emotion [NLK*13, CN19]. By changing these properties in synthesized gestures, we may exert some control over the perceived speaker personality and emotion [AHKB20, HES*22]. Again, speech synthesis provides an analogy, where it was recently shown that simple and easy additions of filler words and pausing can meaningfully and reliably be used to alter listeners’ perception of speaker certainty [KLSG22].

6.4.4. Interlocutor-Aware Gestures

While non-verbal behavior is impacted by internal state of the speaker, the external context also guides the types of gestures a speaker might perform. In a dyadic conversation, the model must be aware of the behavior of the interlocutor while generating the relevant gestures [AMMS19, JKHB20, NC22, YYH20]. This includes modeling appropriate listener behavior as well as speaker behavior. Characters must modify their behavior to react to the content, mood and timing of interlocutors. Characters must be able to be surprised, angered, pleased, etc. based on what their interlocutor may say. Given the increasing availability of dyadic datasets with motion capture for both conversational parties, we expect to see more research in this direction in the next few years.

6.4.5. Spatially Aware Gestures

Even more generally, the context could also include spatial understanding of the environment. For example, the correctness of deictic gestures relies on the information about objects and directions in a scene. To carry communicative value, most of these gestures will therefore require access to visual and/or spatial information beyond what may be contained in the speech – think about a phrase such as “You need to go *that* way”, which completely lacks information about which direction the system should point. People also use spatial configurations in complex ways while gesturing, for example, placing ideas in a referential space in front of them and then referring to ideas by referring to the space they have been located in. While studies that involve external contexts are quite common for downstream tasks like navigation [SKM*19], non-verbal behavior generation in multiple external contexts is up and coming [DWAB22, KNN*22] which makes it a promising research direction, if relevant data can be obtained.

6.4.6. Weak Correlations with Grounding Information

Let’s imagine that we have access to all the variables discussed thus far that impact the dynamics of co-speech gestures, such as acoustics, text, speaker identity, emotional state, and external contexts. Further imagine that we are able to gather large-scale datasets with all these variables, which is unlikely to ever happen due to the combinatorial explosion of possible combinations of different factors. Would having this rich input information and broad data coverage be sufficient to confidently predict the specific co-speech gestures that a given speaker will perform? The best we can likely say is “Maybe!” While large scale datasets may enable us to minimize our epistemic uncertainty about gesticulation, it is unclear how significant the stochasticity is, i.e. aleatoric uncertainty, of these gestures will be. The situation is analogous to the problem of prosody in text-to-speech, where there can be many possible acoustic realizations and intonation contours for the same lexical input [WWK15, LTHY17, WSZ*18]. Significant variation persists even when a speaker is asked to read the same text several times under exactly the same circumstances [HMS*14]. To handle ambiguity in gesture realization, it is compelling to consider probabilistic models, since they can “hallucinate” the missing information and stochastic components of non-verbal behavior, as a

way to resolve the one-to-many problem for motion and gesture generation [HAB20, AHKB20].

6.4.7. Grounding Gestures in Creative Intent

Gesture authoring enables an animator or system creator to design and edit motion, e.g. making a character appear less nervous or stressed, thus grounding the animation within the designer’s creative intent. Typically, animation design intent is captured through key-framing or motion capture. However, these approaches are difficult to scale for nonverbal behavior because the former requires specialized animation skills, while the latter requires expensive camera setups and laborious post-processing. Automatic gesture generation approaches in part solve the scalability issue by the ability to generate abundant motion data, but they struggle with high-level control. For instance, attempts at handling control either bake in mechanistic, low-level control signals like wrist height, wrist velocity, and radial extent [AHKB20], or they generate gestures that deviate from the intended control specifications [HES*22]. Moreover, in multi-speaker scenarios, they are unable to capture the variability of different speakers’ gesticulation, and cannot distinguish between gesture types used in a certain scenario (e.g. deictic gestures for a lecturer in front of display) from gesture style differences between speakers [ALNM20]. Yoon et al. [YPJ*21] recently proposed an innovative approach to this challenge: an authoring toolkit that balances gesture quality and authoring effort. The toolkit combines automatic gesture generation using a GAN-based generative model [YCL*20] and manual controls. The generative model first produces a gesture sequence from speech input, and animator can interactively edit the motion through low-level pose control and coarse-level style parameters. We think similar gesture authoring approaches that maximize design intent and gesture quality, while minimizing authoring effort will be important for grounding non-verbal behavior within the animator’s creative intent.

6.5. Multimodal Synthesis

Human communicative behavior is not only grounded in multiple modalities and information streams, but is also expressed through multiple modalities. A complete virtual agent will need to listen, observe, decide, speak, and move. On the generation side, verbal behavior generation is considered separate from non-verbal behavior, and the generation of non-verbal behavior is in turn typically broken into several smaller sub-problems treated in isolation. Head motion might be treated separately from lip motion, facial expression, and gaze; finger motion might be treated separately from arm motion; and lower-body motion might be separated from the motion of the upper body. A long-term goal would be to bring these sub-problems together, to create more coherent synthetic behavior with a wider range of possible expressions, and eventually unify the synthesis of these expressions with verbal behavior generation. Recent work has explored learning full-body gesture motion (including the head and the lower body), e.g. [AHKB20] and the submissions to the full-body tier of the GENE Challenge 2022 [YWK*22].

Another line of work has considered training verbal (text-to-speech) and non-verbal (speech-to-gesture) synthesis systems on the same data [ASH*20] and, subsequently, merging them into one single network that generates both speech audio and gesture motion [WAG*21]. Given the strides that have been made in generating

convincing speech audio from text [TQSL21], adapting successful text-to-speech methods to simultaneously generate both acoustics and joint rotations, as was done in [WAG*21], seems like a compelling direction for future work. This not only brings advantages in terms of modeling efficiency (the gesture-generation systems will possess information about, e.g. prosodic prominence without having to learn to extract that information from speech audio), but also more closely resembles models of human communication such as the growth-point hypothesis [McN92a], and could enable gestures that not only complement but, as in Kendon’s continuum (see Figure 2), replace or augment speech with novel information. This may require even deeper representations of communicative intent, as approaches that generate gesture based on text and/or audio are restricted to redundant gestures, but gesture that is non-redundant with the spoken audio is a key part of human behavior.

7. Broader Impact

High quality gesture synthesis can advance a range of applications by allowing computational systems to leverage nonverbal communication. This can allow more natural and fluid communication of both functional and affective information, which will prove useful in a range of assistive applications, employing both agents and robots. These include tutors, rehabilitation trainers, relational agents for health and eldercare, and personal assistants. They can also support richly interactive entertainment experiences in which you can have meaningful interactions with virtual characters.

