Speech-based Gesture Generation for Robots and Embodied Agents: A Scoping Review

ANONYMOUS AUTHOR(S)

Humans use gestures as a means of non-verbal communication. Often accompanying speech, these gestures have several purposes but in general aim to convey an intended message to the receiver. Researchers have tried to develop systems to allow embodied agents to be better communicators when interacting with humans via using gestures. With advances in artificial intelligence, the methods used by these systems have evolved throughout the years. In this article, we present a scoping literature review of the methods and the metrics used to generate and evaluate co-speech gestures. After collecting a set of papers using a term search on the Scopus database, we analysed the content of these papers based on methodology (i.e. model, dataset used), evaluation measures (i.e. objective and subjective) and limitations. The results indicate that data-driven approaches are used more frequently with high agility, continuity and naturalness, and deep learning becomes the main idea for the problem of speech-driven gesture generation. However, researchers tend to implement their model on virtual agents rather than real robots. In terms of evaluation, we found a trend of combining objective and subjective metrics, while no standards exist for either. This literature review provides an overview of the research in the area and more specifically insights on the trends and the challenges to be met in building a system to automatically generate gestures for robotic agents.

Additional Key Words and Phrases: literature review, co-speech gestures, gesture generation, robot

ACM Reference Format:
Anonymous Author(s). 2021. Speech-based Gesture Generation for Robots and Embodied Agents: A Scoping Review. In *Human Agent Interaction Conference (HAI), 2021, Online.* ACM, New York, NY, USA, 13 pages. https://doi.org/NA

1 INTRODUCTION

Gestures are a type of non-verbal communication used in human-human interaction along with speech. An extensive work aiming to define and categorise gestures as well as to analyse their impact on human-human communication has been proposed by Kendon [17]. One of the important aspect of gestures, in comparison with other non-verbal communication cues such as proxemics or haptics, is that they aim to convey a message [6]. Co-speech gestures are hence found to enrich, clarify or elaborate on objects and actions descriptions communicated with the verbal message. Other gestures, such as beat gestures are used to keep the rhythm of the speech or emphasise certain words. Gestures are also used to show agreement, disagreement, questioning or doubt associated with the speech utterance. Besides, gestures can be used for speech disambiguation in joint attention and pointing tasks [26]. When expressing gestures, several parts of the body can be involved: one hand (i.e. a good luck gesture with fingers crossed), two hands (i.e. time out gesture, used by sport referees) or hands and other body parts (i.e. the cut-throat gesture to signify a threat). In general, most of the gestures would be using only the upper body parts (fingers, hands, arms, head, and torso).

In 1992, McNeill [25] proposed a classification gestures as: 1) iconic (representing objects or actions), 2) metaphoric (representing abstract concepts), 3) deictic (pointing), and 4) beat (emphasizing on words or keeping with the rhythm of speech). A year after, Streeck found that speech and gestures were tightly temporally coordinated in several cultures [36].

Hence, when dealing with the design of embodied social agent, it becomes crucial to adequately design gestures and to temporally integrate them with speech delivery [24]. Indeed, the risk for embodied agent is to have dissonant non-verbal behaviours from verbal messages that could lead to a particular discomfort for the human interlocutor [15]. In Human-Robot Interaction, co-verbal gestures have been found to enhance robot's likability [32] and to increase future contact intentions compared to when a robot does not use gestures. Besides, Salem et al. [33] found that incongruent speech-gesture influenced negatively the participants' task related performance in their study. On the other hand, during a narration scenario, a robot's use of co-verbal gestures has also helped participants improve information recall and has influence the perceived competence of the robot [3, 4]. In conversational contexts, Stolzenwald et al. [35] proposed to a system able to capture the human gesture-style and use it to generate robot's gestures in order to improve rapport. Designing gestures for every possible speech utterance does not seem to be a scalable solution, especially when considering the variety of embodied agents and robots embodiment constraints (number of joints and degrees of freedom).

Since several years, researchers have been trying to automatically generate gestures for embodied conversational agent. In their recent review, [40] propose to focus on evaluation practices of gesture generation in embodied conversational agents. While the review presents an interesting collection of related research and an in-depth analysis of evaluation methods in the domain, we believe that the lack of discussion on the methods used to generate the gestures, is not allowing to appreciate the choice of authors in terms of evaluations. In particular, what kind of input data are used for co-speech gesture generation? What datasets are used? What features are extracted? What models are used for the generation itself? In order to address these important technical questions, we present in this paper a scoping literature review and analysis of 19 papers that were published in peer-reviewed venues. We first present the method we used to collect and analyse these papers. We then present the results highlighting the trends and research gaps.