The development of the technology also raises potential ethical issues which must be given careful consideration. Some of the issues are common to many deep learning approaches that involve human data. For instance, what kind of bias is in the data that is used? Does it represent the full range of human nonverbal behavior, or only specific language groups, ethnicities and social strata? Will people using these models take care to match the input data with the desired output representation or will the data be mismatched, using the wrong gender, ethnicity, age, etc. on synthesized characters? What are the ownership rights associated with data that may be scraped from a web source? Do you own your gesture style? How can consent be obtained for online data?

The technology could also make it easier to generate deepfakes, i.e., synthetic media that mimics the likeness of real people, especially of politicians and other public figures that have a lot of video data online. Prominent examples include photorealistic lip motion from audio [SSKS17], real time facial expression re-enactment [TZS*16] and talking-head video synthesis [WML21]. The technology can be adapted to create synthetic nonverbal motion for nefarious purposes such as political propaganda, financial fraud and fake news. Moreover, a more unique consideration for nonverbal behavior results from people’s tendency to entrain to their interlocutors. If they entrain to synthetic models they may interact with, does this have any impact on their own behavior? It is important for both researchers and developers of this technology to devise ways to mitigate these risks.

8. Conclusion

This paper summarizes the history of gesture generation, from early work on rule-based systems to the explosion of recent work using deep learning approaches. Deep learning approaches have employed a range of input, including text, audio and various control signals, and used a wide set of architectures. Most systems have focused on monologue generation, but work is beginning to explore dialog and richer notions of context. Despite substantial progress, the field is still young and there are very significant challenges to solve. These include better datasets, improved subjective and objective evaluation practices, higher quality motion, producing more meaningful gestures, adequately addressing the stochasticity of gesture, providing adequate control over the output and matching the rich set of grounding that supports human gesture, from multi-person interaction to adequately representing the spatial context of the conversation. There is much exciting work to come.

Acknowledgments

S. N. was partially supported by an IBM PhD fellowship award. G. E. H. was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. S. N. and M. N. were partially supported by the National Science Foundation on grant IIS 2232066.

The authors are grateful to Stefan Kopp for Figure 4 and to Konrad Tollmar, and anonymous reviewers for reviewing the manuscript.

References

- [AF02] ARIKAN O., FORSYTH D. A.: Interactive motion generation from examples. *ACM Trans. Graph.* 21, 3 (2002), 483–490. doi: 10.1145/566570.566606. 8, 13, 18
- [AGL*22] AO T., GAO Q., LOU Y., CHEN B., LIU L.: Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. URL: <https://doi.org/10.1145/3550454.3555435>, doi:10.1145/3550454.3555435. 8, 10, 18, 19
- [AHKB20] ALEXANDERSON S., HENTER G. E., KUCHERENKO T., BESKOW J.: Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* 39, 2 (2020), 487–496. 4, 8, 9, 10, 13, 15, 19, 20
- [ALIM20] AHUJA C., LEE D. W., ISHII R., MORENCY L.-P.: No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (2020), pp. 1884–1895. 6, 7, 8, 10, 12, 17, 18, 19
- [ALM22] AHUJA C., LEE D. W., MORENCY L.-P.: Low-resource adaptation for personalized co-speech gesture generation. In *Proc. CVPR* (2022), pp. 20566–20576. 6, 10, 15
- [ALNM20] AHUJA C., LEE D. W., NAKANO Y. I., MORENCY L.-P.: Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision* (2020), Springer, pp. 248–265. 4, 9, 10, 13, 15, 17, 18, 19, 20
- [AMMS19] AHUJA C., MA S., MORENCY L.-P., SHEIKH Y.: To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *International Conference on Multimodal Interaction* (2019), pp. 74–84. 4, 10, 14, 19
- [AONB17] ALEXANDERSON S., O’SULLIVAN C., NEFF M., BESKOW J.: Mimebot – investigating the expressibility of non-verbal communication across agent embodiments. *ACM Transactions on Applied Perception (TAP)* 14, 4 (2017), 24:1–24:13. 19
- [ASH*20] ALEXANDERSON S., SZÉKELY E., HENTER G. E., KUCHERENKO T., BESKOW J.: Generating coherent spontaneous speech and gesture from text. In *Proc. IVA* (Glasgow, UK, Oct. 2020), vol. 20, ACM, pp. 1:1–1:3. doi:10.1145/3383652.3423874. 20
- [Bav94] BAVELAS J. B.: Gestures as part of speech: Methodological implications. *Research on language and social interaction* 27, 3 (1994), 201–221. 2
- [BBL*08] BUSO C., BULUT M., LEE C.-C., KAZEMZADEH A., MOWER E., KIM S., CHANG J. N., LEE S., NARAYANAN S. S.: Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359. 7
- [BCB14] BAHDANAU D., CHO K., BENGIO Y.: Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014). 11
- [BCRM21] BHATTACHARYA U., CHILDS E., REWKOWSKI N., MANOCHA D.: Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia* (New York, NY, USA, 2021), MM ’21, Association for Computing Machinery. 19
- [BDK16] BREAZEL C., DAUTENHAHN K., KANDA T.: Social robotics. *Springer handbook of robotics* (2016), 1935–1972. 2
- [BEY15] BOZKURT E., ERZIN E., YEMEZ Y.: Affect-expressive hand gestures synthesis and animation. In *2015 IEEE International Conference on Multimedia and Expo (ICME)* (2015), IEEE, pp. 1–6. 14
- [BGJM17] BOJANOWSKI P., GRAVE E., JOULIN A., MIKOLOV T.: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. 13
- [BHF*19] BHATTACHARYA A., HANSELMANN M., FRITZ M., SCHIELE B., STRAEHLE C.-N.: Conditional flow variational autoencoders for structured sequence prediction. *arXiv preprint arXiv:1908.09008* (2019). 10
- [BK09a] BERGMANN K., KOPP S.: GNetIc—using Bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents* (2009), Springer, pp. 76–89. 4, 7
- [BK09b] BERGMANN K., KOPP S.: Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1* (2009), pp. 361–368. 7
- [BK12] BERGMANN K., KOPP S.: Gestural alignment in natural dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2012), vol. 34. 14
- [BKK*17] BOZKURT E., KHAKI H., KEÇECİ S., TÜRKER B. B., YEMEZ Y., ERZIN E.: The jstkd database: an affective multimodal database of dyadic interactions. *Language Resources and Evaluation* 51, 3 (2017), 857–872. 10
- [BL93] BORKENAU P., LIEBLER A.: Consensus and self-other agreement for trait inferences from minimal information. *Journal of Personality* 61, 4 (1993), 477–496. 3
- [BMR*20] BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., ET AL.: Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901. 17, 19
- [BRB*21] BHATTACHARYA U., REWKOWSKI N., BANERJEE A., GUHAN P., BERA A., MANOCHA D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (2021), IEEE, pp. 1–10. 10, 11, 14

- [BYE20] BOZKURT E., YEMEZ Y., ERZIN E.: Affective synthesis and animation of arm gestures from speech prosody. *Speech Communication* 119 (2020), 1–11. 10, 14
- [CBC*00] CASSELL J., BICKMORE T., CAMPBELL L., VILHJALMSSON H., YAN H.: Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents* (2000), 29–63. 5
- [CBFV16] CHOWANDA A., BLANCHFIELD P., FLINTHAM M., VALSTAR M.: Computational models of emotion, personality, and social relationships for interactions in games. In *The 2016 International Conference on Autonomous Agents & Multiagent Systems* (2016). 19
- [CBK14] CIENKI A., BIETTI L. M., KOK K.: Multimodal alignment during collaborative remembering. *Memory Studies* 7, 3 (2014), 354–369. 14
- [CHS*18] CAO Z., HIDALGO G., SIMON T., WEI S.-E., SHEIKH Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. corr abs/1812.08008 (2018). *arXiv preprint arXiv:1812.08008* (2018). 11
- [Cla16] CLAVET S.: Motion matching and the road to next-gen animation. In *Proc. of GDC* (2016), vol. 2016. 18
- [CM11] CHIU C.-C., MARSELLA S.: How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents* (2011), Springer, pp. 127–140. 4, 8
- [CMM99] CASSELL J., MCNEILL D., MCCULLOUGH K.-E.: Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition* 7, 1 (1999), 1–34. 18
- [CMM15] CHIU C.-C., MORENCY L.-P., MARSELLA S.: Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents* (2015), Springer. 4, 10, 12
- [CN19] CASTILLO G., NEFF M.: What do we express without knowing?: Emotion in gesture. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (2019), International Foundation for Autonomous Agents and Multiagent Systems, pp. 702–710. 2, 19
- [CPB*94] CASSELL J., PELACHAUD C., BADLER N., STEEDMAN M., ACHORN B., BECKET T., DOUVILLE B., PREVOST S., STONE M.: Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques* (1994), pp. 413–420. 4, 5
- [CPPS04] CAROLIS B. D., PELACHAUD C., POGGI I., STEEDMAN M.: Apml, a markup language for believable behavior generation. In *Life-like characters*. Springer, 2004, pp. 65–85. 5
- [CTL*21] CHEN K., TAN Z., LEI J., ZHANG S.-H., GUO Y.-C., ZHANG W., HU S.-M.: Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13. 13
- [CVB01] CASSELL J., VILHJALMSSON H. H., BICKMORE T.: BEAT: The behavior expression animation toolkit. In *Proc. SIGGRAPH* (2001), pp. 477–486. 4, 5, 18
- [CVMG*14] CHO K., VAN MERRIENBOER B., GULCEHRE C., BAH-DANAU D., BOUGARES F., SCHWENK H., BENGIO Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014). 11
- [Dav18] DAVIS R. O.: The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review* 24 (2018), 193–209. 2
- [DCLT18] DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018). 12, 18
- [DGS13] DAEL N., GOUDBEEK M., SCHERER K. R.: Perceived gesture dynamics in nonverbal expression of emotion. *Perception* 42, 6 (2013), 642–657. 2
- [DKD*16] DURUPINAR F., KAPADIA M., DEUTSCH S., NEFF M., BADLER N. I.: Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Transactions on Graphics (TOG)* 36, 1 (2016), 1–16. 2
- [DWAB22] DEICHLER A., WANG S., ALEXANDERSON S., BESKOW J.: Towards context-aware human-like pointing gestures with rl motion imitation. In *Workshop on Context-Awareness in Human-Robot Interaction* (2022), HRI '22 workshop. URL: <https://www.diva-portal.org/smash/record.jsf?pid=diva2:1664509>. 10, 14, 19
- [EF69] EKMAN P., FRIESEN W. V.: Nonverbal leakage and clues to deception. *Psychiatry* 32, 1 (1969), 88–106. 3
- [Ekm92] EKMAN P.: An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200. 3
- [FGPO22] FARES M., GRIMALDI M., PELACHAUD C., OBIN N.: Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *arXiv preprint arXiv:2208.01917* (2022). 6, 10, 13, 15, 19
- [FM18] FERSTL Y., MCDONNELL R.: Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (2018), ACM, pp. 93–98. 4, 6, 7, 10
- [FNM19] FERSTL Y., NEFF M., MCDONNELL R.: Multi-objective adversarial gesture generation. In *ACM SIGGRAPH Conference on Motion, Interaction and Games* (2019), pp. 3:1–3:10. 4, 9, 10
- [FNM20] FERSTL Y., NEFF M., MCDONNELL R.: Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics* 89 (2020), 117–130. 8, 9, 18
- [FNM21] FERSTL Y., NEFF M., MCDONNELL R.: Expressgesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* (2021), e2016. 7, 10, 11, 18
- [FP16] FOURATI N., PELACHAUD C.: Perception of emotions and body movement in the emilya database. *IEEE Transactions on Affective Computing* 9, 1 (2016), 90–101. 2
- [GAL*14] GRATCH J., ARTSTEIN R., LUCAS G., STRATOU G., SCHERER S., NAZARIAN A., WOOD R., BOBERG J., DEVAULT D., MARSELLA S., TRAUM D., RIZZO S., MORENCY L.-P.: The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (2014). 10
- [GBK*19] GINOSAR S., BAR A., KOHAVI G., CHAN C., OWENS A., MALIK J.: Learning individual styles of conversational gesture. In *IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 3497–3506. 4, 7, 9, 10, 15, 17, 18
- [GEB15] GATYS L. A., ECKER A. S., BETHGE M.: A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015). 13
- [GFC22] GHORBANI S., FERSTL Y., CARBONNEAU M.-A.: Exemplar-based stylized gesture generation from speech: An entry to the genea challenge 2022. In *Proceedings of the ACM International Conference on Multimodal Interaction* (2022), ICMI '22, ACM. 10, 15, 19
- [GFD*15] GIRAUD T., FOCONE F., DEMULIER V., MARTIN J. C., IS-ABLEU B.: Perception of emotion and personality through full-body movement qualities: a sport coach case study. *ACM Transactions on Applied Perception (TAP)* 13, 1 (2015), 1–27. 2
- [GFH*] GHORBANI S., FERSTL Y., HOLDEN D., TROJE N. F., CARBONNEAU M.-A.: Zeroeggs: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum*. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgfm.14734>, doi:<https://doi.org/10.1111/cgfm.14734>. 6, 7, 10, 15, 19