2 METHOD

In order to draw the map of the existing literature dealing with speech-based gesture generation for robots, we used a literature scoping approach and followed the guidelines by [30]. This method was previously used in Human-Agent Interaction to identify research trends in terms of usage of novel technology [28]. It starts with the data collection which is performed via queries on an identified database. After querying and filtering based on the exclusion criteria (see section 2.2), we extracted the relevant information from each of the selected publication in order to address our review questions.

2.1 Data collection

We set out by conducting a term search using the Scopus database [1] early Feb 2021.

Our aim was to scope for publications that were looking at the generation of gestures for robots using speech. Our search query was performed on the paper's title, abstract and keywords and included: {robot AND gesture AND generation AND speech}. With this search we gathered 53 papers, from which we removed outputs that were not peer reviewed articles (i.e. front page of conference proceedings), leaving us with 49 papers.

[1] https://scopus.com/

This search was combined with another set of 4 papers which had already been reviewed by the authors and were relevant to the topic of this literature review.

Ultimately, we gathered a set of 53 papers, which were screened afterwards to match the inclusion criteria.

2.2 Exclusion criteria and Screening

In order to focus our review on papers relevant to our topic, we applied the following exclusion criteria. A paper was excluded if:

- it was not focusing on gesture generation (some papers focused on turn-taking or social behavior modelling)
- it was dealing with teleoperation (gesture mapping) rather that autonomous generation of gestures
- it was not using speech as an input (some papers used BCI or music)
- it was about generating more than gesture (such as multi-modal interfaces) or not using arms as an output (only spine or head were used)

After applying these exclusion criteria, we were left with 19 papers to analyse.

2.3 Data extraction

We decided to set a reading grid in order to extract relevant information from the pool of papers selected. The reading scheme aimed to give an overview and contained:

(1) year of publication,
(2) the output agent: virtual avatar or robot,
(3) the type of model used for the generation,
(4) the dataset used as an input,
(5) the features that were used by authors
(6) the metrics, objective and/or subjective used by authors to evaluate their work.

3 RESULTS

In this section, we present the results of our analysis from the 19 publications included (see Table 3).

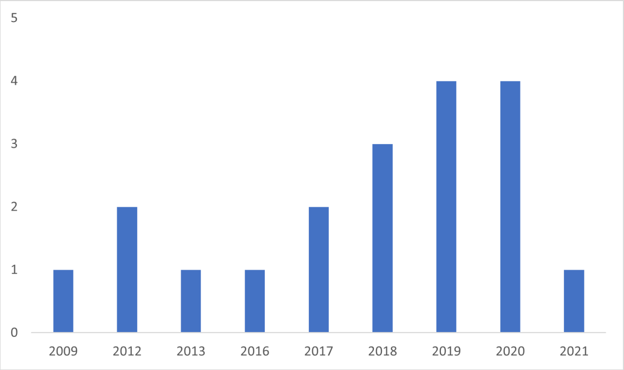

Fig. 1. Number of publications that were included per year

While the number of publications that matched our criteria could seem quite low (compared to other body of research), we observe an increasing number of papers since the past six years (see Figure 1).

Robot	Body Parts	DoF used for gestures
NAO [43]	head, shoulders, elbows and wrists	12
Android Erica [14]	face*, head, neck, torso, arms, hands and fingers* (thumb, index and other fingers together)	16
Engkey [19]	head, shoulders and elbows	8
Stewart Robot [18]	shoulders, elbows and wrists	12
BERTI [3]	neck, waist, shoulders, elbows, wrist, hands and fingers*	15
Pepper [34]	head, shoulders, elbows and wrist	12

Table 1. List of robots used to evaluate the speech-based gesture generation models. The DoF presented is including all the members - head, and two arms if present. (DoF: Degree of Freedom; *: actuators present but not used)

In terms of the output platforms reported, we found that 7 papers were using robots (see Table 1): NAO, Pepper (simulated), Android Erica, Engkey, Steward Robot and BERTI. The other 12 papers used virtual agents as output platforms, but the authors aimed to eventually test their model on a robotic platform. Table 1 gives some details on the movement capacity of each of these robots. We notice that although some of them have a high degree of freedom, none of them is comparable to human capacities (20+ DoF per arm). Besides, if some robots feature actuated fingers and/or face (e.g. Android Erica and BERTI), none of the proposed model tested on robots used them for gesture generation. Similarly, none of the paper reviewed using virtual avatars modelled the fingers in the gestures due to poor data quality [23].