- [GHLY18] GO D.-S., HYUNG H.-J., LEE D.-W., YOON H. U.: Android robot motion generation based on video-recorded human demonstrations. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (2018), IEEE, pp. 476–478. 4
- [GM99] GOLDIN-MEADOW S.: The role of gesture in communication and thinking. *Trends in cognitive sciences* 3, 11 (1999), 419–429. 2
- [GM05] GOLDIN-MEADOW S.: *Hearing gesture: How our hands help us think*. Harvard University Press, 2005. 2
- [GMM99] GOLDIN-MEADOW S., MCNEILL D.: The role of gesture and mimetic representation in making language the province of speech. In *The Descent of Mind: Psychological Perspectives on Hominid Evolution*. Oxford University Press, 03 1999. doi:10.1093/acprof:oso/9780192632593.003.0009. 2
- [GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Red Hook, NY, USA, 2014), NIPS'14, Curran Associates, Inc., pp. 2672–2680. URL: <http://papers.nips.cc/paper/5423-generative-adversarial-nets>. 9
- [HAB20] HENTER G. E., ALEXANDERSON S., BESKOW J.: MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Trans. Graph.* 39, 4 (2020), 236:1–236:14. doi:10.1145/3414685.3417836. 9, 15, 20
- [HCT*22] HUANG W. C., COOPER E., TSAO Y., WANG H.-M., TODA T., YAMAGISHI J.: The VoiceMOS Challenge 2022. In *Proc. Interspeech* (2022), pp. 4536–4540. doi:10.21437/Interspeech.2022-970. 17
- [He22] HE Z.: Automatic quality assessment of speech-driven synthesized gestures. doi:10.1155/2022/1828293. 17
- [HES*22] HABIBIE I., ELGHARIB M., SARKAR K., ABDULLAH A., NYATSANGA S., NEFF M., THEOBALT C.: A motion matching-based framework for controllable gesture synthesis from speech. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings* (2022), pp. 1–9. 5, 9, 10, 15, 19, 20
- [HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems* (2020), pp. 6840–6851. 18
- [HKPP20] HOLDEN D., KANOUN O., PEREPICHKA M., POPA T.: Learned motion matching. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 53–1. 11
- [HKS17] HOLDEN D., KOMURA T., SAITO J.: Phase-functioned neural networks for character control. *ACM Trans. Graph.* 36, 4 (2017), 42:1–42:13. doi:10.1145/3072959.3073663. 15
- [HKS*18] HASEGAWA D., KANEKO N., SHIRAKAWA S., SAKUTA H., SUMI K.: Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *International Conference on Intelligent Virtual Agents* (2018), ACM, pp. 79–86. 4, 9, 10
- [HLW*18] HAN S., LIU B., WANG R., YE Y., TWIGG C. D., KIN K.: Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–10. 15
- [HM05] HOWARD R. A., MATHESON J. E.: Influence diagrams. *Decision Analysis* 2, 3 (2005), 127–143. 7
- [HMS*14] HENTER G. E., MERRITT T., SHANNON M., MAYO C., KING S.: Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. In *Proc. Interspeech* (Singapore, 2014), ISCA, pp. 1504–1508. doi:10.21437/Interspeech.2014-361. 19
- [Hos11] HOSTETTER A. B.: When do gestures communicate? a meta-analysis. *Psychological bulletin* 137, 2 (2011), 297. 2
- [HPK22] HE Y., PEREIRA A., KUCHERENKO T.: Evaluating data-driven co-speech gestures of embodied conversational agents through real-time interaction. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents* (2022), IVA '22, ACM, pp. 8:1–8:8. doi:10.1145/3514197.3549697. 16
- [HRU*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017). 17
- [HS97] HOCHREITER S., SCHMIDHUBER J.: Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735. 9
- [HS16] HOLLADAY R. M., SRINIVASA S. S.: Rogue: Robot gesture engine. In *AAAI Spring Symposia* (2016). 4
- [HXM*21] HABIBIE I., XU W., MEHTA D., LIU L., SEIDEL H.-P., PONS-MOLL G., ELGHARIB M., THEOBALT C.: Learning speech-driven 3d conversational gestures from video. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents* (2021), pp. 101–108. 6, 7, 10, 17
- [IMMI18] ISHI C. T., MACHIYASHIKI D., MIKATA R., ISHIGURO H.: A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robot. Autom. Lett.* 3, 4 (2018), 3757–3764. 10, 11
- [JKEB19] JONELL P., KUCHERENKO T., EKSTEDT E., BESKOW J.: Learning non-verbal behavior for a social robot from YouTube videos. In *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions* (Oslo, Norway, 2019). 6
- [JKHB20] JONELL P., KUCHERENKO T., HENTER G. E., BESKOW J.: Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the International Conference on Intelligent Virtual Agents* (2020), ACM. 14, 16, 19
- [JSCS19] JOO H., SIMON T., CIKARA M., SHEIKH Y.: Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 10873–10883. 6
- [JSL*17] JOO H., SIMON T., LI X., LIU H., TAN L., GUI L., BANERJEE S., GODISART T. S., NABBE B., MATTHEWS I., KANADE T., NOBUHARA S., SHEIKH Y.: Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017). 7
- [JYN*20] JÖRG S., YE Y., NEFF M., MUELLER F., ZORDAN V.: Virtual hands in vr: Motion capture, synthesis, and perception. In *ACM SIGGRAPH 2020 Courses* (New York, NY, USA, 2020), SIGGRAPH '20, Association for Computing Machinery. URL: <https://doi.org/10.1145/3388769.3407494>, doi:10.1145/3388769.3407494. 15
- [JYW*21] JONELL P., YOON Y., WOLFERT P., KUCHERENKO T., HENTER G. E.: HEMVIP: Human evaluation of multiple videos in parallel. In *Proceedings of the International Conference on Multimodal Interaction* (2021). 16
- [KDZ20] KORZUN V., DIMOV I., ZHARKOV A.: The finemove entry to the genea challenge 2020. In *Proc. GENE Workshop* (Oct. 2020). doi:10.5281/zenodo.4088609. 12
- [KDZ21] KORZUN V., DIMOV I., ZHARKOV A.: Audio and text-driven approach for conversational gestures generation. In *Proceedings of Computational Linguistics and Intellectual Technologies* (2021), DIALOGUE '21. URL: <http://www.dialog-21.ru/media/5526/korzunvaplusdimovinpluszharkovaa031.pdf>. 10, 12
- [Ken72] KENDON A.: Some relationships between body motion and speech. *Studies in dyadic communication* 7, 177 (1972), 90. 3, 4
- [Ken83] KENDON A.: Gesture and Speech. How They Interact. In Weinmann, J. M., and Harrison, R. P., eds. *Nonverbal Interaction* (1983), 13–45. 3
- [Ken88] KENDON A.: How gestures can become like words. In *Cross-Cultural Perspectives in Nonverbal Communication* (1988), Hogrefe & Huber Publishers. 3