Dataset	Language	Description
Aoyama Gakuin [38]	Japanese	There are 1049 sentences in this dataset: 68.41% are metaphoric gestures, 23.73% are beat gestures, and others are iconic and deictic gestures. The dataset is 298 minutes long.
TED Gesture[43]	English	Only upper body gestures were extracted; The total duration of the valid data was 97 h. The gesture poses were resampled at 15 frames per second, and each training sample having 34 frames was sampled with a stride of 10 from the valid video sections.
Trinity [7]	English	An actor recorded 25 takes, ranging between 10 and 20 minutes each, totalling over 370 minutes of data. The whole body motion was recorded using a MoCap system with 53 markers.

Table 2. Most commonly used datasets in our collection of publications

As for the most used datasets, we found three: Aoyama Gakuin, TED Gesture and Trinity (see Table 2). Unfortunately, these datasets do not have consistent criteria for the speech gestures data. Also, the description of these datasets does not provide enough details such of types of gestures (only Aoyama Gakuin reports on this) or the number and position of joints considered.

Paper	Agent	Methodology			Evaluation	
		Model Used	Input Dataset	Input Features	Subjective Metrics	Objective Metrics
Wu et al. 2021 [41]	VA	cGAN	Aoyama Gakuin	Prosodic feature	Questionnaire (Naturalness, Time Consistency, Semantics)	Log-Likelihood Standard Error
Yoon et al. 2020 [42]	VA	GAN	TED Gesture	32-D feature vectors for text and audio, one-hot vectors for speaker ID	Questionnaire (Preference, Human-likeness of motion, Speech–gesture match)	FGD
Kucherenko et al. 2020 [23]	VA	Autoregressive Model	Trinity	Semantic, Acoustic features	Questionnaire (Human-likeness , Semantics, Time Consistency)	RMSE, Acceleration, Jerk
Ahuja et al. 2020 [1]	VA	Mixture-Model	PATS	MFCC feature	Questionnaire (Naturalness, style transfer correctness)	Probability of Correct Keypoints, F1, IS
Alexanderson et al. 2020 [2]	VA	Autoregressive model	Trinity	MFCC	Questionnaire (Human-likeness, Appropriateness style control, Full-body gestures)	-
Ferstl et al. 2019 [8]	VA	Mixture-model	Trinity	MFCC, F0	-	-
Kucherenko et al. 2019 [21]	VA	The Denoising Autoencoder Neural Network	Aoyama Gakuin	MFCC, Prosodic, Spectrogram features	Questionnaire (Naturalness, Time consistency and Semantic consistency)	APE
Yoon et al. 2019 [43]	NAO	RNN	TED Gesture	One-hot vector for text	Questionnaire (Anthropomorphism, Likeability, Speech-gesture correlation)	-
Kucherenko et al. 2019 [22]	VA	The Denoising Autoencoder Neural Network	Aoyama Gakuin	MFCC feature	Questionnaire (Naturalness, Time consistency and Semantic consistency)	Jerk
Shimazu et al. 2018 [34]	Simulator of the Pepper	LSTM	TED Talks	Prosodic feature	Questionnaire (Naturalness, Skill of presentation, Utilization of gesture, Vividness, Enthusiasm)	Log-likelihood
Hasegawa et al. 2018 [10]	VA	Bi-Directional LSTM	Aoyama Gakuin	MFCC feature	Questionaire (Naturalness, Time Consistency Semantic Consistency)	APE
Ishi et al. 2018 [14]	Android Erica	Probabilistic models	ATR	Prosody feature	Questionnaire (Human-likeness, Suitability, Naturalness, Frequency, Timing)	-
Takeuchi et al. 2017 [37]	VA	Bi-Directional LSTM	Aoyama Gakuin	MFCC feature	Questionnaire (naturalness, Time consistency Semantic consistency)	-
Ondavs et al. 2017 [29]	NAO	Multimodal dialogue system	Rules for gesture generation	-	Questionnaire (Naturalness, Smoothness, User acceptance)	-
Kadono et al. 2016 [16]	VA	Classification model	Iconic Gestures Dictionary	-	-	-
Mlakar et al. 2013 [27]	VA	Language Independent Engine	Expressive Dictionary of Gestures	-	Questionnaire (Content Matching, Synchronization, Fluidness, Speech-Gesture Matching, Execution Speed, Amount of Gesticulation)	-
Kim et al. 2012 [19]	Engkey	Mixture-Model	2012 TED conference	-	Questionnaire (Plausibility, Naturalness, Repetitiveness)	-
Kim et al. 2012 [18]	Steward Robot	Hierarchical Structured Model	Word-gesture database	-	Questionnaire (Suitability of Gestures, Synchronization, Scheduling)	-
Bremner et al. 2009 [3]	BERTI	Universal Beat Gesture Rules	Chat show videos	-	Pilot study for the time spent with focused attention	-

Table 3. List of publications included and analysed according to the Methodology and Evaluation dimensions. (VA: Virtual Agent, Aoyama Gakuin: Collected by [38], Trinity: Collected by [7])

3.1 Methodology

There are 13 papers in our pool that presented a data-driven generation method while the other 6 papers applied rule-based generation methods.