- [Ken94] KENDON A.: Do gestures communicate? a review. *Research on language and social interaction* 27, 3 (1994), 175–200. 2
- [KG10] KOPPENSTEINER M., GRAMMER K.: Motion patterns in political speech and their influence on personality ratings. *J. Res. Pers.* 44, 3 (2010), 374–379. 19
- [KGP02] KOVAR L., GLEICHER M., PIGHIN F.: Motion graphs. *ACM Trans. Graph.* 21, 3 (2002), 473–482. doi:10.1145/566654.566605. 8, 13, 18
- [KHH*19] KUCHERENKO T., HASEGAWA D., HENTER G. E., KANEKO N., KJELLSTRÖM H.: Analyzing input and output representations for speech-driven gesture generation. In *International Conference on Intelligent Virtual Agents* (Paris, France, July 2019), ACM, pp. 97–104. doi:10.1145/3308532.3329472. 4, 8, 9, 10, 12
- [KHK*21] KUCHERENKO T., HASEGAWA D., KANEKO N., HENTER G. E., KJELLSTRÖM H.: Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human-Computer Interaction* (2021). doi:10.1080/10447318.2021.1883883. 4, 17
- [Kin14] KING S.: Measuring a decade of progress in text-to-speech. *Loquens* 1, 1 (2014), e006. doi:10.3989/loquens.2014.006. 16, 18
- [Kip01] KIPP M.: Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology* (2001). 6, 7
- [Kip05] KIPP M.: *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005. 2, 3, 4, 6, 7
- [KJLW03] KOPP S., JUNG B., LESSMANN N., WACHSMUTH I.: Max-a multimodal assistant in virtual reality construction. *KI* 17, 4 (2003), 11. 5
- [KJvW*20] KUCHERENKO T., JONELL P., VAN WAVEREN S., HENTER G. E., ALEXANDERSON S., LEITE I., KJELLSTRÖM H.: Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction* (2020). 4, 8, 10, 12, 18, 19
- [KJY*20] KUCHERENKO T., JONELL P., YOON Y., WOLFERT P., HENTER G. E.: The GENE Challenge 2020: Benchmarking gesture-generation systems on common data. In *Proceedings of the International Workshop on Generation and Evaluation of Non-Verbal Behaviour for Embodied Agents* (2020), GENE '20. URL: <https://genea-workshop.github.io/2020/>. 9, 10, 12
- [KJY*21] KUCHERENKO T., JONELL P., YOON Y., WOLFERT P., HENTER G. E.: A large, crowdsourced evaluation of gesture generation systems on common data. In *Proceedings of the Annual Conference on Intelligent User Interfaces* (2021). Accepted for publication. 7, 16, 17, 18
- [KKM*06] KOPP S., KRENN B., MARSELLA S., MARSHALL A. N., PELACHAUD C., PIRKER H., THÓRISSON K. R., VILHJÁLMSOHN H.: Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents* (2006), Springer, pp. 205–217. 3, 6
- [KKW02] KRANSTEDT A., KOPP S., WACHSMUTH I.: Murml: A multimodal utterance representation markup language for conversational agents. In *AAMAS'02 Workshop Embodied conversational agents-let's specify and evaluate them!* (2002). 5
- [KLA*20] KARRAS T., LAINE S., AITTALA M., HELLSTEN J., LEHTINEN J., AILA T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 8110–8119. 17
- [KLSG22] KIRKLAND A., LAMERIS H., SZÉKELY E., GUSTAFSON J.: Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence. *Proc. Interspeech 2022* (2022), 4990–4994. 19
- [KNN*22] KUCHERENKO T., NAGY R., NEFF M., KJELLSTRÖM H., HENTER G. E.: Multimodal analysis of the predictability of hand-gesture properties. In *Proceedings of the IEEE conference on Virtual Reality and 3D User Interfaces* (2022), AAMAS '22, IFAAMAS, pp. 770–779. 5, 6, 7, 8, 10, 13, 18, 19
- [KPB20] KOBYZEV I., PRINCE S. J., BRUBAKER M. A.: Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 3964–3979. 9
- [KW02] KOPP S., WACHSMUTH I.: Model-based animation of co-verbal gesture. In *Proceedings of Computer Animation 2002 (CA 2002)* (2002), IEEE, pp. 252–257. 4, 5
- [KW04] KOPP S., WACHSMUTH I.: Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds* 15, 1 (2004), 39–52. 4, 7
- [KW13] KINGMA D. P., WELING M.: Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* (2013). 9
- [LAM21] LEE D. W., AHUJA C., MORENCY L.-P.: Crossmodal clustered contrastive learning: Grounding of spoken language to gesture. In *Companion Publication of the 2021 International Conference on Multimodal Interaction* (2021), ICMI '21 Companion, ACM, pp. 202–210. doi:10.1145/3461615.3485408. 10, 13, 18
- [LB85] LIBERMAN M., BUCHSBAUM A.: *Structure and usage of current Bell Labs text to speech programs*. Tech. rep., Technical Memorandum TM 11225-850731-11, AT&T Bell Laboratories, 1985. 5
- [LBH*13] LÜCKING A., BERGMAN K., HAHN F., KOPP S., RIESER H.: Data-based analysis of speech and gesture: The Bielefeld speech and gesture alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces* 7, 1 (2013), 5–18. 7, 13
- [LCR*02] LEE J., CHAI J., REITSMA P. S. A., HODGINS J. K., POLLARD N. S.: Interactive control of avatars animated with human motion data. 491–500. 8, 13, 18
- [LDM*19] LEE G., DENG Z., MA S., SHIRATORI T., SRINIVASA S. S., SHEIKH Y.: Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 763–772. 6, 7, 15
- [LFZ*22] LIANG Y., FENG Q., ZHU L., HU L., PAN P., YANG Y.: Seeg: Semantic energized co-speech gesture generation. In *Proc. CVPR* (2022), pp. 10473–10482. 10, 12, 18
- [LHP11] LE Q. A., HANOUNE S., PELACHAUD C.: Design and implementation of an expressive gesture model for a humanoid robot. In *2011 11th IEEE-RAS International Conference on Humanoid Robots* (2011), IEEE, pp. 134–140. 4, 6
- [Lip98] LIPPA R.: The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: A lens model analysis. *J. Res. Pers.* 32, 1 (1998), 80–107. 19
- [LKP*21] LI J., KANG D., PEI W., ZHE X., ZHANG Y., HE Z., BAO L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/VCF International Conference on Computer Vision* (2021), pp. 11293–11302. 4, 9, 10
- [LKTK10] LEVINE S., KRÄHENBÜHL P., THRUN S., KOLTUN V.: Gesture controllers. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 124. 4, 8
- [LM06] LEE J., MARSELLA S.: Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents* (2006), Springer, pp. 243–255. 4
- [LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16. 11
- [LMSJ21] LIU Y., MOHAMMADI G., SONG Y., JOHAL W.: Speech-based gesture generation for robots and embodied agents: A scoping review. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (2021), pp. 31–38. 2