3.1.1 Rule-based. Common speech gestures are associated with speech context, speech audio and interactive objects. Among these, speech context has become the main content for studying the generation of speech gestures because of its importance and relevance [17]. A simple way to generate gestures from speech content is to associate speech words with gestures [3], which is the basic idea of the rule-based generation methods. These studies used experts, manually defined gesture generation rules [18, 19, 27]. The flow chart shown in Figure 2 presents an overview of the rule-based method. As mentioned before, the speech texts have become the main input for the generation of speech gestures. In the speech text analyzing phase, researchers [16, 18, 27] applied morphological analysis methods to disassemble each word into a set of functions and content words. One work by Kim et al. [19] utilized dialogue sentence analysis method to extract correspond punctuation marks of a sentence; mainly used for beat gestures. After text analyzing, gestures can be selected from the Word-to-Gestures database or dictionary by the gesture selection rules. Finally, motion sequence can be generated by referring to starting and ending time of utterance of each syllable from the Text-to-Speech (TTS) processor. A special study by Kadono et al. [16] created its own gestures dictionary for iconic gestures from online images by using image processing and clustering techniques, but the processing of gesture generation is still rule-based. By exploiting that machine learning method, it could solve the problem of repetitive and labour-intensive creation for gestures dictionary.

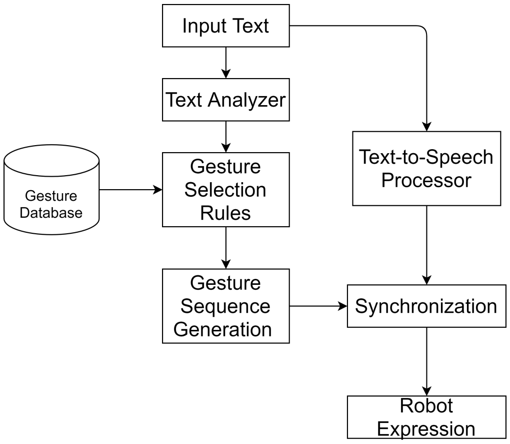

Fig. 2. Overview of the rule-based gesture generation system

3.1.2 Data-driven. As for the data-driven generation methods, 10/13 of the studies employed unimodal input: 8 studies used speech audio and 2 studies used speech text, and only three studies applied multimodal input which combined speech audio and text or speaker identity as optional. An overview of data-driven methods is given in Figure 3. We can frame the data-driven method for gesture generation as follows: for the given speech features S that are extracted from a frame of speech text or audio or both of text and audio at a fixed intervals i, the task is to generate a mapping motion sequence $\hat{G}$ that can be accompanied with uttering speech. The speech features S typically include: semantic features

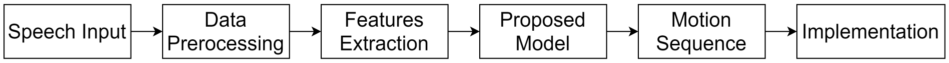

Fig. 3. Overview of the data-driven gesture generation system

from speech texts, prosodic features such as F0 (fundamental frequency), energy and their derivatives, and MFCC (Mel-Frequency Cepstral Coefficients) which mainly used in speech recognition. In addition, the ground truth G and generated gestures $\hat{G}$ represented as skeleton coordinates in Euler angles sequences. Moreover, because the sequence lengths of audio and text are different in most scenarios, that caused a challenge issue for multimodal input methods: audio-text alignment. In order to address this, researchers [23] encode the silence and filler words (excluding semantic features) as fixed vectors which only have a constant negative number. Similarly, Yoon et al. [42] insert padding tokens into the text sequence to make it has the same length as the audio sequence, but this method requires that the exact time at which the words are spoken. Three state-of-the-art studies [8, 41, 42] adopted GAN models (Generative Adversarial Networks) [9]. Their main idea of GAN (shown in the Figure 4) is based on game theory scenario, in which the generator must compete with its opponent, discriminator. The generator learns to generate reasonable motion sequences as input of the discriminator which is for training the negative samples. The discriminator tries to distinguish the samples from real dataset and samples from generator, and indicates whether the input samples are real training samples or fake samples from generator. Through backpropagation, the classification of discriminator provides a signal, and then the generator updates its weight. Finally, if the generator is trained successfully, the discriminator's ability of distinguishing true from false will become worse. Specially, one of them [8] used multiple discriminators for output optimization. They showed merit of adversarial training in that the generated motion appears less damped with large moving range when comparing with other data-driven models. When human express the same language in different situations, they have similar but different gestures. Unlike the one-to-one mapping structure of other data-driven models, the one-to-many mapping structure of GAN models can achieve multiple gestures generation for one same speech input with multiple random noises. Two studies from KTH proposed their own autoregressive models to dispose of the alignment between the speech and gesture [1, 23]. Both RNN (Recurrent Neural Network) studies adopted LSTM (Long Short-Term Memory) architecture [12] except the model from Yoon et al. [43] which used GRU (Gated Recurrent Unit) architecture [5]. RNN faces short-term memory problems, this is due to the problem of vanishing gradient. Compared with other neural network structures, RNN is more prone to suffer from gradient disappearance as it processes more steps. To avoid gradient disappearance, two specialised versions of RNN were created: GRU and LSTM. LSTM uses memory cells to store the activation values of the previous word in a long sequence (see Figure 5). Therefore, it should be efficient for specific data, such as gestures related to the whole sentence or gestures beyond a sentence [10]. When comparing LSTM

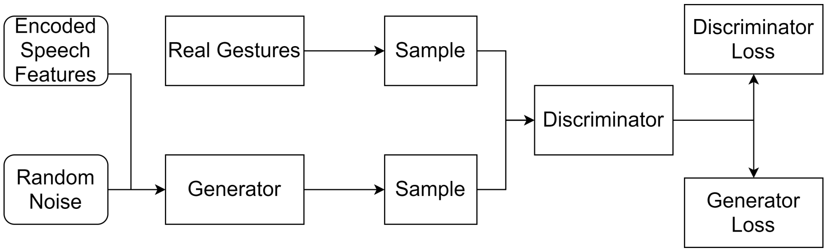

Fig. 4. Overview of GAN structure

with RNN, LSTM introduces more gates, which control the mixing of per trained weights and stream. Therefore, it brings flexibility in controlling the output as well as the complexity. GRU is pretty much similar to LSTM, but has less gates, and needs less time for processing. Yoon et al. [43] employed this architecture to achieve a working model on NAO in real time.

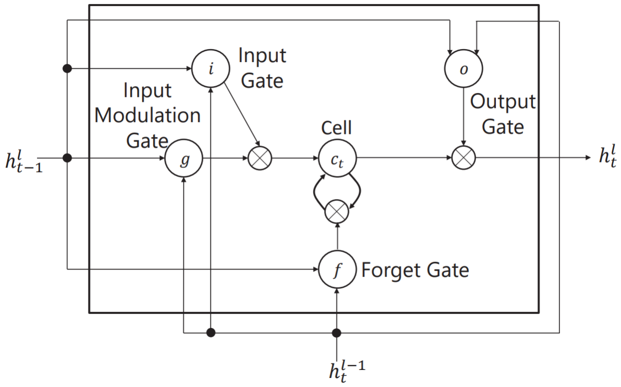

Fig. 5. Internal Architecture of LSTM [10]

3.2 Evaluations

Since there is no perceptual quality metrics for human gesture, it is difficult to evaluate performances of the gesture generation models objectively, and only 8/19 of papers implemented an objective evaluation. Although participants can assess gestures through subjective evaluation, objective evaluation indicators can help to compare fairly and repeatedly gestures generated by advanced models. At present, the authors of these papers indicated that there are no unified evaluation metrics for gesture generation. The study by [42] introduced a novel objective metrics Frechet Gesture Distance (FGD) inspired from Frechet Inception Distance (FID) metric [11] widely used in image generation. Authors proposed to evaluate the FGD metric looking at agreement with human raters on preference, human-likeness and speech–gesture match. Their metric reaches a maximum of 70% agreement which was lower than the between human agreement. The authors explain this deviation by the fact that their measure might fail to capture subtle motions and does not allow to assess motion quality and diversity as separate dimensions. Before that, Average Position Error (APE), Log-Likelihood Standard Error, RMSE (Root-Mean-Square Error) plus acceleration and jerk were the metrics the most commonly used for gesture generation models. These metrics focuses on measuring the difference between original and generated gestures.