- [LTHY17] LUONG H.-T., TAKAKI S., HENTER G. E., YAMAGISHI J.: Adapting and controlling DNN-based speech synthesis using input codes. In *Proc. ICASSP* (2017), pp. 4905–4909. doi:10.1109/ICASSP.2017.7953089. 19
- [LTK09] LEVINE S., THEOBALT C., KOLTUN V.: Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.* 28, 5 (Dec. 2009), 1–10. 8
- [LWH*12] LEVINE S., WANG J. M., HARAUX A., POPOVIĆ Z., KOLTUN V.: Continuous character control with low-dimensional embeddings. *ACM Trans. Graph.* 31, 4 (2012), 28:1–28:10. doi:10.1145/2185520.2185524. 12
- [LZCvdP20] LING H. Y., ZINNO F., CHENG G., VAN DE PANNE M.: Character controllers using motion VAEs. *ACM Trans. Graph.* 39, 4 (2020). doi:10.1145/3386569.3392422. 15
- [LZI*22] LIU H., ZHU Z., IWAMOTO N., PENG Y., LI Z., ZHOU Y., BOZKURT E., ZHENG B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297* (2022). 7, 17
- [MCCD13] MIKOLOV T., CHEN K., CORRADO G., DEAN J.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013). 18
- [McN92a] MCNEILL D.: *Hand and Mind: What Gestures Reveal about Thought*. 1992. 3, 20
- [McN92b] MCNEILL D.: *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992. 3
- [McN00] MCNEILL D.: *Language and Gesture*. Language Culture and Cognition. Cambridge University Press, 2000. doi:10.1017/CBO9780511620850. 3
- [McN05] MCNEILL D.: *Gesture and thought*. University of Chicago press, 2005. 3
- [Meh68] MEHRABIAN A.: Some referents and measures of nonverbal behavior. *Behavior Research Methods & Instrumentation* 1, 6 (1968), 203–207. 3
- [MXL*13] MARSELLA S., XU Y., LHOMMET M., FENG A., SCHERER S., SHAPIRO A.: Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2013), pp. 25–35. 4, 6, 18
- [MYL*16] METALLINO A., YANG Z., LEE C.-C., BUSSO C., CARNICKE S., NARAYANAN S.: The usc creativeit database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Language resources and evaluation* 50, 3 (2016), 497–521. 6, 7, 10
- [NC22] NGUYEN T. V. T., CELIKTUTAN O.: Context-aware body gesture generation for social robots. In *ICRA 2022 Workshop on Prediction and Anticipation Reasoning for Human-Robot Interaction* (2022). 10, 14, 19
- [Nef16] NEFF M.: Hand gesture synthesis for conversational characters. *Handbook of Human Motion* (2016), 1–12. 4
- [NGDJ21] NG E., GINOSAR S., DARRELL T., JOO H.: Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proc. CVPR* (2021), pp. 11865–11874. 10, 15, 17
- [NGM15] NOVACK M., GOLDIN-MEADOW S.: Learning from gesture: How our hands change our minds. *Educational psychology review* 27, 3 (2015), 405–412. 2
- [NKAS08] NEFF M., KIPP M., ALBRECHT I., SEIDEL H.-P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24. 4, 7
- [NKM*21] NAGY R., KUCHERENKO T., MOELL B., PEREIRA A., KJELLSTRÖM H., BERNARDET U.: A framework for integrating gesture generation models into interactive conversational agents. In *Proceedings of the IEEE conference on Virtual Reality and 3D User Interfaces* (2021), AAMAS '21, IFAAMAS, pp. 1779–1781. 16
- [NLK*13] NORMOYLE A., LIU F., KAPADIA M., BADLER N. I., JÖRG S.: The effect of posture and dynamics on the perception of emotion. In *Proc. SAP* (2013), pp. 91–98. 2, 19
- [Nob00] NOBE S.: Where do most spontaneous representational gestures actually occur with respect to speech. *Language and gesture* 2 (2000), 186. 4
- [NTB*11] NEFF M., TOOTHMAN N., BOWMAN R., FOX TREE J. E., WALKER M. A.: Don't scratch! self-adaptors reflect emotional stability. In *International Workshop on Intelligent Virtual Agents* (2011), Springer, pp. 398–411. 2, 3
- [NTHLO10] NG-THOW-HING V., LUO P., OKITA S.: Synchronized gesture and speech production for humanoid robots. In *International Conference on Intelligent Robots and Systems* (2010), IEEE/RSJ. 4
- [NWAU10] NEFF M., WANG Y., ABBOTT R., WALKER M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents* (2010), Springer, pp. 222–235. 2
- [OB16] OBEN B., BRÔNE G.: Explaining interactive alignment: A multimodal and multifactorial account. *Journal of Pragmatics* 104 (2016), 32–51. 14
- [PALvdP18] PENG X. B., ABBEEL P., LEVINE S., VAN DE PANNE M.: Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14. 14
- [PB03] PELACHAUD C., BILVI M.: Computational model of believable conversational agents. In *Communication in multiagent systems*. Springer, 2003, pp. 300–317. 4
- [PCDC*02] PELACHAUD C., CAROFIGLIO V., DE CAROLIS B., DE ROSIS F., POGGI I.: Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2* (2002), pp. 758–765. 4, 5
- [PKS*04] PIWEK P., KRENN B., SCHRÖDER M., GRICE M., BAUMANN S., PIRKER H.: Rrl: A rich representation language for the description of agent behaviour in neca. *arXiv preprint cs/0410022* (2004). 5
- [PMA*21] PENG X. B., MA Z., ABBEEL P., LEVINE S., KANAZAWA A.: Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–20. 14
- [PNR*19] PAPAMAKARIOS G., NALISNICK E., REZENDE D. J., MOHAMED S., LAKSHMINARAYANAN B.: Normalizing flows for probabilistic modeling and inference, 2019. *arXiv:1912.02762*. 9
- [PSM14] PENNINGTON J., SOCHER R., MANNING C.: GloVe: Global vectors for word representation. In *Proc. EMNLP* (2014), pp. 1532–1543. 11
- [QTZ*21] QIAN S., TU Z., ZHI Y., LIU W., GAO S.: Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/VCF International Conference on Computer Vision* (2021), pp. 11077–11086. 10, 15
- [RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695. 17
- [RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022). 17
- [RG91] RAO A. S., GEORGEFF M. P.: Modeling rational agents within a bdi-architecture. *KR* 91 (1991), 473–484. 5
- [RGP21] REBOL M., GÜTI C., PIETROSZEK K.: Passing a non-verbal turing test: Evaluating gesture animations generated from speech. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces* (2021), VR '21, IEEE, pp. 573–581. doi:10.1109/VR50410.2021.00082. 9, 10, 16, 18

- [RK97] REITHINGER N., KLESEN M.: Dialogue act classification using language models. In *Fifth European Conference on Speech Communication and Technology* (1997). 7
- [RPCM18] RAVENET B., PELACHAUD C., CLAVEL C., MARSELLA S.: Automating the production of communicative gestures in embodied characters. *Frontiers in Psychology* 9 (2018). 6
- [RPG*21] RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M., SUTSKEVER I.: Zero-shot text-to-image generation. In *International Conference on Machine Learning* (2021), PMLR, pp. 8821–8831. 17
- [RR77] RUBIN S. M., REDDY R.: The locus model of search and its use in image interpretation. In *IJCAI* (1977), vol. 2, pp. 590–595. 11
- [RSD81] RUTTER D., STEPHENSON G., DEWEY M.: Visual communication and the content and style of conversation. *British Journal of Social Psychology* 20, 1 (1981), 41–52. 2
- [SB18] SADOUGHI N., BUSO C.: Novel realizations of speech-driven head movements with generative adversarial networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (2018), IEEE, pp. 6169–6173. 9
- [SB19] SADOUGHI N., BUSO C.: Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100. 10, 14, 18
- [SBM21] SAUND C., BÎRLĂDEANU A., MARSELLA S.: Cmcf: An architecture for realtime gesture generation by clustering gestures by motion and communicative function. In *Proceedings of the IEEE conference on Virtual Reality and 3D User Interfaces* (2021), AAMAS '21, IFAA-MAS, pp. 1136–1144. 10, 13
- [SCNW19] SMITH H. J., CAO C., NEFF M., WANG Y.: Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2 (2019). doi: 10.1145/3340254. 15
- [SDWMG15] SOHL-DICKSTEIN J., WEISS E., MAHESWARANATHAN N., GANGULI S.: Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning* (2015), pp. 2256–2265. 18
- [SE19] SONG Y., ERMON S.: Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems* (2019). 18
- [SER*13] SALEM M., EYSSEL F., ROHLFING K., KOPP S., JOUBLIN F.: To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323. 2
- [SGZ*16] SALIMANS T., GOODFELLOW I., ZAREMBA W., CHEUNG V., RADFORD A., CHEN X.: Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016). 17
- [SKM*19] SAVVA M., KADIAN A., MAKSYMETS O., ZHAO Y., WIMJANS E., JAIN B., STRAUB J., LIU J., KOLTUN V., MALIK J., ET AL.: Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9339–9347. 19
- [SKW*12] SALEM M., KOPP S., WACHSMUTH I., ROHLFING K., JOUBLIN F.: Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics* 4, 2 (2012), 201–217. 2
- [SKWJ09] SALEM M., KOPP S., WACHSMUTH I., JOUBLIN F.: Towards meaningful robot gesture. In *Human centered robot systems*. Springer, 2009, pp. 173–182. 4
- [SLB15] SADOUGHI N., LIU Y., BUSO C.: Msp-avatar corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2015), vol. 7, IEEE, pp. 1–6. 10
- [SN17] SMITH H. J., NEFF M.: Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12. 2, 19
- [SN19] SAUNDERSON S., NEJAT G.: How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11, 4 (2019), 575–608. 2
- [SSKS17] SUWAJANAKORN S., SEITZ S. M., KEMELMACHER-SHLIZERMAN I.: Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 95. 20
- [SZGK18] SOHN S. S., ZHANG X., GERACI F., KAPADIA M.: An emotionally aware embodied conversational agent. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (2018), pp. 2250–2252. 19
- [TH09] TAYLOR G. W., HINTON G. E.: Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning* (2009), pp. 1025–1032. 8
- [Th696] THÓRISSON K. R.: *Communicative humanoids: a computational model of psychosocial dialogue skills*. PhD thesis, Massachusetts Institute of Technology, 1996. 5
- [TKS*17] TAKEUCHI K., KUBOTA S., SUZUKI K., HASEGAWA D., SAKUTA H.: Creating a gesture-speech dataset for speech-based automatic gesture generation. In *Proceedings of the International Conference on Human-Computer Interaction* (2017), Springer, pp. 198–202. 6, 7, 9, 10
- [TLP07] TREUILLE A., LEE Y., POPOVIĆ Z.: Near-optimal character animation with continuous control. In *ACM SIGGRAPH 2007 papers*. 2007, pp. 7–es. 8, 18
- [TMMK08] THIEBAUX M., MARSELLA S., MARSHALL A. N., KALLMANN M.: Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1* (2008), pp. 151–158. 4, 6
- [TQSL21] TAN X., QIN T., SOONG F., LIU T.-Y.: A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561* (2021). 18, 20
- [TRG*22] TEVET G., RAAB S., GORDON B., SHAFIR Y., BERMANO A. H., COHEN-OR D.: Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022). 18
- [Tui93] TUIE K.: The production of gesture. *Semiotica* (1993), 83–105. 3
- [Tve07] TVERSKY B.: Communicating with diagrams and gestures. *Research trends in science, technology, and mathematics education* (2007). 2
- [TWGM21] TAYLOR S., WINDLE J., GREENWOOD D., MATTHEWS I.: Speech-driven conversational agents using conditional flow-vaes. In *Proceedings of the ACM European Conference on Visual Media Production* (2021), CVMP '21, ACM, pp. 6:1–6:9. doi:10.1145/3485441.3485647. 10
- [TZS*16] THIES J., ZOLLHOFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2387–2395. 20
- [VCC*07] VILHJÁLMSOHN H., CANTELMO N., CASSELL J., E CHAFAI N., KIPP M., KOPP S., MANCINI M., MARSELLA S., MARSHALL A. N., PELACHAUD C., ET AL.: The behavior markup language: Recent developments and challenges. In *International Workshop on Intelligent Virtual Agents* (2007), Springer, pp. 99–111. 6
- [VDLRBM14] VOLKOVA E., DE LA ROSA S., BÜLTHOFF H. H., MOHLER B.: The mpi emotional body expressions database for narrative scenarios. *PLoS one* 9, 12 (2014), e113647. 7, 10, 11, 14
- [vdODZ*16] VAN DEN OORD A., DIELEMAN S., ZEN H., SIMONYAN K., VINYALS O., GRAVES A., KALCHBRENNER N., SENIOR A., KAVUKCUOĞLU K.: WaveNet: A generative model for raw audio, 2016. arXiv:1609.03499. 17
- [VMD*14] VOLKOVA E. P., MOHLER B. J., DODDS T. J., TESCH J., BÜLTHOFF H. H.: Emotion categorization of body expressions in narrative scenarios. *Frontiers in psychology* 5 (2014), 623. 2