While these objective metrics can give an indication of the coordination between speech and motion, they do not seem to be enough as they do not reflect the quality of generated gestures correctly (i.e. naturalness, credibility). In that sense, subjective metrics are necessary for the assessment of generated gestures.Out of 19 studies, 17 conducted human perceptual study using questionnaires. Specially, only one which used the time spent with focused attention when looking at the robot [3] to assess human perception of the gestures. Three studies [1, 23, 42] conducted study by employing crowdsourcing solutions (Amazon MTurk). In terms of dimensions evaluated via human feedback, authors aimed to evaluate several aspects of the gestures: 1) the accuracy of the dynamics (timing and meaning of the gesture),

2) the naturalness, and 3) the competency of the embodied agent. For the accuracy of the dynamics, participants were often asked to rate gestures in terms of time consistency, semantics and/or speech-gesture match. The naturalness was also surveyed, asking participants to rate the human-likeness, the anthropomorphism, the fluidity and the plausibility of the gestures. Finally, as the scenario for which the gestures were generated was often public speaking, we found that some authors evaluated the presentation skills, vividness and enthusiasm of the avatar when performing a talk ([34]). These last type of qualitative measure seems to be relevant as they evaluate the gestures generated in a context (here public speech).

3.3 Limitations

Only 6 papers realised their implementation of gesture generation on the real robots while one [34] was implemented on robot simulator. Most of these studies are limited by the hardware constraints (limited degree of freedom, not direct mapping of the joints, and limited execution of the gestures). However, this could be addressed in the near future with the rapid growth of machinery and robotics.

For the rule-based generation methods, the most significant limitation is that it takes a tremendous cost and effort to manually specify a gesture shape for each word. It causes the trend that recent works focus on data-driven generation rather than rule-based generation. Also, in the generated motion from lexis to gestures handcrafted mapping showed less diverse expression [18].

For the data-driven generation methods, the drawbacks of single modal input are obvious, they results poor generation results. Specifically, lack of semantics caused by using speech audio as a single modality is still a considerable barrier for excellent performances. Likewise, using only speech text as a single input causes the generated motion to appear less prosodic and could not be tightly coupled with beat gestures for instance. However, multimodal generation models can achieve both semantic-aware and prosodic-aware gesture generation (see [23, 34, 42]). One common limitation of data-driven generation methods is that a large sample of parallel speech and gestures training data with sufficient quality and diversity is needed in order to achieve a good performance.

In terms of evaluation, we found that the main issue is a lack of common metrics to assess accuracy (semantic and timing), diversity and quality of the generated gestures. We saw that combining objective and subjective measures together seemed to better capture the overall performance of the generative model. A clear description of the metrics (objective and questionnaire) will be necessary to ensure the reproducibility of the studies and will hopefully lead to a standardisation of measures in the field.

4 DISCUSSION

Through these studies, findings revealed a shift in terms of methodology from rule-based generation methods to data-driven generation methods. More specifically, deep learning becomes the main stream in the data-driven generation methodology, and it starts from simple RNN to LSTM or GRU, then develops to Bi-Directional LSTM, and the most advanced model GAN. This can be explained by both the availability of open datasets and the apparition of more sophisticated DNNs architecture allowing to handle the sequential constraints of co-speech gesture generation.

An interesting finding is that none of the studies talked about the influence of culture or language on gesture generation. On the flip side of co-speech gesture are symbolic gestures (emblems gestures), which can convey language information independently. The symbolic gestures may lead to misunderstanding in cross-culture. For example, "OK" gesture expresses agreement, approval and that everything is well in Australia, however, it indicates money in the

Japanese culture. So symbolic gestures vary by culture, current studies have not considered this gestures due to cross-cultural misunderstanding.

Findings indicated that most of the rule-based generation methods lack random variation and naturalness which can be attributed to the gestures generated from predefined gestures dictionaries. In addition, most studies which adopted unimodal input (speech audio or text) tend to produce defective output, e.g. as it mentioned in the paper [43], the approach encountered an unnatural mapping problem that the synthesized audio and the generated gestures can not be tightly coordinated in the experiments. Ishi et al.[14] experienced the same shortfall due to both of them adopting speech text-based method; likewise, results from [41] encountered lack of semantics for some meaningful motion by applying the speech audio-based method.

More and more studies begin to use both subjective and objective evaluation, however, there is no unified and standard objective evaluation metrics for generated gestures. Our analysis also found that recent state-of-the-art publications preferred to apply generated gestures on virtual agents rather than the real robots. While the use of simulation in robotics is common, it comes also with the drawback of not being able to accurately capture the physical constraints as present in a real robot platform [13]. More research should be conducted to assess how effectively these simulated models could be transferred to real robots.

While gestures are heavily dependent on fingers movement [17], we found a lack of use of fingers in the gesture models that are proposed in the current literature. With better and higher quality capture of fingers motion, one could image to achieve more accurate and natural gesture generation, especially seeing the improvement of sign language recognition models [31]. We argue that there is an urgent need for a dataset with substantial amounts of quality speech gestures of various speakers, and we believe that this area will benefit from a sires of standardized objective evaluation metrics which can make generated gestures have comparability across different studies.

5 CONCLUSION

In this survey, we review the studies of speech-driven gesture generation for robots, focusing on the methods, evaluation measures and limitations. In conclusion, while there are no standard evaluation metrics, more and more works are aiming to exploit multimodal model based on deep learning method for generating better speech gestures.

This scoping review is subject to some limitations. Some relevant studies not listed in the Scopus database, or not using our search terms could have been missed. In addition, other words such as “humanoid” or “android” can be used to address robots, which may have prevented the retrieval of potential relevant publications during the initial search by terms. Second, due to our exclusion criteria, we only reported publications which focused on gesture generation and reported in details the method and evaluation used. Other publications which focused on interactive aspects of human-robot interaction might have been excluded while still linked to our research topic. We conclude by presenting several potential future directions that have been identified as research gaps.

One possible area for future work is to investigate natural and richer set of gestures. As shown with this literature review, a lot of the datasets cover the same context - talks and presentation - gestures, however co-speech gestures are present in a variety of contexts that haven’t been investigated so far. This would necessitate the use of novel datasets covering various types of gestures and not only specific to presentation or talks. Another challenge at the moment is the quality of the datasets and in particular the capture of high quality hand and finger motion. Indeed, a lot of gestures rely on hand and finger motions (e.g. pointing and enumerating gestures). To be able to generate similar gestures, one would need to capture both large upper body movements and distal finger motions precisely. One would also need a robot that has enough degrees of freedom to render these types of gestures. While capturing more precisely all upper body segments seem feasible, only two robots currently used in this field of work would be able to render finger motions (see Table 1, BERTI and Erica). We also found that most of the work published in the field relied on only few datasets. While this is beneficial for the research community to benchmark and to compare performances, it also presents some drawbacks; one of which is that most of these models are limited in terms of language and culture that they cover due to the fact that the datasets used are only in English or in Japanese. We know that co-speech gestures are heavily culture and language dependent [20]. Therefore, novel datasets would be needed to investigate the cultural differences and build a cross or multi languages and culture gesture generative model.

This review also highlighted the fact that most data-driven model where not featuring any semantic models. Because co-speech gestures occur in various contexts and for various purposes, it could be interesting to investigate how gestures classification could be used in the generative model, and evaluate the generated outputs per type of gestures (i.e.iconic, beat, ...). Other approach could be considered such as model-based data-driven approaches that could combine intention of communication to speech as an input of the generative models. This would for instance allow to integrate the gesture generative model in a robot cognitive architecture based on the communication acts.

Another possible area of future work is to do some affective computing on the input speech. For a same sentence, different people may speak with various emotions and speed according to the pronunciation, stress and intonation, this should correspond to different co-gestures. Moreover, for the future model structure, the Transformer [39] is a possible solution, which is a new neural network architecture based on self attention mechanism that is good at processing language understanding tasks and requires less computational power, thus improving the training speed by an order of magnitude.

Another challenge that was revealed by this review was the lack of standard objective and subjective metrics. In terms of objective metrics, several aspects would need to be considered: the adequacy of the gestures with regard to the speech, the timing and synchronicity and the variability of the gesture. Indeed, speech and gestures do not have a one-to-one mapping - several gestures could go with a speech utterance and vice-versa. Subjective metrics are also very important to evaluate the quality of the gestures. We found that more work would be needed to build a standard questionnaire allowing to assess the perception of the gestures in context.

Finally, while our review focused on generation of gestures for robots, we found that only few papers were presenting an evaluation of their system on a real robot. Although it can be challenging to do so, it will be the only way to build generative models that take into account the social robot's motion constraints.

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*. Springer, 248–265.

[2] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[3] Paul Bremner, Anthony G Pipe, Mike Fraser, Sriram Subramanian, and Chris Melhuish. 2009. Beat gesture generation rules for human-robot interaction. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 1029–1034.