- [VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. In *Advances in Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates, Inc., pp. 5998–6008. URL: <https://papers.nips.cc/paper/7181-attention-is-all-you-need>. 11, 12, 14
- [WAG*21] WANG S., ALEXANDERSON S., GUSTAFSON J., BESKOW J., HENTER G. E., SZÉKELY É.: Integrated speech and gesture synthesis. In *Proc. ICMI* (Montréal, QC, Oct. 2021), vol. 23, ACM, pp. 177–185. doi:10.1145/3462244.3479914. 20
- [WGKB21] WOLFERT P., GIRARD J. M., KUCHERENKO T., BELPAEME T.: To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In *Proceedings of the International Conference on Multimodal Interaction* (2021). 16
- [Whi03] WHITTAKER S.: Theories and methods in mediated communication: Steve whittaker. In *Handbook of discourse processes*. Routledge, 2003, pp. 246–289. 2
- [WHTL22] WEI Y., HU D., TIAN Y., LI X.: Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579* (2022). 2
- [WLI21a] WU B., LIU C., ISHI C. T., ISHIGURO H.: Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-gan and unrolled-gan. *Electronics* 10, 3 (2021), 228. 9, 10
- [WLI21b] WU B., LIU C., ISHI C. T., ISHIGURO H.: Probabilistic human-like gesture synthesis from speech using gru-based wgan. In *Companion Publication of the 2021 International Conference on Multimodal Interaction* (2021), ICMI '21 Companion, ACM, pp. 194–201. doi:10.1145/3461615.3485407. 9, 10
- [WMK14] WAGNER P., MALISZ Z., KOPP S.: Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014), 209–232. 5
- [WML21] WANG T.-C., MALLYA A., LIU M.-Y.: One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2021). 20
- [WN13] WANG Y., NEFF M.: The influence of prosody on the requirements for gesture-text alignment. In *International Workshop on Intelligent Virtual Agents* (2013), Springer, pp. 180–188. 4
- [WRB22] WOLFERT P., ROBINSON N., BELPAEME T.: A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems* (2022). 16
- [WSRS*17] WANG Y., SKERRY-RYAN R., STANTON D., WU Y., WEISS R. J., JAITLEY N., YANG Z., XIAO Y., CHEN Z., BENGIO S., LE Q., AGIOMYRGIANNAKIS Y., CLARK R., SAUROUS R. A.: Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech* (Grenoble, France, 2017), INTERSPEECH'17, ISCA, pp. 4006–4010. doi:10.21437/Interspeech.2017-1452. 17
- [WSZ*18] WANG Y., STANTON D., ZHANG Y., SKERRY RYAN R., BATTENBERG E., SHOR J., XIAO Y., JIA Y., REN F., SAUROUS R. A.: Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proc. International Conference on Machine Learning* (2018), pp. 5180–5189. 19
- [WTGM22a] WINDLE J., TAYLOR S., GREENWOOD D., MATTHEWS I.: Arm motion symmetry in conversation. *Speech Communication* 144 (2022), 75–88. 6
- [WTGM22b] WINDLE J., TAYLOR S., GREENWOOD D., MATTHEWS I.: Pose augmentation: Mirror the right way. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents* (2022), IVA '22, ACM, pp. 33:1–33:3. doi:10.1145/3514197.3549677. 6
- [WWK15] WATTS O., WU Z., KING S.: Sentence-level control vectors for deep neural network speech synthesis. In *Proc. Interspeech* (2015), pp. 2217–2221. 19
- [XBHN13] XU J., BROEKENS J., HINDRIKS K., NEERINCX M. A.: Mood expression through parameterized functional behavior of robots. *IEEE*, 2013. 2
- [YCL*20] YOON Y., CHA B., LEE J.-H., JANG M., LEE J., KIM J., LEE G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16. 4, 8, 9, 10, 12, 14, 17, 18, 19, 20
- [YKJ*19] YOON Y., KO W.-R., JANG M., LEE J., KIM J., LEE G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proc. ICRA* (2019), pp. 4303–4309. 4, 6, 7, 10, 11, 12, 17, 18
- [YPJ*21] YOON Y., PARK K., JANG M., KIM J., LEE G.: Sgtoolkit: An interactive gesture authoring toolkit for embodied conversational agents. In *ACM Symposium on User Interface Software and Technology (UIST)* (2021). 20
- [YWJ21] YE Z., WU H., JIA J.: Human motion modeling with deep learning: A survey. *AI Open* 3 (2021), 35–39. doi:10.1016/j.aiopen.2021.12.002. 2
- [YWK*22] YOON Y., WOLFERT P., KUCHERENKO T., VIEGAS C., NIKOLOV T., TSAKOV M., HENTER G. E.: The GENE Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction* (2022), ICMI '22, ACM. 7, 8, 10, 13, 16, 17, 18, 19, 20
- [YYH20] YANG Y., YANG J., HODGINS J.: Statistics-based motion synthesis for social conversations. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 201–212. 4, 8, 19
- [ZBC22] ZHOU C., BIAN T., CHEN K.: Gesturmaster: Graph-based speech-driven gesture generation. In *Proc. ICMI* (2022), ICMI '22, ACM. 5, 10, 13, 17, 18
- [ZQZ*22] ZHUANG W., QI J., ZHANG P., ZHANG B., TAN P.: Text/speech-driven full-body animation. Demo track. 10, 14
- [ZRMOL22] ZABALA U., RODRIGUEZ I., MARTÍNEZ-OTZETA J. M., LAZKANO E.: Modeling and evaluating beat gestures for social robots. *Multimedia Tools and Applications* 81, 3 (2022), 3421–3438. 9
- [ZYL*22] ZHOU Y., YANG J., LI D., SAITO J., ANEJA D., KALOGERAKIS E.: Audio-driven neural gesture reenactment with video motion graphs. In *Proc. CVPR* (2022), pp. 3418–3428. 10, 11, 13, 18