[4] Paul Bremner, Anthony G. Pipe, Chris Melhuish, Mike Fraser, and Sriram Subramanian. 2011. The effects of robot-performed co-verbal gesture on listener behaviour. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*. 458–465. https://doi.org/10.1109/Humanoids.2011.6100810

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture* (1969), 57–106.

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

[8] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. 1–10.

[9] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).

[10] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500* (2017).

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[13] Sebastian Höfer, Kostas Bekris, Ankur Handa, Juan Camilo Gamboa, Florian Golemo, Melissa Mozifian, Chris Atkeson, Dieter Fox, Ken Goldberg, John Leonard, C. Karen Liu, Jan Peters, Shuran Song, Peter Welinder, and Martha White. 2020. Perspectives on Sim2Real Transfer for Robotics: A Summary of the R:SS 2020 Workshop. (Dec. 2020). https://www.arxiv-vanity.com/papers/2012.03806/

[14] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764.

[15] Wafa Johal, Gaëlle Calvary, and Sylvie Pesty. 2015. Non-verbal Signals in HRI: Interference in Human Perception. In *International Conference on Social Robotics*. Springer, 275–284.

[16] Yuki Kadono, Yutaka Takase, and Yukiko I Nakano. 2016. Generating iconic gestures based on graphic data analysis and clustering. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 447–448.

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

[18] Heon-Hui Kim, Yun-Su Ha, Zeungnam Bien, and Kwang-Hyun Park. 2012. Gesture encoding and reproduction for human-robot interaction in text-to-gesture systems. *Industrial Robot: An International Journal* (2012).

[19] Jaewoo Kim, Woo Hyun Kim, Won Hyong Lee, Ju-Hwan Seo, Myung Jin Chung, and Dong-Soo Kwon. 2012. Automated robot speech gesture generation system based on dialog sentence punctuation mark extraction. In *2012 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 645–647.

[20] Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes* 24, 2 (2009), 145–167.

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

[22] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2019. On the importance of representations for speech-driven gesture generation. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2072–2074.

[23] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250.

[24] Quoc Anh Le, Souheil Hanoune, and Catherine Pelachaud. 2011. Design and implementation of an expressive gesture model for a humanoid robot. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 134–140.

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

[26] Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. 2014. Exploring a model of gaze for grounding in multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 247–254.

[27] Izidor Mlakar, Zdravko Kačič, and Matej Rojc. 2013. TTS-driven synthetic behaviour-generation model for artificial bodies. *International Journal of Advanced Robotic Systems* 10, 10 (2013), 344.

[28] Mohammad Obaid, Wafa Johal, and Omar Mubin. 2020. Domestic Drones: Context of Use in Research Literature. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (Virtual Event, USA) *(HAI '20)*. Association for Computing Machinery, New York, NY, USA, 196–203. https://doi.org/10.1145/3406499.3415076

[29] Stanislav Ondáš, Jozef Juhár, Matúš Pleva, Peter Ferčák, and Rastislav Husovskỳ. 2017. Multimodal dialogue system with NAO and VoiceXML dialogue manager. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 000439–000444.

[30] Micah DJ Peters, Christina M Godfrey, Hanan Khalil, Patricia McInerney, Deborah Parker, and Cassia Baldini Soares. 2015. Guidance for conducting systematic scoping reviews. *JBI Evidence Implementation* 13, 3 (2015), 141–146.

[31] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2020. Sign language recognition: A deep survey. *Expert Systems with Applications* (2020), 113794.

[32] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot. In *International conference on social robotics*. Springer, 31–41.

[33] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

[34] Akihito Shimazu, Chie Hieida, Takayuki Nagai, Tomoaki Nakamura, Yuki Takeda, Takenori Hara, Osamu Nakagawa, and Tsuyoshi Maeda. 2018. Generation of gestures during presentation for humanoid robots. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 961–968.

[35] Janis Stolzenwald and Paul Bremner. 2017. Gesture mimicry in social human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 430–436. https://doi.org/10.1109/ROMAN.2017.8172338

[36] Jürgen Streeck. 1993. Gesture as communication I: Its coordination with gaze and speech. *Communications Monographs* 60, 4 (1993), 275–299.

[37] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 365–369.

[38] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction*. Springer, 198–202.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[40] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2021. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. arXiv:2101.03769 [cs.HC]

[41] Bowen Wu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2021. Modeling the Conditional Distribution of Co-Speech Upper Body Gesture Jointly Using Conditional-GAN and Unrolled-GAN. *Electronics* 10, 3 (2021), 228.

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

[43] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.